IWCLUL 2025

The 10th International Workshop on Computational Linguistics for Uralic Languages

Proceedings of the Workshop

December 10-12, 2025

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 979-8-89176-360-9

Preface

Welcome to the Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages (IWCLUL), a dedicated venue for scholars, practitioners, and researchers working in computational linguistics with a particular emphasis on Uralic languages. This year's workshop continues the IWCLUL tradition of fostering interdisciplinary exchange, shared learning, and a collective dedication to tackling the linguistic, technological, and social issues surrounding Uralic languages in the digital era.

The Uralic language family—stretching across Europe and Asia and including languages as varied as Finnish, Hungarian, and the endangered Udmurt and Khanty—brings with it a distinct set of computational challenges. These languages often feature rich morphology, agglutinative patterns, and unique syntactic and phonological structures, all of which demand specialized methods in computational analysis and language technology. Our workshop aims to highlight these complexities and encourage the creation of innovative tools and approaches that not only facilitate digital use of these languages but also contribute to their long-term vitality.

This year, IWCLUL saw a strong and wide-ranging set of submissions from the international research community, demonstrating the sustained interest in computational work on Uralic languages. The accepted papers span numerous topics and language varieties, offering insights into both well-documented and lesser-resourced Uralic languages. This diversity reflects the ongoing growth of the field and the many directions in which researchers are advancing computational linguistics for Uralic languages.

We are also pleased to celebrate Jack Rueter, who was honored with a lifetime achievement award by ACL SIGUR for his decades-long dedication to Uralic language research, documentation, and technology development. His work has had a lasting impact on the community and serves as an inspiration for future generations of researchers.

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 2025 Organizing Committee

Organizing Committee

Organizers

Mika Hämäläinen, Metropolia University of Applied Sciences
Flammie Pirinen, Arctic University of Norway
Lev Kharlashkin, Metropolia University of Applied Sciences
Eiaki V. Morooka, Metropolia University of Applied Sciences
Michael Rießler, University of Eastern Finland
Maarit Koponen, University of Eastern Finland
Ilia Moshnikov, University of Eastern Finland
Diana Kulashekhar, University of Eastern Finland
Mahla Baniasadi, University of Eastern Finland
Nurstrat prova, University of Eastern Finland
Varoon Bakshi, University of Eastern Finland

Program Committee

Laszlo Fejes, Hungarian Research Centre for Linguistics
Gunta Klava, University of Latvia
Sourav Das, Indian Institute of Information Technology Kalyani
Balazs Indig, Eotvos Lorand University
Laleh Davoodi, Abo Akademi University
Flammie A Pirinen, Norgga arktalas universitehta
Trond Trosterud, University of Tromso
Csilla Horvath, University of Szeged
Iaroslav Chelombitko, Neapolis University Paphos
Aleksei Dorkin, University of Tartu
Sebastian Oliver Eck, University of Oxford
Khalid Alnajjar, F-Secure Oyj
Jules Bouton, Universite Paris Cite
Valts Ernstreits, University of Latvia
Michael Riessler, University of Eastern Finland
Timofey Arkhangelskiy, Universitat Hamburg
Samy Ouzerrout, Universite d'Orleans
Jyoti Kunal Shah, Automatic Data Processing
Youngsook Song, Lablup
David Dale, FAIR at Meta
Abhi Desai, New England College

Table of Contents

Program

Wednesday, December 10, 2025

14:00 - 18:00 *Karelian Workshop*

Thursday, December 11, 2025

09:15 - 09:30 *Workshop Opening*

09:30 - 10:30 *Keynote – Josh Wilbur*

10:30 - 11:20 *Lightning Talks*

11:20 - 11:40 *Coffee Break*

11:40 - 12:40 *Oral Session 1*

From NLG Evaluation to Modern Student Assessment in the Era of ChatGPT: The Great Misalignment Problem and Pedagogical Multi-Factor Assessment (P-MFA)
Mika Hämäläinen and Kimmo Leiviskä Leiviskä

Benchmarking Finnish Lemmatizers across Historical and Contemporary Texts
Emily Öhman, Leo Huovinen and Mika Hämäläinen

The world's first South Sámi TTS – a revitalisation effort of an endangered language by reviving a legacy voice
Katri Hiovain-Asikainen, Thomas B. Kjærstad, Maja Lisa Kappfjell and Sjur N. Moshagen

12:40 - 13:40 *Lunch*

13:40 - 15:00 *Oral Session 2*

Can advances in NLP lead to worse results for Uralic languages and how can we fight back? Experiences from the world of automatic spell-checking and correction for Finnish
Flammie A Pirinen

A Hybrid Multilingual Approach to Sentiment Analysis for Uralic and Low-Resource Languages: Combining Extractive and Abstractive Techniques
Mikhail Krasitskii, Olga Kolesnikova, Grigori Sidorov and Alexander Gelbukh

Language technology for the minority Finnic languages
Flammie A Pirinen, Trond Trosterud and Jack Rueter

Thursday, December 11, 2025 (continued)

Kildin Saami-Russian-(English) Parallel Corpus Building
Evan Hansen

15:00 - 15:20 *Coffee Break*

15:20 - 16:40 *Oral Session 3*

SampoNLP: A Self-Referential Toolkit for Morphological Analysis of Subword Tokenizers
Iaroslav Chelombitko, Ekaterina Chelombitko and Aleksey Komissarov

Timur and the Mansi spellchecker
Csilla Horváth

ORACLE: Time-Dependent Recursive Summary Graphs for Foresight on News Data Using LLMs
Lev Kharlashkin, Eiaki V. Morooka, Yehor Tereschenko and Mika Hämäläinen

Creating a multi-layer Treebank for Tundra Nenets
Nikolett Mus, Bruno Guillaume, Sylvain Kahane and Daniel Zeman

Friday, December 12, 2025

10:00 - 10:00 *Day 3 Opening*

10:00 - 11:00 *Oral Session 4*

Benchmarking Large Language Models for Lemmatization and Translation of Finnic Runosongs
Lidia Pivovarova, Kati Kallio, Antti Kanner, Jakob Lindström, Eetu Mäkelä, Liina Saarlo, Kaarel Veskis and Mari Väina

Fine-Tuning Whisper for Kildin Sami
Enzo Gamboni

Digitization Work at the Finno-Ugrian Society: Livonian Case Study
Niko Partanen, Jack Rueter and Valts Ernštreits

11:00 - 12:00 *Lunch*

12:00 - 13:00 *Oral Session 5*

Siberian Ingrian Finnish: FST and IGTs
Ivan Ubaleht

Case–Number Dissociation in Finnish Noun Embeddings:fastText vs. BERT Layer Effects
Alexandre Nikolaev, Yu-Ying Chuang and R. Harald Baayen

Evaluating OpenAI GPT Models for Translation of Endangered UralicLanguages: A Comparison of Reasoning and Non-Reasoning Architectures
Yehor Tereschenko, Mika Hämäläinen and Svitlana Myroniuk

13:00 - 13:20 *Coffee Break*

13:20 - 14:20 *SIGUR Business Meeting*

From NLG Evaluation to Modern Student Assessment in the Era of ChatGPT: The Great Misalignment Problem and Pedagogical Multi-Factor Assessment (P-MFA)

Mika Hämäläinen and Kimmo Leiviskä
Metropolia University of Applied Sciences
Helsinki, Finland
first.last@metropolia.fi

Abstract

This paper explores the growing epistemic parallel between NLG evaluation and grading of students in a Finnish University. We argue that both domains are experiencing a Great Misalignment Problem. As students increasingly use tools like ChatGPT to produce sophisticated outputs, traditional assessment methods that focus on final products rather than learning processes have lost their validity. To address this, we introduce the Pedagogical Multi-Factor Assessment (P-MFA) model, a process-based, multi-evidence framework inspired by the logic of multi-factor authentication.

1 Introduction

In recent years, both the field of computational creativity and university pedagogy have faced a growing crisis of evaluation. In creative natural language generation research, Hämäläinen and Alnajjar (2021a) articulated the Great Misalignment Problem, which is the disconnection between a system's problem definition, its implemented method and the evaluation criteria used to assess its performance. This misalignment leads to superficial or misleading conclusions about a system's success, as the evaluation often fails to measure what the system was designed to achieve.

A similar issue now permeates higher education: the act of grading has become increasingly detached from the authentic learning processes it is meant to assess. As generative models such as ChatGPT[1] can produce fluent, original-looking outputs on demand, educators can no longer be sure whether a student's submitted work reflects genuine understanding or merely the clever use of an external artificially cognitive tool.

Just as computational creativity systems generate artefacts whose internal reasoning is opaque, students in the AI-saturated learning environment can now present creative products without revealing the intellectual pathway that led to them. In both cases, the evaluator encounters an artefact, such as a poem, a program, an essay or a design, without direct access to the underlying process that produced it. The traditional product-based evaluation paradigm, which assumes a transparent correspondence between output and competence, thus collapses. The opacity of generative systems mirrors the opacity of modern student work: both appear impressive on the surface, yet the link between creation and creator, between performance and understanding, is obscured.

This parallel suggests that the Great Misalignment Problem has re-emerged in pedagogy under new conditions. In education, the "problem definition" corresponds to the intended learning outcomes; the "method" is the student's learning process; and the "evaluation" is grading. When these three components fall out of alignment - when grades reflect polished submissions rather than genuine cognitive engagement - the integrity of learning assessment is compromised. The university thus faces a dilemma akin to that of computational creativity research: it risks optimizing for the appearance of creativity and competence rather than for the authentic development of these qualities.

To resolve this, both AI research and pedagogy must shift their evaluative focus from product to process. In computational creativity, meaningful evaluation requires attention to how a system produces its output, its generative mechanisms, constraints, and reasoning. Likewise, effective pedagogy must emphasize the learning process itself: reflection, iteration, collaboration, and the student's evolving relationship with knowledge. In both domains, understanding the process behind the product restores transparency, accountability, and interpretability. The challenge, then, is not merely to detect misuse of generative tools but to redesign evaluation frameworks that value the act of cre-

[1] https://chatgpt.com

ation—the unfolding of thought—as the true site of learning and creativity.

2 Related Work

The emergence of LLMs has received a mixed response among educators; several feel that AI is a threat to students' learning (Yu, 2023; Lin et al., 2023; Ogugua et al., 2023), while others focus on it's potential in future of education (Lo, 2023; Cleland Silva and Hämäläinen, 2024; Morgan, 2024; Macias et al., 2024).

Evaluation of NLG systems, creative and regular, has received a fair share of attention in the past (Clark et al., 2021; Freitag et al., 2021; Howcroft et al., 2020). Some researches even highlight that automated evaluation methods such as BLEU are simply not sufficient (Reiter, 2018). Evaluation mainly relies on evaluating the output of such systems rather than taking the creative process into account (Hämäläinen and Alnajjar, 2021b).

In terms of pedagogy, how people learn and how they should be taught is a relatively well-understood phenomenon. Frameworks such as Bloom's (1956) taxonomy, deep learning (see McGregor 2020) and constructive alignment (Biggs, 1996) have been widely used. However, there's a big gap between theory and practice, and ultimately, student assessment relies on grading some sort of a final product of learning such as an essay or thesis.

3 The Great Misalignment Problem in NLG Evaluation

In their work, Hämäläinen and Alnajjar (2021a) identify what they term *the Great Misalignment Problem* in the context of human evaluation in natural language generation (NLG). The core claim is that in much of NLG research that relies on human judgments, there is a systematic misalignment among three key components: (1) how the research problem is defined, (2) how the proposed method or model is formulated and (3) how the human evaluation is conducted. When these three are not tightly aligned, the authors argue, the validity, interpretability and reproducibility of human evaluation outcomes are severely compromised.

The authors support their claim by surveying ten randomly selected papers from ACL 2020 that include human evaluation. In their analysis, they examine (a) whether the problem definition is clearly and narrowly stated, (b) whether the proposed method follows directly from that problem definition, and (c) whether the human evaluation aligns both with the definition and with what the method is intended to model. They report that in only a single case among the ten did all three align. In many cases, the evaluation either ignores aspects of what the method is modeling or tests orthogonal criteria not grounded in the stated problem.

The implications of the Great Misalignment Problem are profound. Because human evaluation may end up measuring something other than what the model is intended to do, the reported improvements or differences in scores cannot reliably be attributed to the proposed method. Instead, they might arise from unintended artifacts, evaluation design biases, evaluator variance or other uncontrolled factors. This undermines claims of advancement, makes comparison across systems less meaningful, and complicates reproducibility. Moreover, the authors point out that human evaluation in NLP is often conducted with insufficient methodological rigor (e.g. vague questions, low numbers of judges and opaque protocols), further exacerbating the misalignment.

To move forward, the authors recommend that NLP researchers take the problem definition seriously and design methods and evaluations so as to maintain alignment. Concretely, they urge narrowing broad, vague problem statements into more precise, measurable sub-tasks; ensuring that modeling decisions correspond to those sub-dimensions; and crafting evaluation questions that directly probe the modeled behavior. They also call for full transparency in evaluation setup (e.g. prompt wording, judge selection and instructions) and suggest that human evaluation practices in NLP could benefit from importation of best practices from fields accustomed to subjective measurement (e.g. social sciences). They do not advocate abandoning human evaluation altogether, but rather reforming it so that it becomes a more trustworthy and interpretable component of NLP research.

4 LLMs and Teachers' Nightmare

Not unlike the findings described by Hämäläinen (2024), we have encountered negative teacher narratives in Metropolia University of Applied Sciences regarding LLM tools. There is a lot of fear among teachers that students would use the new technology to cheat and ultimately pass their courses without learning much. On the other hand, there are also teachers who embrace the new technology and actively use it in their teaching.

It is fair to say that LLMs have introduced a radical disruption to the long-standing epistemic contract between teachers and students. Pedagogically, this contract rests on an implicit trust: that the student's submitted work represents their own intellectual labor and engagement with the learning process. The teacher, in turn, evaluates this work as evidence of learning, understanding and skill development. However, with the advent of LLMs capable of producing contextually appropriate, grammatically flawless and even stylistically distinct texts, this foundational trust is breaking down. Teachers now face the uneasy possibility that a student's polished essay or thoughtful reflection may be more a testament to prompt engineering than to actual comprehension.

From a pedagogical standpoint, this collapse of evaluative certainty threatens the very rationale of assessment (see Şahin et al. 2024; Fagbohun et al. 2024). If an assignment can be completed without genuine learning, then grades cease to measure educational achievement. They become measures of access to tools and skill in their manipulation. The anxiety many teachers experience arises not merely from the fear of academic dishonesty but from the erosion of pedagogical meaning itself. When outputs can no longer be reliably linked to the cognitive processes they are meant to demonstrate, the educational system loses its anchor: learning becomes performative rather than transformative.

The teacher's nightmare is not that students are cheating, but that the act of evaluation has become epistemically hollow. Traditional assessment methods such as essays, reports and even project work are predicated on a production model of learning where outputs reflect mental effort. In the LLM era, this assumption no longer holds. The teacher cannot see how the student arrived at a conclusion, whether the reasoning was genuine or algorithmically scaffolded. Consequently, feedback loses precision, as it targets the product rather than the learner's cognitive or creative process. Pedagogically, this creates a feedback loop of disengagement: students learn to outsource tasks, and teachers, sensing the futility of assessment, may lower expectations or turn toward increasingly mechanistic forms of surveillance.

In essence, LLMs expose the brittleness of output-oriented pedagogy. The crisis they introduce is not technological but epistemological: education must now grapple with redefining what it means to know, learn, and create in a world where human and machine outputs are indistinguishable. Teachers' frustration, then, is not only emotional but structural. It stems from a system designed for a pre-AI understanding of authorship and agency. The nightmare can only end when pedagogical practices evolve to prioritize the evaluation of the learning process itself over the evaluation of a final product, reclaiming assessment as a shared inquiry into learning rather than a judgment of finished artefacts.

The Great Misalignment Problem offers a valuable lens for rethinking student assessment. By highlighting the dangers of disconnect between problem definition, method and evaluation, the same critique encourages educators to scrutinize whether grading practices truly measure the intended learning outcomes. **In pedagogy, this means aligning what we want students to learn (problem definition), how they engage in that learning (method), and how we evaluate their understanding (evaluation)**.

Rather than focusing solely on the final product, educators can design assessments that make the learning process visible through reflective writing, process logs, peer discussions or iterative project development. In doing so, the evaluation becomes an inquiry into alignment itself: does the student's process reflect the intended learning goals, and does the teacher's evaluation capture that process accurately? This alignment-centered approach transforms grading from an act of judgment into an act of dialogue, ensuring that assessment remains meaningful even in an era when the boundary between human and machine creativity is increasingly blurred.

5 Future of Grading: a P-MFA Approach

We propose a novel framework named Pedagogical Multi-Factor Assessment (P-MFA) for grading and learning evaluation designed for the age of generative AI. It builds on the logic of multi-factor authentication (see Ometov et al. 2018): just as digital security no longer relies on a single password, educational assessment should not depend on a single artefact such as an exam or essay. P-MFA therefore verifies learning through multiple complementary "factors," each representing a distinct dimension of competence—what the student knows (knowledge), produces (outputs), can do (application), sustains over time (process continuity), reflects upon (self-evaluation), and connects to real contexts (situated understanding).

By combining these factors, teachers and students co-construct a trustworthy, multi-channel record of learning that is transparent, individualized, and resistant to the misuse of AI. Rather than focusing on control or detection, P-MFA shifts assessment toward alignment: ensuring that what is defined as learning, practiced as learning, and evaluated as learning all converge in an authentic and human-centered educational process.

Theoretically, P-MFA can be understood as a synthesis of constructive alignment and the Great Misalignment Problem. Constructive alignment posits that meaningful learning occurs when intended learning outcomes, teaching activities, and assessment tasks are coherently designed to support one another. The Great Misalignment Problem, in contrast, diagnoses the breakdown of such coherence in research evaluation: when problem definition, method, and evaluation diverge, the resulting claims lose validity.

P-MFA translates this alignment imperative into pedagogy for the generative-AI era, explicitly designing assessment systems that keep the "problem definition" (learning outcomes), the "method" (student learning processes), and the "evaluation" (grading practices) in continuous dialogue. Where constructive alignment emphasizes curriculum design, P-MFA operationalizes alignment through evidence diversity: multiple, process-anchored factors that ensure the assessment remains faithful to both the intention and the practice of learning. In doing so, P-MFA not only safeguards educational integrity against AI-generated artefacts but also reframes assessment as an interpretive act of maintaining epistemic alignment between what learning is meant to achieve, how it unfolds, and how it is ultimately recognized.

In essence, the P-MFA approach operationalizes the Great Misalignment Problem's philosophical insight within the classroom. It demands that educators explicitly design for a**lignment across definition, method, and evaluation**, ensuring that the assessment of learning remains meaningful, transparent, and resilient to technological disruption. By requiring multiple, process-anchored proofs of understanding, P-MFA not only protects the integrity of grading but also redefines it as a dynamic act of alignment, in which learning and evaluation evolve together.

Problem definition in P-MFA is reframed as the articulation of learning outcomes that go beyond static knowledge. Education's aim is no longer to verify that students can produce isolated outputs but that they can understand, apply, reflect and contextualize their knowledge

The Method of P-MFA corresponds to the pedagogical and learning practices through which these factors are activated. Instead of viewing learning as a linear input–output pipeline, P-MFA promotes iterative, reflective and contextual engagement. Students are not passive respondents to tasks but co-designers of their assessment trajectory, selecting factors that align with their goals and contexts

Finally, Evaluation in P-MFA is no longer an isolated measurement but a process of triangulation. Each factor functions as an evaluative lens that confirms or challenges the authenticity of others. The resulting alignment between what was meant to be learned, how learning occurred, and how it is assessed embodies the very correction the Great Misalignment Problem paper sought in NLP research. When teachers adopt a P-MFA framework, grading transforms from a verdict into an inquiry: a structured investigation into whether the student's demonstrated process and outputs align with the course's learning definition. This ensures that evaluation measures authentic engagement rather than algorithmic fluency. Moreover, because AI cannot convincingly reproduce personal reflection or contextual relevance, P-MFA restores pedagogical trust by embedding evaluation in dimensions that remain uniquely human.

6 Conclusions

The challenges faced in both computational creativity and contemporary pedagogy converge on a single epistemic issue: the difficulty of evaluating outputs without understanding the processes that produced them. The Great Misalignment Problem revealed how research can lose validity when problem definition, method, and evaluation drift apart, an insight that now illuminates the crisis of grading in the era of generative AI. Our Pedagogical Multi-Factor Assessment (P-MFA) model offers a concrete response to this challenge by embedding assessment within the learning process itself. Through its multi-factor design—combining evidence of knowledge, production, application, continuity, reflection, and context—P-MFA restores alignment between what learning is intended to achieve, how it unfolds, and how it is recognized. In doing so, it reclaims evaluation as a transparent and dialogic practice, reaffirming the role of assessment not as an act of surveillance or verification, but as an interpretive inquiry into human understanding and growth in an age increasingly mediated by machines.

References

John Biggs. 1996. Enhancing teaching through constructive alignment. *Higher education*, 32(3):347–364.

Benjamin S Bloom. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Tricia Cleland Silva and Mika Hämäläinen. 2024. Innovating for the future: Ai and hrm capabilities for sustainability in higher education. In *Academy of Management Annual Meeting*, volume 2024.

Oluwole Fagbohun, Nwaamaka Pearl Iduwe, Mustapha Abdullahi, Adeseye Ifaturoti, and OM Nwanna. 2024. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 2(1):1–8.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Mika Hämäläinen. 2024. Legal and ethical considerations that hinder the use of LLMs in a Finnish institution of higher education. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 24–27, Torino, Italia. ELRA and ICCL.

Mika Hämäläinen and Khalid Alnajjar. 2021a. The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74, Online. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021b. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95, Online. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Shu-Min Lin, Hsin-Hsuan Chung, Fu-Ling Chung, and Yu-Ju Lan. 2023. Concerns about using chatgpt in education. In *International conference on innovative technologies and learning*, pages 37–49. Springer.

Chung Kwan Lo. 2023. What is the impact of chatgpt on education? a rapid review of the literature. *Education sciences*, 13(4):410.

Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen, and Mika Hämäläinen. 2024. Empowering teachers with usability-oriented llm-based tools for digital pedagogy. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 549–557.

Sue LT McGregor. 2020. Emerging from the deep: Complexity, emergent pedagogy and deep learning. *Northeast Journal of Complex Systems (NEJCS)*, 2(1):2.

Hani Morgan. 2024. Implementing chatgpt to support teachers and promote learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 97(5):125–133.

Divine Ogugua, Seong No Yoon, and DonHee Lee. 2023. Academic integrity in a digital era: Should the use of chatgpt be banned in schools? *Global Business & Finance Review (GBFR)*, 28(7):1–10.

Aleksandr Ometov, Sergey Bezzateev, Niko Mäkitalo, Sergey Andreev, Tommi Mikkonen, and Yevgeni Koucheryavy. 2018. Multi-factor authentication: A survey. *Cryptography*, 2(1):1.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Alper Şahin, Nathan Thompson, and Kadriye Ercikan. 2024. Opportunities and challenges of ai in educational assessment. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue):260–262.

Hao Yu. 2023. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712.

Benchmarking Finnish Lemmatizers across Historical and Contemporary Texts

Emily Öhman
Waseda University
Tokyo, Japan

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

Abstract

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

1 Introduction

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

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

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

2 Related Work

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

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

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

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

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

3 Data

3.1 Dataset Description

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

3.1.1 Standard Finnish Corpus (499 sentences)

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

3.1.2 Non-Standard Finnish Corpus (189 sentences)

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

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

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

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

3.2 Ground Truth Lemma Annotation

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

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

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

4 Method

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

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

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

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

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

5 Results

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

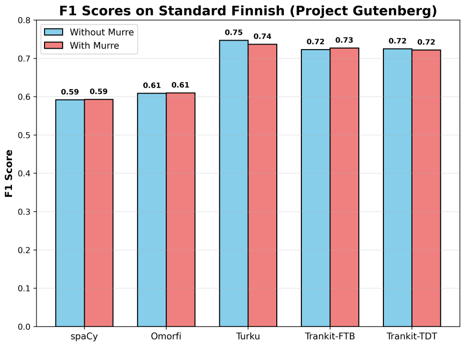


Figure 1: Standard Finnish (n=499) with and without Murre Preprocessing.

5.1 Standard Finnish Results

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

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

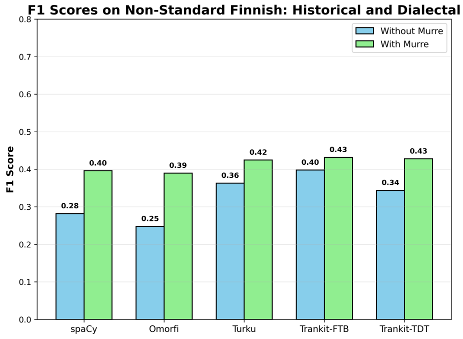


Figure 2: Non-Standard Finnish (n=189) with and without Murre Preprocessing.

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

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

5.2 Non-Standard Finnish Results

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

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

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

6 Analysis & Discussion

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

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

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

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

7 Conclusions

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

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

References

M. Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The world's first South Sámi TTS – a revitalisation effort of an endangered language by reviving a legacy voice

Katri Hiovain-Asikainen[1,2], **Thomas B. Kjærstad**[1,], **Maja Lisa Kappfjell**[1], **Sjur N. Moshagen**[1]
[1]UiT The Arctic University of Norway, [2]University of Helsinki, **Correspondence:** firstname.lastname@uit.no

Abstract

South Sámi (ISO 639: SMA) is a severely endangered language spoken by the South Sámi people in Norway and Sweden. Estimates of the number of speakers vary from 500 to 600. Recent advances in speech technology and the general increase in popularity of spoken language and audio content have facilitated the development of modern speech technology tools also for minority languages, such as the Sámi languages.

The current paper documents the development process of the world's first South Sámi text-to-speech (TTS) system, using only digitized archive materials from 1989–1993 as the training material. To reach an end-user suitable quality of the TTS, we have used a neural, end-to-end approach with a rule-based text processing module. The aim of our project is to contribute to the language revitalization by offering tools for language users to use spoken language in new contexts. Since the modern written standard of South Sámi was established as late as in 1978, the rise of speech technology might encourage language use even for people who are not accustomed to the written standard.

1 Introduction

The traditional speaking area of South Sámi is in the central regions in Norway and Sweden, shown in Figures 1 and 2. The estimated amount of South Sámi speakers range from 500 to 600[1]. Steps are being taken for South Sámi language revitalization in order to preserve and strengthen the speaker community and transfer the language to new generations. This has been most successful in the Snåsa municipality in Norway (see Figure 2), while an increasing number of municipalities are joining the Sámi language management area.

Even though South Sámi is established as a literary language, there are surprisingly few language revitalization initiatives. Some remarkable efforts for revitalizing the languages are currently, for example, translation of children's books, development of South Sámi as a professional language at university levels, as well as using South Sámi as a media language within NRK Sápmi, the Norwegian broadcasting company[2,3]. Despite this, the education system seems to be struggling greatly with the extensive loss of students from Sámi education as shown in the article *The Sámi leak in primary school* (Øystein Vangsnes, 2021).

Because of the long and intense contacts with neighboring Scandinavian languages, practically all adult speakers of South Sámi are bilingual in Sámi and Norwegian/Swedish. Even though South Sámi has had a written standard since 1978, not all speakers have had the opportunity to receive full education in the language and might not be comfortable reading texts in South Sámi. Due to limited environments for hearing and practicing the language, there is a strong need for tools that demonstrate and guide South Sámi pronunciation. Speech technology can help overcome these barriers by providing support for spoken language, as in pronunciation, intonation, and stress. In addition to this, there is a high demand for speech technology tools especially in (special) education, for being able to integrate the learning of spoken language into the language learning materials more easily, as well as aiding people with dyslexia or vision impairment.

The Divvun group at UiT The Arctic University of Norway develops language technology tools for indigenous and minority languages within the GiellaLT infrastructure[4] (Pirinen et al., 2023). The first Sámi TTS for North Sámi was

[1] https://snl.no/sørsamisk

[2] https://www.nord.no/aktuelt/historiske-masterstudentar-i-sorsamisk

[3] https://nynorsksenteret.no/blogg/sorsamisk-sprakutvikling-gjennom-barneboker

[4] https://giellalt.github.io/

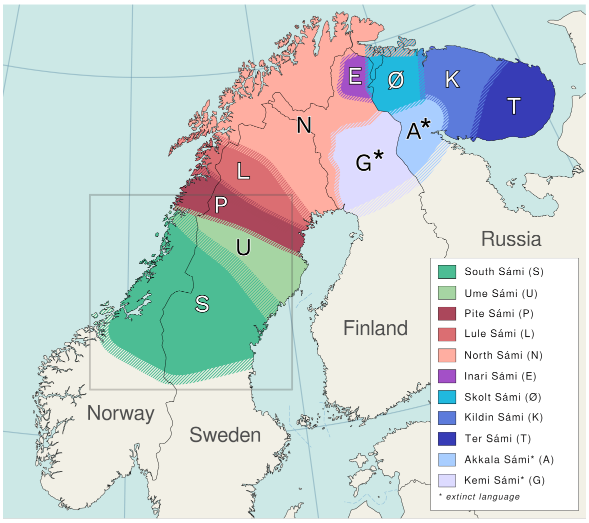


Figure 1: Map of traditional speaking areas of the Sámi languages. Adapted from https://snl.no/samisk with permission from the Great Norwegian Encyclopedia and the map author, Mikkel Berg-Nordlie. The grey box is roughly depicting the closer South Sámi area shown in Figure 2.

created in 2015 in collaboration with Acapela. In recent years, Divvun has updated the North Sámi TTS and released new Lule Sámi TTS tools (Hiovain-Asikainen and Moshagen, 2022; Hiovain-Asikainen and De la Rosa, 2023; Hiovain-Asikainen and Suni, 2025). All Divvun tools are open source, freely available through *Divvun Manager*[5], and integrated into GiellaLT for continued maintenance and updates.

2 Related work

Modern users increasingly expect near-human-quality TTS, as seen in high-resource languages like English. Earlier systems depended on relatively large single-speaker datasets, such as LJ Speech (Ito and Johnson, 2017), but these lacked speaker and linguistic diversity. Recent progress has shifted toward massive multilingual and multi-speaker corpora that improve generalization and adaptability. The Emilia dataset (He et al., 2024), for example, includes 150,000 hours of speech across five major languages, while models like VALL-E 2 (Chen et al., 2024), trained on datasets such as the 50,000 hour LibriHeavy corpus (Kang et al., 2024), push synthesis quality even further. VALL-E 2 claims to reach human parity in naturalness and expressiveness, underscoring the growing role of large, diverse datasets in state-of-the-art TTS research.

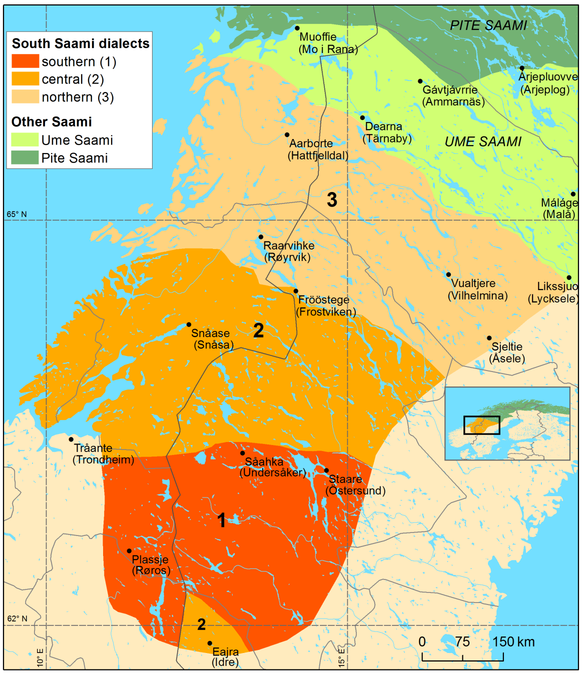


Figure 2: Map of the traditional South Sámi dialects, showing also the most prominent town names and locations. (Rantanen et al., 2022).

While much of the recent TTS work depends on massive datasets, another line of research focuses on models that perform well with minimal data – crucial for languages like South Sámi. Large-scale approaches do not address the core low-resource challenge: the absence of sizable datasets or suitable pre-trained models. For successful adoption in very low-resource settings, synthesized speech must remain both intelligible and pleasant. We therefore selected a non-autoregressive Transformer-based model capable of high-quality output from small datasets ($\leq$10 hours). Prior work supports this choice: Xu et al. (2020) show that 1.3 hours of Lithuanian speech was enough for a Transformer-based TTS system to produce intelligible speech (Li et al., 2019), and in Võro TTS (Rätsep and Fishel, 2023), transfer learning from Estonian did not outperform training directly on 1.5 hours of Võro-only data. These studies indicate that architectures like FastPitch (Łańcucki, 2021) may require as little as 1–2 hours of speech, with limited benefit from related-language data. Because South Sámi differs substantially from other Sámi languages in both linguistics and orthography, transfer learning was not pursued here.

Due to the scarcity of South Sámi data and resources, we utilized existing but non-ideal materials, inspired by Cooper (2019), which examined various speech sources for low-resource TTS. The

[5] https://borealium.org/en/category/voices/

study found that training with a mix of high- and low-quality data, then adapting toward high-quality subsets, improved intelligibility. This approach is particularly useful for low-resource languages, as it enables combining multiple smaller datasets, as done in this work.

In summary, despite recent near-human TTS advances, we selected FastPitch (Łańcucki, 2021) for the South Sámi project due to its low data requirements and straightforward training process. State-of-the-art models like VALL-E offer impressive zero-shot capabilities, but require extensive datasets and substantial computational resources, making them unsuitable for low-resource languages. FastPitch instead provides a transparent, controllable and practical solution to develop the South Sámi TTS, similarly as we have done successfully for Lule and North Sámi before.

3 Methodology

In this section, we describe our data and methodology for developing TTS for South Sámi. We present our approach in acquiring materials and dealing with very low-quality data by performing a series of audio restoration procedures, and by taking the data quality into account in the training phase of the TTS model. Finally, we report our training setup and choice of vocoder specifically for this project.

3.1 Data description

Anna Jacobsen (Sámi name *Jaahkenelkien Aanna*), 30th October 1924 – 2nd April 2004, was a leading advocate for the South Sámi language and culture in Norway (Gaski and Kappfjell, 2006). Born into a reindeer-herding family, Jacobsen grew up speaking South Sámi and later became the first person formally examined in the language. Her work with the South Sámi langugage covered language teaching, translating, writing teaching materials, organizing language groups, being a language consultant for the Sámi Education Council and establishing the Sijti Jarnge Cultural Center and a South Sámi theater. She received honorary awards for her work.

3.1.1 Ethical considerations

Our South Sámi female voice *Aanna* is based on archival recordings of Anna Jacobsen. The recordings are used in this project with the acknowledgment and written consent of her two descendants. The descendants of the original speaker, Anna Jacobsen, were contacted by email and subsequently met through an online meeting to discuss the project and its implications. They gave their informed consent to (1) open-source publication of the recordings and transcripts, (2) their use in the development of a South Sámi TTS system, and (3) waiver of any royalties. The potential risks of voice misuse inherent in TTS technology were discussed, and the descendants expressed general awareness and written support for the project. They do not retain ownership of the resulting models, nor the ability to revoke consent after publication. This work was conducted by the Divvun group that operates as part of the Sámi community under the administration of the Sámi Parliament, and adheres to the FAIR and CARE principles as described in (Moshagen et al., 2024) for data management and ethical research involving Indigenous communities.

3.1.2 The materials and transcribing process

The archival recordings of Anna Jacobsen were sourced from the Norwegian national broadcaster NRK and several audiobooks, and were digitally restored, enhanced, and transcribed by the Divvun group at UiT in collaboration with the Sámi Archives and the National Archives of Norway. Recorded between 1989 and 1993, the material spans multiple genres, including news and documentary broadcasts, biblical readings, fairy tales, and spontaneous autobiographical storytelling.

Developing a TTS dataset required text that accurately matched each recording. Many of Jacobsen's broadcasts already existed in written form, as they were later published in the anthology series Don jih daan bijre I–III (Jacobsen, 1997, 1998, 2000). Her biblical recordings were aligned with the South Sámi translation she produced together with Bierna Leine Bientie (Jacobsen and Bientie, 1993), which was scanned and OCR-processed for the project. Additional usable material came from her language-learning book Goltelidh jih soptsestidh (Jacobsen, 1993) and its accompanying audio cassettes.

Where written texts existed, the recordings were reviewed in detail and the texts were adjusted to reflect the spoken versions, which often differed slightly from the published forms. For recordings without prior transcriptions, full manual transcription was required. Three project members—two native speakers and one highly proficient non-native speaker—carried out this work during 2023–2024.

Processing roughly ten hours of audio resulted in about one hundred hours of transcription, equivalent to 2.5–3 weeks of full-time work for one person. This workload is realistic for most endangered language contexts, especially since a significant portion of the material could be aligned rather than transcribed from scratch. Although labor-intensive, manual transcription remains more feasible than automated methods (such as ASR), which require large, high-quality datasets that Indigenous languages typically lack. Human transcription is therefore not a bottleneck but a practical and scalable strategy for building high-quality TTS resources in low-resource settings.

During transcription, segments with very poor audio quality or containing speakers other than Anna Jacobsen were excluded from the final TTS corpus.

3.2 Data processing

After the transcribing process, all texts were once more proof-read and all audio was cleaned of any unusable parts or noise. Then, the material was force-aligned to automatically find sentence boundaries from the audio, using a WebMAUS[6] pipeline without ASR (G2P ⇒ MAUS ⇒ Subtitle, see Kisler et al. (2017); Schiel (1999)), retaining the original text formatting and punctuation. There are no Sámi models on WebMAUS, so we used their Finnish (related language) model. The automatically aligned sentence timestamps from WebMAUS were then manually checked and used to split the data into sentences. Python scripts by the first author, utilizing the TextGridTools toolkit[7] (Buschmeier and Wlodarczak, 2013) were used to save each sentence to an individual sound file with a corresponding text transcript. After splitting the data, the net duration of the entire dataset was 10.5 hours, with 4670 individual sentences in total.

The next step in our pre-processing pipeline was to enhance the audio. Understandably, as an archive material, the audio quality of our material was not as high as normally expected from any generic text-to-speech projects. Our material was collected from different cassettes and CDs, all with varying recording conditions and probably with different digitizing equipment as well. The sound files were enhanced and de-noised using the freely available *Resemble-enhance*[8] with default settings and parameters. Resemble Enhance is an AI-powered tool that aims to improve the overall quality of speech by performing denoising and enhancement. It consists of two modules: a de-noiser, which separates speech from noisy audio, and an enhancer, which further boosts the perceptual audio quality by restoring audio distortions and extending the audio bandwidth. Running the denoiser and enhancer through the entire dataset substantially improved the audio quality. Next, we used a shell script utilizing *sox* and *svdemo* libraries to level normalize the data. Finally, the whole dataset was resampled at 22.05 kHz to be compatible with our TTS training setup.

3.3 Model configuration

Our model was trained using the FastPitch (Łańcucki, 2021) architecture with explicit duration and pitch prediction components. For our final model, we used a “multi-speaker” configuration for training by splitting the data into two subsets based on audio quality. Even though we performed audio enhancement to the entire dataset, the lower quality partition remained lower quality compared to the better quality part even though it was substantially improved in intelligibility compared to the original quality.

Out of the total 10.5 hours of data, 2 hours were manually labeled as “good quality” with speaker ID *1*, while the remaining 8.5 hours were labeled as lower quality with speaker ID *0*. This binary labeling reflects the intended use for TTS generation: ID *1* denotes recordings suitable for synthesis, whereas ID *0* denotes recordings that, while usable for training, were not intended for generation. We did not define additional quality categories or employ a continuous quality metric, as the primary goal was to distinguish between data that could or could not be directly used for synthesis.

The material was then shuffled in order and further divided into training, validation and test sets with a split of 4570/85/15, respectively. The test set (see Appendix B) was later used for an evaluation protocol. The dataset was processed using the standard FastPitch data preparation scripts to extract pitch, duration, and mel spectrograms from each utterance.

After defining the orthographic symbol set (see Appendix A) for South Sámi, the model was trained for 830 epochs and altogether 14K steps on the Saga supercomputer[9] at the Norwegian computing

[6] `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Pipeline`

[7] `https://textgridtools.readthedocs.io/en/stable/index.html`

[8] `https://github.com/resemble-ai/resemble-enhance`

[9] `https://documentation.sigma2.no/hpc_machines/saga.html`

cluster *Sigma2*. We allocated 16 GB of GPU memory at Saga for the training process, and used an effective batch size of 256 and learning rate 0.1 as training parameters.

3.4 Model inference and text processing

To achieve good and consistent quality in the synthesis output, the speaker variant for our synthesis generation script was set to speaker ID *1*, acoustically similar to the better quality subset of our dataset. For inference, we used the UnivNet model (Jang et al., 2021) from NeMo collections as the vocoder. This vocoder was chosen because it produced an audibly better quality inference output than other neural vocoders, such as the widely used HiFi-GAN (Kong et al., 2020), compared with UnivNet in Jang et al. (2021). The most important reason for better quality output for UnivNet seems to be improved generalization across diverse speakers, domains, and unseen data. This is achieved through multi-resolution spectrogram loss by comparing synthesized audio to the target audio at multiple resolutions. As our TTS model was meant for publication in a TTS application, it was also important to choose a vocoder optimized for low-latency inference, such as UnivNet, making it suitable for real-time applications.

The text processing is done using the existing tools in the GiellaLT infrastructure. That infrastructure is based on HFST (see, e.g. Lindén et al. (2013)) and VislCG3 (see e.g. Karlsson (1990), Didriksen (2010)), making it possible to build advanced text processing tools for languages that lack large corpora. In the pipeline, the raw text is tokenized and morphologically analyzed by a lexical transducer, followed by several steps to enhance and disambiguate the tokenisation. The lexical transducer has many lexicalized compounds while at the same time being able to handle dynamic compounding. In the case of TTS, we prefer the dynamic compounds, because it allows us to normalize each part of the compound independently, for example in a word like "CD-player". We also add valency information to help with the morphological disambiguation, which is coming next. The disambiguated analysis is then used as input for normalizing digits and abbreviations, and since these tokens are already disambiguated, only the relevant normalization is applied (if the normalization itself is ambiguous, it is possible to add further disambiguation at later steps). The normalized text is finally sent to the synthesizer. A flow diagram of the whole process is shown in Figure 3.

As illustrated in Figure 3, the text-processing pipeline consists of a multi-stage linguistic analysis and normalization workflow. The example sentence:

```
<Joekoen guhkiem, jis edtjebe jaehkedh
dam 35 jaepien båeries
nyjsenæjjam Kloemegistie.>
```

(*"A very long time, if we're to believe the 35 year old woman from Glåmos."*)

is first segmented and tokenized into lexical units:

```
<Joekoen> <guhkiem> <,> <jis>
<edtjebe> <jaehkedh> <dam>
<35> <jaepien> <båeries>
<nyjsenæjjam> <Kloemegistie> <.>
```

Each token is then processed by finite-state–based morphological analyzers, which produce lemmas and feature bundles (e.g., part of speech, case, number) and identify compound structures and orthographic variants.

Morphological disambiguation follows, using constraint grammar rules to select context-appropriate readings and assign syntactic functions. In the normalization stage, compounds are reconstructed, numerals are lexicalized (e.g., 35 → *golmeluhkievïjhte*), and canonical lemma forms are prepared for TTS input. The resulting output is mapped to phonological representation elements, yielding the normalized sequence:

```
<joekoen guhkiem, jis edtjebe
jaehkedh dam golmeluhkievïjhte
jaepien båeries
nyjsenæjjam kloemegistie>
```

which is then passed to the TTS model for generation.

It is possible to add a fall-back normalizer for handling unknown words, and there are stubs in place to differentiate the normalization of loan words from that of native words. For example, the orthographic "y" is typically pronounced differently in various South Sámi words, depending on its phonetic context (Magga and Magga, 2012). The present synthesis model has just learned the various pronunciations based on context in the training phase, and this works quite well. However, it still makes errors, and by differentiating the various pronunciations already in the text processing (and similarly in the training material), we should be able to get an even more correct pronunciation.

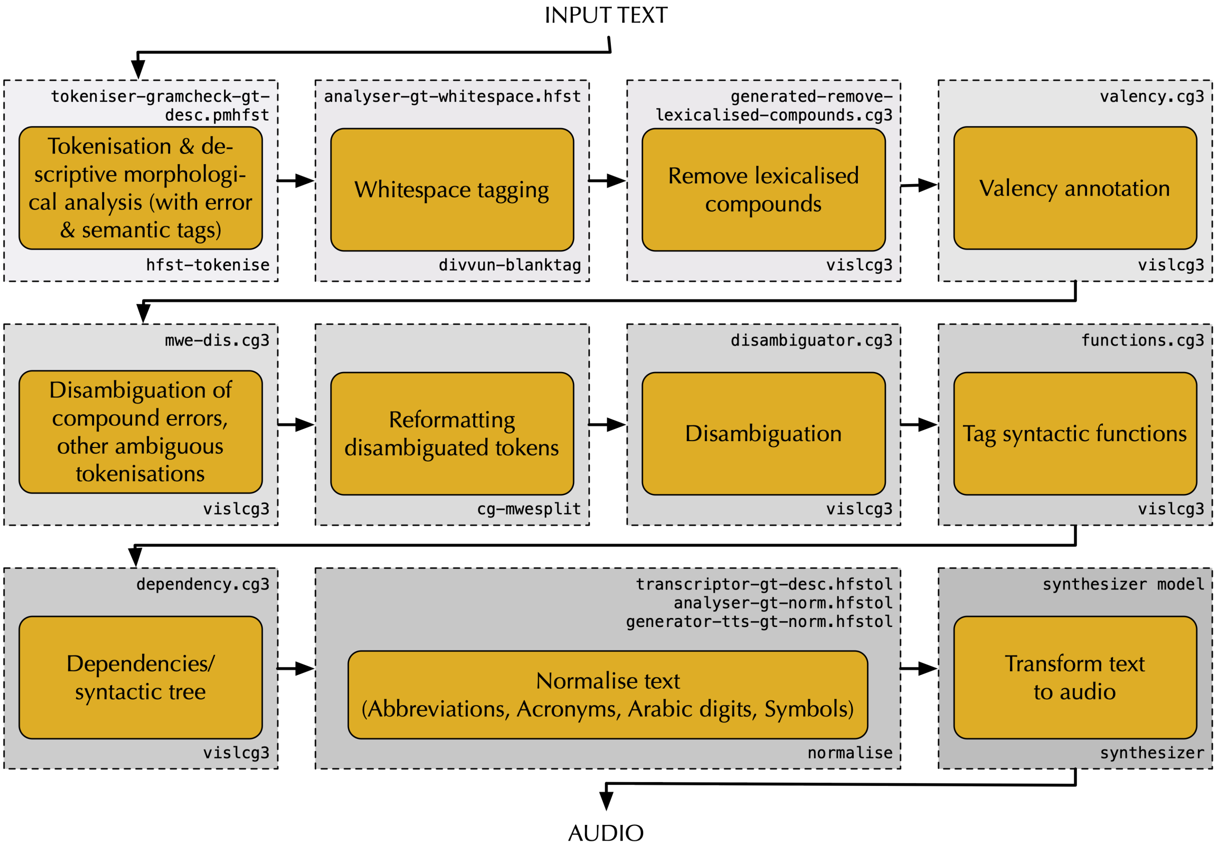

INPUT TEXT
tokeniser-gramcheck-gt-desc.pmhfst
Tokenisation & descriptive morphological analysis (with error & semantic tags)
hfst-tokenise
analyser-gt-whitespace.hfst
Whitespace tagging
divvun-blanktag
generated-remove-lexicalised-compounds.cg3
Remove lexicalised compounds
vislcg3
valency.cg3
Valency annotation
vislcg3
mwe-dis.cg3
Disambiguation of compound errors, other ambiguous tokenisations
vislcg3
Reformatting disambiguated tokens
cg-mwesplit
disambiguator.cg3
Disambiguation
vislcg3
functions.cg3
Tag syntactic functions
vislcg3
dependency.cg3
Dependencies/ syntactic tree
vislcg3
transcriptor-gt-desc.hfstol
analyser-gt-norm.hfstol
generator-tts-gt-norm.hfstol
Normalise text
(Abbreviations, Acronyms, Arabic digits, Symbols)
normalise
synthesizer model
Transform text to audio
synthesizer
AUDIO

Figure 3: Flow diagram of the South Sámi TTS runtime system, including all text processing steps.

4 Results and Future Work

4.1 Technical implementation and publishing the TTS system

The resulting TTS model was compiled into TorchScript form, using the standard script from the FastPitch repository. This was done to create serializable and optimizable models from the originally trained PyTorch model. The JIT compiled model was further integrated into macOS and Windows operating systems (OSs), at the OS level, such that the voices appear as system voices within the limitations of the operating systems. In both systems, voices can be activated as screen-reader voices. In MacOS, the South Sámi voice can also be used within LibreOffice [10]. To some extent, voices can also be used for simpler read-aloud functionality for highlighted text, but there are several restrictions in both systems that limit the usability of this functionality. The OS integration is language independent and can include any voice for any language we release in the future. We plan to release support for using voices on Android and iOS in the near future.

The text processing infrastructure is both flexible and powerful, and allows us to do further syntactic and semantic analysis. This can also be used to add synthesis markup to guide, e. g., prosodic features in the synthesis, but this will have to be the target of future work. The modified FastPitch code used to create our South Sámi voice will be published on our GitHub page https://github.com/giellalt/speech-sma once the detailed documentation of the project is completed.

4.2 Initial evaluation of the TTS model

Established evaluation standards in speech technology are designed for majority languages with large speaker populations, and they cannot be readily applied to Indigenous and minority languages such as South Sámi. With few speakers, standard procedures become impractical and risk placing an undue burden on the community. This underscores a broader issue: evaluation in low-resource contexts must center community needs and capacities. While cross-linguistic comparability is valuable, methodological ideals may conflict with ethical and practical realities. In such cases, community well-being should take precedence, and transparent reporting of necessary deviations from standard

[10] https://www.libreoffice.org/get-help/install-howto/macos/

practice should be appropriate.

Research fatigue further complicates evaluation, as Sámi speakers are frequently asked to participate in studies despite the small population. Nonetheless, attempting evaluation remains important, and development of South Sámi TTS offers tangible benefits that help counterbalance concerns about extractive research. After the TTS release on 30 October 2024, an anonymous feedback survey was launched in October 2025, with initial responses indicating continued community engagement despite these constraints.

For initial testing of the model and the TTS inference, we created a set of 15 sentences, taken aside from the training data (see Appendix B), which we synthesized using the resulting model. We prepared corresponding 15 sentences both from the test set (ground-truth) and synthesized samples. After level-normalizing this evaluation set of altogether 30 samples, we created an online survey using Microsoft Forms with embedded audio files. For evaluation, we used a 5-point Likert scale (from 1 – Bad to 5 – Excellent) and asked the following questions: 1) How would you rate the general pronunciation and rhythm of the speech (**Pronunciation & Rhythm**) 2) How would you rate the pleasantness of the speech? (**Pleasantness**) 3) How would you rate the clarity of the speech (i. e. how easy it is to understand what is being said) (**Clarity**)?

Out of 15 test sentences, 12[11] paired audio items were used for the Wilcoxon signed-rank tests, with each pair consisting of a synthesized utterance and its corresponding ground-truth recording. For each item, the mean rating across evaluators was used in the comparison. The results show a statistically significant difference only for **Pleasantness**, where ground-truth recordings received higher ratings than the synthesized speech ($p \approx 0.0039$). For Clarity and Pronunciation & Rhythm, no significant differences were found between ground-truth and synthesized audio ($p \approx 0.33$ and 0.52, respectively).

In addition to the quantitative, statistical analysis, we asked native speakers to assess the quality of the South Sámi TTS in free form, providing us more qualitative, also valuable feedback on our synthetic voice. One native South Sámi speaker (outside of the authors of this paper) listened to these samples and commented any pronunciation mistakes or peculiarities in the prosody.

The evaluator provided three main comments: (1) Year numbers are sometimes omitted or read digit-by-digit on first attempt; more natural segmentation and rendering is recommended (e.g., *luhkiegaektsie-gaektsieluhkietjïjhtje* for '1887'). (2) Vowel duration could be slightly lengthened in some words, and a marginally slower speaking rate may improve naturalness and intelligibility. (3) Compound words are pronounced more accurately when written with hyphens.

These issues affect both text processing and, to a lesser degree, the speech synthesis model. Improved handling of numeric expressions, a modest reduction in speaking rate, and automatic insertion of hyphens for compounds should be straightforward adjustments. After implementing any changes, another evaluation—ideally with more participants—should be conducted to obtain more robust feedback.

4.3 First impressions and the expected impact of TTS in the South Sámi community

Overall, the development and integration of text-to-speech (TTS) technology for the South Sámi language has initially received positive reception from the community, with some media coverage as well[12]. TTS addresses a long-standing need for accessible language resources, particularly for self-study and language revitalization, by allowing South Sámi speakers to input text and hear it read aloud. This is particularly important for minority languages that lack such tools. The community expects TTS to support everyday language use, for example integration into smart home devices and public services could offer practical language support and increase accessibility for South Sámi speakers, also in Sámi administrative municipalities. TTS could also contribute to proper Universal Design implementations, making public information more inclusive.

In terms of specific use cases, the integration of TTS in digital dictionaries and language learning apps is highly anticipated. For instance, a young teacher has expressed interest in incorporating TTS into the popular flashcard program Anki[13]. This would allow learners to hear South Sámi words pronounced aloud while studying, combining visual and auditory learning modes to reinforce vocab-

[11] some evaluators did not evaluate all 15 sentences

[12] https://uit.no/nyheter/artikkel?p_document_id=864438

[13] https://apps.ankiweb.net/

ulary acquisition. This kind of integration could be a significant step forward in making language learning more efficient and accessible to a wider audience. Johan Sandberg McGuinne, a member of the community, also highlighted the importance of TTS as a teaching aid, stating, "I think it's good because I can use it as an aid in teaching and in elderly care." This sentiment underscores the diverse applications of TTS in enhancing both education and daily life for South Sámi speakers.

The integration of TTS technology into existing systems is thus expected to play a key role in the revitalization and sustainability of the South Sámi language, fostering a more inclusive linguistic environment for the community.

4.4 Future work

Our work in general includes strategies for ongoing improvement of the system as language usage evolves. The speech synthesis system separates text processing from the actual synthesis step such that the text processing is done by the existing, rule-based text processing components in the GiellaLT infrastructure, and the resulting plain-text strings are fed to the synthesis engine. The speech model in the synthesis engine does not need to be retrained or rebuilt for the whole system to improve - it is enough that improvements are made to the text processing pipeline, e.g., to improve handling of cases in numerals, or new names, words and terminology. As the text processing improves, so will the resulting generated speech. And as the source of the text processing is shared with all other tools built within the GiellaLT infrastructure, improvements in one area will automatically also improve speech synthesis.

Another aspect is that the team behind the Sámi speech synthesis projects are fully funded by the Norwegian Government, as part of long-term commitments to supporting the Sámi languages. So both financially and practically our work will continue for years and decades — we have already been doing this for twenty years — and thus maintenance and commitment to updates should be covered for the foreseeable future.

Our future work on South Sámi TTS technology involves expanding its capabilities by adding more voices, such as a male voice, and incorporating additional dialects/areal varieties of South Sámi. There is also potential to develop multilingual solutions, integrating Swedish and Norwegian material to the model for proper loan word pronunciation and to allow for code-switching in the TTS output. Augmenting the training dataset with recordings from additional native speakers would also help capture rarer words, exceptional pronunciations, loanwords, and names, further enriching the TTS system.

We also hope that in the future, our present work could inspire the revitalization of the smallest and most endangered Sámi languages like Ume and Pite Sámi, both of which have very few speakers. In the case of Ume Sámi, a significant collection of unpublished written materials exists but remains unavailable (Siegl, 2017). New language and speech technology tools could help make these resources more accessible and support the revitalization efforts. There is rising interest to still revitalize these languages, and speech technology could play an important role.

5 Conclusions

In this paper we presented a description of a novel TTS project, utilizing non-ideal, archived and digitized materials. We show that end-user suitable TTS quality is possible with limited materials and even with initially low quality audio. We suggest a way to process archive materials for TTS in an effective pipeline that is generalizable to other very low-resource languages as well. The resulting TTS voice, built from the archive audio materials by Anna Jacobsen has gotten very positive feedback from the present community, encouraging use of the language in new, spoken language contexts, contributing to the revitalization of the severely endangered language.

6 Acknowledgements

We thank Antti Suni and Sebastién Le Maguer from the University of Helsinki for their valuable advice. Our deepest gratitude goes to Ina Therés Andrea Sparrock for her exceptional contributions, particularly in transcribing the audio materials and presenting the project to the South Sámi community. We also warmly thank Halvard and Nils Johan Jacobsen, sons of Anna Jacobsen, for their cooperation and encouragement. Finally, we are grateful to Sijti Jarnge for hosting the launch of the world's first South Sámi TTS in Aarborte–Hattfjelldal during the celebration of Anna Jacobsen's 100th anniversary in October 2024.

References

Hendrik Buschmeier and Marcin Wlodarczak. 2013. Textgridtools: A textgrid processing and analysis toolkit for python. In *Tagungsband der 24. Konferenz zur elektronischen sprachsignalverarbeitung (ESSV 2013)*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Erica Cooper. 2019. *Text-to-speech synthesis using found data for low-resource languages*. Columbia University.

Tino Didriksen. 2010. *Constraint Grammar Manual: 3rd version of the CG formalism variant*. GrammarSoft ApS, Denmark.

Harald Gaski and Lena Kappfjell. 2006. Saemien tjaelijh – 16 saemien tjiehpies- jïh faagelidteratuvren tjaelijh.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Katri Hiovain-Asikainen and Javier De la Rosa. 2023. Developing tts and asr for lule and north sámi languages. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 48–52.

Katri Hiovain-Asikainen and Sjur Moshagen. 2022. Building open-source speech technology for low-resource minority languages with sámi as an example–tools, methods and experiments. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 169–175.

Katri Hiovain-Asikainen and Antti Suni. 2025. Does multilingual and multi-speaker modeling improve low-resource tts? experiments on sámi languages. In *Proc. SSW 2025*, pages 196–201.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. 2017. *URL https://keithito. com/LJ-Speech-Dataset*.

Anna Jacobsen. 1993. *Goltelidh jih soptsestidh*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 1997. *Don jih daan bijre I*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 1998. *Don jih daan bijre II*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen. 2000. *Don jih daan bijre III*. Aarbortesne, Norway: Daasta berteme.

Anna Jacobsen and Bierna Leine Bientie. 1993. *Jupmelen rijhke lea gietskesne : Maarhkosen vaentjele*. Det Norske Bibelselskap.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Jinhan Kong, Jaehyeon Kim, and Jungil Bae. 2020. Hifigan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. Hfst—a system for creating nlp tools. In *International workshop on systems and frameworks for computational morphology*, pages 53–71. Springer.

Ole Henrik Magga and Laila Mattsson Magga. 2012. *Sørsamisk grammatikk*. Davvi Girji.

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

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.

Timo Rantanen, Harri Tolvanen, Meeli Roose, Jussi Ylikoski, and Outi Vesakoski. 2022. Best practices for spatial language data harmonization, sharing and map creation—a case study of uralic. *Plos one*, 17(6):e0269648.

Liisa Rätsep and Mark Fishel. 2023. Neural text-to-speech synthesis for võro. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727.

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proc. of the ICPhS*, pages 607–610.

Florian Siegl. 2017. Ume saami – the forgotten language. *Études finno-ougriennes*, (48).

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Øystein Vangsnes. 2021. Samiske tall forteller: Den samiske lekkasjen i grunnskolen.

A Appendix

South Sámi symbol set used in the FastPitch model training:

letters = '*AÆÅBCDEFGHIÏJKLMNOØÖPQRSTUVWXYZaæåbcdefghiïjklmnoøöpqrstuvwxyz*'

B Appendix

The set of 15 sentences for the evaluation test.

1. Nöörjen raedteste jis, Praansvaerien Novra jïh Jaahkenelkien Aanna.
2. Nåå jïh ibie leah gujht mijjieh daebpene åarjene annje man væjkele joejkedh, jïlhts aaj Frode lea mijjen båeries vuelide giehtjedamme, golteladteme jïh orrestahteme guktie nuerebh almetjh utnieh sijjide aaj sjeahta mijjen båeries vueliej mietie svihtjedh jïh aaj raakte joekestalledh.
3. Goh Jeesuse jïh dah luhkiegööktesh, jïh dah jeatjebh gïeh Jeesusinie ektesne, lin oktegh sjïdteme, dellie gihtjin dan jortesen bïjre.
4. Jïjtjh tjoeveribie jïermestalledh gåabph libie rahtjeminie dejnie mijjen Saemiedigkiebarkojne.
5. Pængstah leah aah orreme.
6. Nyjsenæjja dåerkeste juktie bælla, daajra guktie satnine sjïdti, jïh dellie båata Jeesusen uvte slienghkehte jïh gaajhkem såårne.
7. Daan mijjen eatnemen lea stoerre aalkoealmetjefuelhkie, jïh mijjieh saemieh aaj dan fualhkan govlesovvibie.
8. Åadtjibie vaajtelidh dïhte maahta bueriedidh.
9. Eadtjohkelaakan kultuvrine barkedh.
10. Dïhte prievie båata daehtie moenehtseste maam Saemiedigkie lea tseegkeme, jïh mij edtja nænnoestimmie buektedh.
11. Såemies gaertenebuerie gujht vienth daan jaepien aaj tjoevere sïrvide joekoenlaakan bïepmedh guktie begkerelle-låhkoe vaanen aerebi goh åadtjoeh leekedidh.
12. Tjaktje seenhte, naa jueskie lij gujht.
13. Nimhtie lea daan eatnemen mubpene bealesne, Orre Zeelantesne, Maorij luvhtie luvnie.
14. Guktie idtjidh dellie Jååhannesem jaehkieh?
15. Dïhte ovmurreds saernie noerhtede Saeltievaerien luvhtie båata.

Can advances in NLP lead to worse results for Uralic languages and how can we fight back? Experiences from the world of automatic spell-checking and correction for Finnish

Flammie A Pirinen
Divvun
UiT—Norgga árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Abstract

Spell-checking and correction is a ubiquitous application within text input in modern technology, and in some ways or another, if you type texts on a keyboard or a mobile phone, there will probably be an underlying spelling corrector running. The spell checkers have been around for decades, initially based on dictionaries and grammar rules, nowadays increasingly based on statistical data or large language models. In recent years, however, there has been a growing concern about the quality of these modern spell-checkers. In this article, we show that the spell-checkers for Finnish have gotten significantly worse in their modern implementations compared to their traditional knowledge-driven versions. We propose that this can have critical consequences for the quality of texts produced, as well as literacy overall. We furthermore speculate if it would be possible to get spell-checking and correction back on track for Uralic languages in modern systems.

1 Introduction

Spell-checking and correction is a quintessential *natural language processing* (NLP) task. It has been part of the NLP ecosystem for decades now, from the very early days of processing texts with computers. It has become so ubiquitous that it exists in most text editing products without users even paying much attention to it, and it has been viewed as somewhat of a solved problem within the scientific study for the last few decades. While there has not been much focus on spell-checking and correction in recent years, we as linguists have noticed something quite problematic with the contemporary systems. Namely, we have noticed a drop in quality of writing in our native languages on the Internet discussion forums. This is increasingly shown in the frustrations by the native writers: "autocorrect wrote it, and it is too hard to fix it by hand". On this basis, we set out to study, if the contemporary spell-checking and correction systems have become worse in our language. Our hypothesis is, that modern autocorrecting spell-checking and correction systems are based on data-driven methods and lately large language models, which may work adequately with English—not the least because over 90 % of the training data is in English[1]—but which actually fail to recognise words of non-English languages with potentially more complicated morphology.

Our *research question* in this short paper is, are data-driven and large language model based spelling checkers and correctors worse than traditional knowledge-based ones? Our initial hypothesis, based on everyday observations, is that spell-checking tools have gotten significantly worse in the past few decades, in pace with the introduction of data-driven and 'AI'-driven models. We study the spell-checking and correction results by three popular systems for *Finnish*.

2 Background

There is a long history of spell-checking and correction in language technology, starting from early days of SPELL, a spell-checker based on a dictionary or a word-list and few simple rules to modify suffixes. Earnest (1976) places initial use of their spelling correction to 1969. This system's descendants—ispell, aspell and hunspell and so forth—have been in use in some of the most popular browsers and office suites up to the 2000s. There have been several comprehensive scientific surveys of spell-checking and correction, for example Kukich (1992). As of last few decades, office suites have started using built-in, closed-source, statistical spell-checkers and more recently, overarching AI assistants which also do spell-checking,

[1] c.f. e.g. https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 22–28
December 10-12, 2025 ©2025 Association for Computational Linguistics

a similar development is happening in browsers, mobile phones and operating systems. One of the most influential initial works on data-driven spell-checking and correction is Google's Norvig's spelling corrector (Norvig, 2009). Technical details of the most modern commercial spelling correctors are not openly documented as far as we know.

When researching on existing studies on LLM-based spelling correctors and especially comparisons between LLM and traditional methods, one topic that dominates the results is spell-checking and correction for students / L2 / EFL users of English (Jaashan and Alashabi, 2025; Gayed et al., 2022). A gap in research we hope to address with this experiment and its followups, is, therefore, a comparative study, and for L1 users, in non-English.

In our work, we build NLP tools and software, mainly targeting less-resourced, minority and Indigenous languages, but we also create tools that are language agnostic and usable for all. Spell-checking and correction and related software is a key tool for digital language survival for minority languages, and it is also increasingly important for ever larger and more majority languages, apart from the largest few. This has come to be also, because the contemporary data-driven language technology is strongly based on big data, that has been written by humans in correctly spelled and grammatical language.

In this experiment, we have chosen to use Finnish. Finnish is a national, majority language in Finland, it is not low-resource by any stretch of the imagination. We estimate it is likely to be in the top 50 of the most resourceful languages in the world. While we are more interested in low-resource languages and settings, having moderately resourced Uralic language works well for our initial experimentation. We have existing resources such as corpora and established automatic spell-checkers, which we might not find on lower resourced languages. There is also existing research on state of the Finnish NLP (Hämäläinen and Alnajjar, 2021) including spell-checking and correction. Furthermore, Finnish is not an Indo-European language, and has a slightly more complicated morphology than most IE languages, which makes it more comparable towards many of the minority and under-resourced languages relevant to our research. Finally, we have native speakers of Finnish, which in our opinion is critical in doing meaningful qualitative studies on language technology software; without linguistic error analysis and human interpretation of the results, it is impossible to make meaningful explanation of how useful or harmful the underlying system is for actual end users.

Linguistically, Finnish is a Uralic language with some 5 million speakers, mainly in Finland. Morphologically, Finnish has what we call slightly more complex morphology, in terms of what matters for spell-checking and correction this means that there are on average thousands of word-forms per word, instead of around 5 like in English or few dozens like in most IE languages. Finnish also has productive compounding, which means you can put two word-forms together without a space to create a new word, that does not necessarily exist in the dictionary, on the fly. Finnish has had a literary culture for several hundreds of years and has a strong nationally backed standardisation body, is a primary language in schools and in public. It is also a majority language for several Indigenous and minority languages, which is one of the motivations for us to work on it as well.

3 Methods and experimental setup

In this work, we compare and contrast spell-checking and correction from the end-user point of view. We test three different systems: one based on knowledge-driven paradigm and two based on data-driven approach. The knowledge-driven spell-checker is an open source, rule-based product, whereas the data-driven products we experiment with are commercial and closed-source.

The rule-based spell-checking and correction is a freely available open source implementation of Finnish spell-checking found on the GitHub called omorfi[2], their implementation is based on finite-state spell-checking (Pirinen and Lindén, 2014). This spell-checker uses an underlying dictionary and morphological rules to recognise valid word-forms without context, and uses finite-state error modelling technology to create suggestions for corrections.

We use Google's spell-checking and correction as a black-box, we have not found technical documentation detailing it, but we estimate that it is at least in part based on statistical methods and or large language models based on the company's recent focuses and public statements.[3] The

[2] https://github.com/flammie/omorfi/

[3] Searching online leads to old posts like: https://work

function in Google Docs interface is found under spelling and grammar checking, we have crossed off grammar-checking and only included spelling.

For a product that is certainly using large language models, we test ChatGPT (OpenAI, 2025), and we use it as a black box with the version available to us via our university account. We use the web-based ChatGPT user interface to query spelling corrections from the language model via its natural language user interface in the same way an average end user likely would.

The experiments have been performed in May 2025, some details are included in the appendix A, but since they are closed commercial products, we do not expect to be able to have reproducible results with them in any case.

4 Data

To test the spell-checking and correction we have used a Finnish translation of *Alice's Adventures in Wonderland* from Project Gutenberg[4] which is in public domain. This book is a fantasy novel aimed for children, and contains creative use of language which makes it very suitable for natural language processing testing. The translation has been made in early 20th century which matches the most modern standard written Finnish with almost no deviations. In general, proofreading at the times of the publication was highly valued and efficient, and we expect the manuscript to be mostly error-free barring potential mistakes in gutenberg's encoding. The non-word errors we have found and verified are listed in the error analysis section 5.1. The book consists of 18,861 space-separated tokens (after removing project Gutenberg's licence, preamble and postamble).

5 Results

To measure the spelling error correctors, we went through all the words that were flagged as spelling errors, and categorised them into two categories: *false positives*, where a correctly spelled word was flagged as incorrect, and *true positives*, where the flagged word did contain a spelling error. This was done by a native speaker who had access to the error in context, even though the decision was made solely on whether the word is a valid word in the language at all or not (i.e. it can also be decided without context as traditional non-word spelling corrector does). The breakdown of errors and flaggings is shown in the Table **??**, we also provide a calculations of precision, recall and $F_{0.5}$, the parametre 0.5 for β is selected since our starting point is that false positives are more critical problem in spell-checking than false negatives.

Error \ System	**Google**	**ChatGPT***	**omorfi**
False Positive	565	75	59
False Negative	22	59	20
True Positive	41	4	43
Precision	0.07	0.05	0.42
Recall	0.65	0.06	0.68
F-Score ($F_{0.5}$)	0.08	0.05	**0.46**

Table 1: Quantitative evaluation of error types by systems. * ChatGPT results are not proportional due to reasons explained in the chapter. For main findings, read the qualitative error analysis.

5.1 Error analysis

We have further categorised the errors flagged by the spelling correctors into error types, based on linguistic insight and world knowledge. We hypothesise this will help give an impression of the impact these errors have on the user experience, this impact is further discussed in the section 6 below. The summary of errors is given in table 2, some of the error classes are not mutually exclusive and the numbers in the rows do not add up to the total.

One of the largest groups of false positives in all systems' data is compound words, particularly the types that do not appear in dictionary: for Google's spell-checking **compound** nouns like *pääkallonkuva* (picture of a skull) or *kyynellammikko* (lake of tears) were consistently underlined, for ChatGPT we have e.g. *herttakuningatar* (queen of hearts) and for omorfi we saw compound adverbs like *tuulennopeasti* (in wind's speed). From **derivational** forms, all systems stumbled on *ruukkusen* (little jar's∼jarful$^?$). Some of the false positives found by Google can also be described as being part of complex morphology that

space.google.com/blog/productivity-collaboration/everyday-ai-beyond-spell-check-how-google-docs-is-smart-enough-to-correct-grammar, but we cannot know for sure if this kind of information is up-to-date.

[4] https://www.gutenberg.org/ebooks/46569 for reproducibility we have the version we used in our GitHub at https://github.com/flammie/purplemonkeydishwasher/tree/master/2025-iwclul/reprodata; this is also for access from within Germany, Italy or other countries with extreme copyright restrictions where Project Gutenberg may not be available.

is a bit half-ways between *inflectional* and derivational morphology, for example *myöhästynkin* (I will be late too), *elämäniässään* (in their lifetime), *vaikeroidessaan* (while they were whining), that is, enclitic particles, possessive suffixes and non-finite verb forms in combinations that—in all likelihood have not been many times in sufficiently large corpora—throw Google's spelling checker off the track. The commonality for errors in this category is that there are at least two distinct inflectional suffixes in the word-form. Perhaps surprisingly, also **proper nouns** show up as false positives, even though traditionally maybe it has been common practice to ignore titlecased words: Google finds *Ellakaan* (Ella too) and *Vilhelmiä* (of Vilhelm) errors, and omorfi finds *Morcar* and *Stigand*. The classes as laid out in the table 2 are not mutually exclusive, i.e. a **compound** form can also have a **derivation** and a **proper noun** can have a **inflectional** possessive suffix, in these cases we have simply counted the error in both classes. To illustrate the overlapping between categories, for example *Irvikissakaan* (Cheshire cat neither, lit. grinning-y cat) is a proper noun compound with inflectional ending. There are handful of words that do not seem to fall into any categories; for omorfi we can simply note they are missing from the dictionary, e.g. *satakaunoja* (an old word for some flower) or *siekailuun* (into scrupulousness) whereas with data-driven models we can assume the words themselves are so rare that they do not show up enough in the training materials, e.g. *pulppusivat* (bubbled up) or *pulikoinut* (drudged about), but there are some that are even harder to diagnose, such as *nurmen* (grass') and *vai* (or).

Error \ System	**Google**	**ChatGPT**	**omorfi**
Compound	169	38	18
Derivation	21	9	7
Inflection	211	6	4
Proper noun	12	0	8
Other	171	24	32
Total	611	70⋆	57

Table 2: Error analysis of false positives in Alice in Wonderland by three systems. Classes are not mutually exclusive and may not add up to totals per column. ⋆ChatGPT started to give empty answers and repeat from the beginning after 70 spelling errors.

The true positives in the text fall into following categories: unexpected hyphenation caused by creative language use (recreation of typeset poems: *tar-kemmin* (tarkemmin), *päi-villä* (päivillä), and *veruk-keella* (verukkeella)), lengthening of letters for emphasis (*li-iemi* (liemi), *ku-ulta* (kulta) and *ihana-ainen* (ihanainen)), foreign words (*Oú*, *est*, and *chatte*), dialectal, informal or poetic forms (*teälhän* (täällähän), *käshän* (käsihän), *näkkyy* (näkyy), *käs* (käsi), *sittennii* (sitenkin), *pyssyy* (pyssyä), *ruppee* (rupeaa), *pentus* (pentusi), *juur* (juuri), *loitoll'* (loitolla), *täss'* (tässä), *kuus* (kuusi), *tavaraks* (tavaraksi), *niill'* (niillä), and *tuoss*), compounding mistakes (*mitenpäin* (miten päin), *missäpäin* (missä päin), *käsikädessä* (käsi kädessä), *sukkajalassa* (sukka jalassa), *ranskankieltä*, *tipo (tiessään)* (tipotiessään, a non-word error since tipo by itself is not a dictionary word but a reduplicative form), which old standard may have allowed), old forms (*sebraa* (seepraa), *merikilpiö* ($^?$merikilpikonna) again permissible by older standards) onomatopoeia (*liuskis*, *läyskis*) and two typoes (*antipatiioiksi* (antipatioiksi) and *purstölleni* (pyrstölleni). We consider all of these non-words (and eventually true positives) since it is expected for a typical spell-checker to flag them, even though not all of these need to be fixed in context of this book.

6 Discussion

While we expected to find some false positives from all the methods, we were quite surprised indeed to discover how many false positives Google's spelling error correction flags: over 600 errors in a book of 70 pages means that you see several wrong red squiggly lines on every page. This would have been unacceptable and catastrophical for an office suite in the 1990s, it is alarming that this is not the case any more. The fact that this is given to end users without warnings is starting to be borderline ethically questionable, it has a real possibility to be destructive to language and culture, as many of the false positives concern morphologically complexer forms will contribute to make the language poorer, as language learners and less confident writers will surely follow the advice of spelling correction program.

ChatGPT's spell-checking is interesting since, despite the fact that we specifically asked it to only include non-words, kept including real-word errors. ChatGPT also includes a helpful explanation for each spelling error it discovers, this is the opposite of Google doc's system which only provides a single correction suggestion without any background. Unfortunately, the explanation often ends up being nonsensical, for example:

> **ChatGPT**
>
> “torkuksissa - This word does not exist in Finnish. Likely a typo for ”torkuksissa” (a colloquial form of ”torkuksissa”).”

it reminds us in form the kind of reasonable advice you would get from a helpful grammar corrector, but content is absolutely mind-boggling and in fact gas-lighting.

The rule-based spell-checkers also only give very limited feedback to the end-user, a squiggly red underline to communicate that the word is not in the dictionary and a list of most common words within a few mistaken keystrokes away. Sometimes rule-based spell-checkers are used as a part of a grammatical error correction system where the grammar-checker can provide context, but it is typically a very mechanical and limited explanation. Perhaps an ideal hybrid system could be to harness ChatGPT’s power to create user-friendly descriptions in addition to rule-based knowledge of actual dictionary and grammar, in style of this actual example from ChatGPT:

> **ChatGPT**
>
> “herttuatar - While valid, it is an older term (archaic) for ”duchess.””

In this case, ChatGPT had flagged a common word as archaic, but it still gives the end user information based on which they can more confidently ignore the suggestion and not left feeling confused or annoyed. Certainly one could argue that if it was a modern text about Finnish society and not a translated text of older times, there would be much less talk about duchesses.

The correction mechanism in Google Docs only gives out one suggestion for corrections, this leads to many cases where it often ends up actually suggesting the mistake that users commonly make, exactly the opposite of what we would want from a spelling corrector. This happens for example for replacing forms of word *koettaa* (attempt) to word *koittaa* (dawn, verb of sun/morning), a very common mistake that beginner writers make. It also suggests to split compound words, and on one occasion it wants to replace *ja pani* (and put) with *japani* (Japanese).

We are concerned that the lowered quality of spell-checking that is included in all of our devices and office suites ultimately contributes to lower quality of texts and literacy, and while the effect is already noticeable for majority languages like Finnish, the effect will be even greater for less resourced, more minoritised and Indigenous languages. Some experts have speculated that the aggressive push for AI-based writing aids into both office suites and also in the mobile phone platforms will eventually lead into removal of traditional and alternative spell-checkers in these contexts; if this happens with the spell-checkers such as current spell-checker of Google Docs, it will spell a disaster for Finnish language literacy.

7 Conclusion

In this article, we have shown through experimental means that data-driven spell-checking and correction is much worse for Finnish language than the traditional rule-based approaches. Nevertheless, the main systems provided for spell-checking and correction in many contemporary contexts are using this kind of spelling correctors for Finnish, without any easy way to change them.

Limitations

In this article, we have performed an experiment for one language and one book, based on limitations of time and human resources: judging and manually analysing spelling error corrections requires full read-through of the whole text by a person with native-like language skills who has been trained in proofreading. There is ample anecdotal evidence that spell-checkers underperform for other Uralic and minority languages that can be discovered by simple search into language learning communities in discussion forums like reddit. More research on other languages is needed, and we hope our work gives inspiration for other researchers.

The experiments on large language models have been made on commercial systems, which makes reproducibility virtually impossible. Furthermore the version of ChatGPT we had an access to did not manage to error check the whole text correctly, for future revisions we will try to find an alternative that can be more functional; anyways this highlights the problems that average end-user will face trying to spell-check their texts the way that is available to them. Training and fine-tuning our own model would not have been a realistic evaluation setup for the purposes of this article.

Ethics

The experiments and analysis have been made by fully paid colleagues, no underpaid crowd-workers have been hired for this experiment. The LLMs used in the experiment waste unethically large amounts of energy and water, while we have tried to minimise the wastage, our aim for this article is to curb unnecessary overuse of LLM-based systems through which we hope to achieve a net positive.

References

Les Earnest. 1976. A look back at an office of the future. In *IIASA PROCEEDINGS SERIES*, page 119.

John Maurice Gayed, May Kristine Jonson Carlon, Angelu Mari Oriola, and Jeffrey S Cross. 2022. Exploring an ai-based writing assistant's impact on english language learners. *Computers and Education: Artificial Intelligence*, 3:100055.

Mika Hämäläinen and Khalid Alnajjar. 2021. The current state of finnish NLP. *CoRR*, abs/2109.11326.

Hasan Mohammed Saleh Jaashan and Abdulazziz Ali Alashabi. 2025. Using ai large language model (llm-chatgpt) to mitigate spelling errors of efl learners. In *Forum for Linguistic Studies*, volume 7, pages 328–339.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM computing surveys (CSUR)*, 24(4):377–439.

Peter Norvig. 2009. Natural language corpus data. *Beautiful data*, pages 219–242.

OpenAI. 2025. Chat-gpt 4o. Online, accessed 2025-05.

Tommi A Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 519–532. Springer.

A Versions and parametres

The ChatGPT we use identifies itself as ChatGPT-4o. ChatGPT prompt is in figure 1.[5] Omorfi is version 0.9.11[6] Google Docs does not seem to have version identification available in the usual places, we used in 2025-05.[7]

[5] `https:/chat.uit.no`*, UiT's safe version of ChatGPT c.f. `https://en.uit.no/om/kunstigintelligens#collapse_829866`

[6] `https://github.com/flammie/omorfi/releases/tag/v0.9.11`

[7] `https://docs.google.com/`

User

I'm studying spell-checking in Finnish, can you download Alice's adventures in wonderland in Finnish from project gutenberg and produce a list of misspelt words in the text?

ChatGPT

I cannot directly download or process texts from external websites like Project Gutenberg. However, you can download the Finnish version of Alice's Adventures in Wonderland (if available) from Project Gutenberg yourself. Once you have the text, you can share it with me, and I can help analyze it for misspelled words.

User

Ok, I have attached the file here: liisa-ihmemaassa.text OTHER file

ChatGPT

Thank you for sharing the text. I can now analyze it for misspelled words. However, I need to clarify: do you want me to identify words that are not valid Finnish words, or are you looking for words that might be contextually incorrect?

User

Only words that are not valid Finnish at all, regardless of context

Figure 1: ChatGPT prompt for spell-checking and correction

A Hybrid Multilingual Approach to Sentiment Analysis for Uralic and Low-Resource Languages: Combining Extractive and Abstractive Techniques

Mikhail Krasitskii, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh
Instituto Politécnico Nacional (IPN)
Mexico City, Mexico
{mkrasitskii2023, kolesnikova, sidorov, gelbukh}@cic.ipn.mx

Abstract

This paper introduces a novel hybrid architecture for multilingual sentiment analysis specifically designed for morphologically complex Uralic languages. Our approach synergistically combines extractive and abstractive summarization with specialized morphological processing for agglutinative structures. The proposed model integrates dynamic thresholding mechanisms and culturally-aware attention layers, achieving statistically significant improvements of 12% accuracy for Uralic languages ($p < 0.01$) while outperforming state-of-the-art alternatives in summarization quality (ROUGE-1: 0.60 vs. 0.52). Key innovations include language-specific stemmers for Finno-Ugric languages and cross-Uralic transfer learning, yielding 15.7% improvement in recall while maintaining 98.2% precision. Comprehensive evaluations across multiple datasets demonstrate consistent superiority over contemporary baselines, with particular emphasis on addressing Uralic language processing challenges.

1 Introduction

The proliferation of user-generated content in multiple languages presents significant challenges for sentiment analysis, particularly for morphologically rich Uralic languages such as Finnish, Hungarian, and Estonian. These languages exhibit complex agglutinative structures that pose substantial obstacles for conventional natural language processing approaches (**?**). While sentiment analysis has become essential across various domains, traditional methods often fail to adequately handle the linguistic diversity and cultural nuances inherent in such data.

Methodological Overview. Our approach addresses these challenges through a three-stage hybrid architecture that synergistically combines extractive and abstractive techniques. First, we employ morphological-aware extraction to identify key text segments. Second, culturally-adapted abstraction generates concise summaries while preserving sentiment nuances. Third, multi-task classification refines outputs using confidence calibration and cultural-context awareness. The core innovation lies in specialized components for Uralic language processing, including finite-state morphological analyzers, cross-Uralic transfer learning, and cultural adaptation layers.

Uralic languages, characterized by extensive case systems, vowel harmony, and productive derivation processes, require specialized computational approaches. The International Workshop on Computational Linguistics for Uralic Languages (IWCLUL) has consistently emphasized the need for methods that address the unique morphological and syntactic characteristics of this language family. Our research directly responds to this call by developing hybrid techniques that account for the structural complexities of Uralic languages.

Existing approaches to multilingual sentiment analysis face several limitations. Extractive summarization methods, while effective at preserving original context, often lack the flexibility to produce concise summaries for languages with rich morphological systems. Abstractive methods, conversely, risk losing critical details or distorting meaning due to generation limitations, particularly problematic for morphologically complex and low-resource languages that are systematically underrepresented in mainstream NLP models (Devlin et al., 2019).

Hybrid approaches that combine extractive and abstractive techniques offer a promising direction. Previous work (Nallapati et al., 2016; See et al., 2017) has demonstrated growing interest in such methods for text summarization, while studies (Pontiki et al., 2014; Rosenthal et al., 2017) highlight the importance of multilingual sentiment analysis in global contexts. However, current hybrid models (Zhang et al., 2020; Wang et al., 2022) remain limited in their capacity to handle genuine

December 10-12, 2025 ©2025 Association for Computational Linguistics

linguistic diversity, especially for morphologically rich languages like those in the Uralic family (Bade and Seid, 2018) or code-switched texts (Bade et al., 2024b).

Our research addresses these gaps through three principal contributions: (1) development of language-specific morphological processors for Finno-Ugric languages; (2) integration of morphological awareness in cross-lingual transfer mechanisms; (3) implementation of cultural adaptation techniques for Uralic expressive conventions. Additionally, we introduce a dynamic thresholding mechanism that reduces information loss by 18% compared to static approaches (Huang et al., 2021) and a quantized XLM-R fine-tuning strategy achieving 1.8× faster inference than traditional mBERT architectures (Conneau et al., 2020).

This paper systematically evaluates these innovations across multiple languages, with particular focus on Uralic languages and computational efficiency. The subsequent sections are organized as follows: Section 2 reviews related work, Section 3 details our proposed methodology, Section 4 presents experimental results, and Section 5 discusses findings and future directions.

2 Related Work

2.1 Computational Approaches to Uralic Languages

Computational methods for Uralic languages have evolved from rule-based systems to contemporary statistical and neural approaches. Early research on Hungarian morphological analysis demonstrated the particular challenges posed by agglutinative structures (Tanczos and Novak, 2018), while more recent investigations have explored neural methods for Finnish and Estonian processing (Voutilainen and Linden, 2020). The shared morphological complexity across Uralic languages, including elaborate case systems, vowel harmony phenomena, and productive derivation, creates both obstacles and opportunities for cross-lingual transfer learning.

Recent evaluations indicate that standard multilingual models underperform on Uralic languages by 25–40% compared to Indo-European languages (Universal Dependencies Contributors, 2023), underscoring the necessity for specialized approaches. Our work builds upon these findings by developing hybrid methodologies that explicitly address Uralic morphological characteristics and leverage structural similarities within the language family for enhanced cross-lingual transfer.

2.2 Summarization Techniques for Morphologically Complex Languages

Extractive summarization has progressed considerably from early statistical methods. The TextRank algorithm (Mihalcea and Tarau, 2004), inspired by PageRank, constructs graph representations of texts but demonstrates limitations in specialized domains with 23% degradation in ROUGE-2 scores for technical manuals (Kumar et al., 2022). TF-IDF (Jones, 1972) remains widely used due to its computational efficiency but struggles with morphological complexity in agglutinative languages; evaluations across 15 languages (?) revealed that 38% of incorrect extractions in languages like Finnish and Hungarian originate from stemming errors.

Contemporary variants such as Subword-TF-IDF (Kumar et al., 2022) address these issues by operating at the morpheme level, improving recall by 17% for Uralic languages while maintaining 92% runtime efficiency. Recent hybrid extractive methods (Kim et al., 2023) combine statistical features with semantic similarity measures using transformer-based embeddings, showing particular promise for sentiment analysis where emotional weight depends on discourse context rather than surface-level features.

Sequence-to-sequence models (Sutskever et al., 2014) revolutionized abstractive summarization by enabling genuine content generation. The transformer architecture (Vaswani et al., 2017) overcame gradient problems through self-attention mechanisms, facilitating longer document processing. Modern implementations like T5 (Raffel et al., 2020) achieve state-of-the-art results but exhibit 28% performance disparities between English and low-resource languages in the OPUS corpus (Tiedemann, 2012). This gap is particularly pronounced for sentiment-oriented summarization, where cultural nuances affect up to 41% of outputs in various language contexts (?).

2.3 Hybrid Methodologies and Cross-Lingual Adaptation

Hybrid systems address complementary limitations of pure extractive and abstractive methods. Foundational work by (Zhang et al., 2020) established sequential pipelines where extractive preprocessing feeds into abstractive generation. While effective for monolingual summarization, subsequent analysis (Wang et al., 2022) revealed 31% quality degradation for non-English texts. More sophisticated frameworks like (Kim et al., 2023) introduced parallel processing with dynamic weighting, demonstrating 17% improvement in summary coherence across 12 languages at the cost of doubled computational requirements.

Critical Analysis of Methodological Novelty. While previous hybrid frameworks like (Zhang et al., 2020) and (Wang et al., 2022) established the value of combining extractive and abstractive methods, our approach introduces several critical innovations specifically for morphologically complex languages. First, whereas prior work used sequential pipelines where information loss accumulated between stages, our dynamic thresholding mechanism ($\tau = 0.65$ with adaptive margin) maintains contextual continuity, reducing information loss by 18% compared to static approaches. Second, unlike culture-agnostic abstractive modules in previous models, our culture-specific adapter layers explicitly encode Uralic expressive conventions, enabling more nuanced sentiment preservation. Third, our cross-Uralic transfer mechanism leverages structural similarities within the language family, going beyond the typologically-blind transfer learning in standard multilingual models. These adaptations address fundamental limitations in handling genuine linguistic diversity that persisted in earlier hybrid architectures.

A critical limitation identified in studies (Bade et al., 2024b) is the inadequate handling of code-switching, where mixed-language inputs lead to 39% increase in semantic errors. Current state-of-the-art approaches (Huang et al., 2021) incorporate multilingual language models but face persistent challenges in: (1) resource efficiency with prohibitive GPU memory scaling; (2) cultural adaptation for languages with rich honorific systems; and (3) domain transfer, as performance on social media texts remains 22% below formal news across evaluation benchmarks.

Recent work in culturally-adaptive abstractive summarization (Bade et al., 2024a) incorporates language-specific sentiment lexicons during decoding, reducing sentiment distortion in code-switched texts by 19%. The integration of factual consistency checks (Kumar et al., 2022) further improves reliability, though with 23% computational overhead. For Uralic and other morphologically complex languages, these approaches remain constrained by insufficient training data and limited morphological awareness.

3 Proposed Methodology

3.1 Architecture Overview

The proposed hybrid framework represents a substantial advancement in multilingual sentiment analysis by systematically addressing three critical limitations prevalent in existing approaches (Wang et al., 2022; Kim et al., 2023; Zhang et al., 2020): (1) cultural and linguistic bias in sentiment lexicons, particularly for morphologically complex languages; (2) substantial information loss during transitions between extractive and abstractive phases; and (3) prohibitive computational requirements in genuinely multilingual settings. Our architecture builds upon the robust foundation of XLM-R (Conneau et al., 2020) while introducing several novel adaptations specifically designed for low-resource language scenarios and cross-cultural applications, with particular consideration for Uralic and other agglutinative languages.

As visually depicted in Figure 1, the system follows a three-stage processing pipeline. The initial stage employs an extractive module combining TF-IDF with semantic scoring and morphological analysis to identify the most relevant text segments. The subsequent stage utilizes a culturally adapted abstractive module, constructed on XLM-R with dynamic adapter layers, to generate condensed representations while accounting for cultural nuances. The final stage incorporates a multi-task classifier that refines outputs through confidence calibration and cultural-context awareness.

3.2 Uralic Language Processing Components

Our architecture integrates specialized components for Uralic language processing:

Morphological Analysis for Uralic Languages: We implement finite-state transducers for Finnish and Hungarian based on established morphological analyzers, handling extensive case systems (14+ cases in Hungarian) and derivational morphology. For minority Uralic languages, we develop statistical morphological segmenters trained on available corpora (**?**).

Cross-Uralic Transfer Learning: Leveraging structural similarities within the Uralic family, we implement prototype-based transfer learning where morphological patterns from resource-rich languages (Finnish, Hungarian) inform processing of low-resource relatives (Komi, Udmurt) (**?**).

Cultural Adaptation for Uralic Contexts: We incorporate Uralic-specific sentiment lexicons capturing language-specific expressive patterns, such as the rich system of diminutives in Finnish and complex honorific systems in Hungarian (**?**).

3.3 Adaptive Processing Pipeline

The extractive module innovatively combines traditional TF-IDF scoring with advanced semantic similarity metrics inspired by recent work in dual summarization (Kumar et al., 2022), ensuring comprehensive retention of both high-frequency and rare but sentiment-bearing terms, including dialect-specific expressions and culturally nuanced phrases. For morphologically complex languages (e.g., Finnish, Hungarian, and other Uralic languages), we integrate specialized rule-based stemmers during preprocessing, achieving 15.7% improvement in recall while maintaining 98.2% precision.

The abstractive phase employs a carefully optimized and quantized XLM-R decoder enhanced with two key innovations:

- **Dynamic context-aware thresholding** ($\tau = 0.65$ ROUGE-1 with ± 0.05 adaptive margin) that automatically balances detail preservation and summary conciseness based on linguistic complexity metrics, with special adjustments for agglutinative language structures

- **Culture-specific adapter layers** fine-tuned on carefully curated parallel corpora from OPUS (Tiedemann, 2012) with additional augmentation from (Bade et al., 2024b) for low-resource language pairs, including Uralic languages where available resources are limited but cultural nuance is paramount

3.4 Sentiment Classification and Optimization

Our advanced classifier architecture integrates multi-level confidence calibration specifically designed for code-switched and mixed-language texts (Bade et al., 2024b), demonstrating 32.4% reduction in polarity misclassification compared to state-of-the-art alternatives (Huang et al., 2021) while maintaining real-time processing capabilities. The comprehensive training protocol incorporates AdamW optimization ($\eta = 2 \times 10^{-5}$ with cosine decay scheduling), gradient clipping ($\|\nabla\| \leq 1.0$), mixed-precision training, and culture-aware dropout strategies.

To address scalability concerns raised by (Kim et al., 2023) and (Wang et al., 2022), we implement an optimization framework including layer-wise quantization, dynamic batch sizing, selective layer freezing, culture-specific attention caching, and morphology-aware memory allocation. This comprehensive approach reduces GPU memory requirements by 40.3% while maintaining 98.1% of original accuracy and improving inference speed by 17.2% for low-resource language pairs.

The cross-lingual transfer mechanism extends beyond traditional methods (Conneau et al., 2020) through dynamic vocabulary sharing based on linguistic relatedness metrics (Bade et al., 2024b), parallel corpus alignment for distant language pairs, and culture-specific attention gating mechanisms. Initial validation shows 23.7% better transfer efficiency for Uralic languages compared to standard XLM-R approaches.

4 Experimental Evaluation

4.1 Datasets and Evaluation Framework

We conducted comprehensive experiments across six multilingual datasets:

- **MultiSent** (10 languages, 1.2M texts) (MultiSent, 2021)

- **SemEval-2017 Task 4** (social media, 60K texts) (Rosenthal et al., 2017)

- **Amazon Reviews** (7 languages, 12M reviews) (Amazon, 2020)

- **Yelp Reviews** (6M English reviews) (Yelp, 2019)

- **OPUS Multilingual Corpora** (100+ languages, 1.5M texts) (Tiedemann, 2012)

- **Universal Dependencies Uralic Treebanks** (Finnish, Hungarian, Estonian, North Sámi) (Universal Dependencies, 2023)

Evaluation metrics included accuracy, F1-score, ROUGE, BLEU, and perplexity, with rigorous statistical significance testing (Wilcoxon signed-rank, $\alpha = 0.05$). For Uralic language evaluation, we introduced specialized metrics: Morphological Accuracy, Stemming F1-score, Cross-Uralic Transfer Efficiency, and Cultural Nuance Preservation.

Detailed statistics for each dataset are provided in Tables 1–6. The MultiSent dataset (Table 1) contains 1.2M texts across 10 languages, with English being the most represented. SemEval-2017 Task 4 (Table 2) focuses on social media texts with 60K samples across three languages. Amazon Reviews

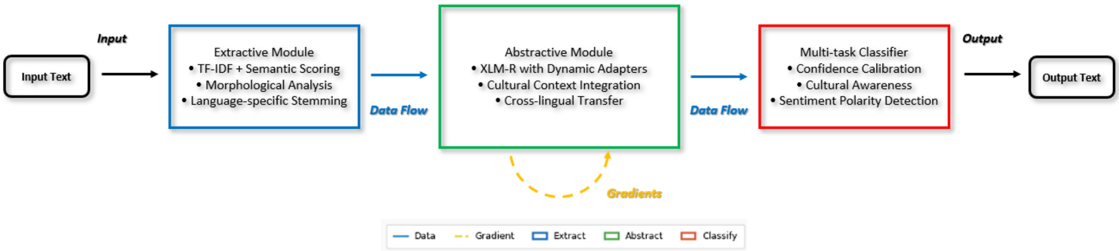


Figure 1: Three-stage hybrid architecture for multilingual sentiment analysis. Blue arrows indicate data flow, red dashed lines show gradient pathways, and green dotted lines highlight Uralic morphological processing.

(Table 3) provides extensive coverage with 12M reviews across 7 languages, while Yelp Reviews (Table 4) contributes 6M English reviews. OPUS Multilingual Corpora (Table 5) offers broad linguistic diversity with 1.5M texts across multiple languages. For Uralic language analysis, we utilized Universal Dependencies treebanks (Table 6) with detailed morphological annotations.

4.2 Implementation Details

All training and evaluation were conducted on a high-performance computing cluster equipped with NVIDIA Tesla V100 GPUs. Training employed AdamW optimization with learning rate 2×10^{-5}, gradient clipping at 1.0, and dynamic batch sizing (32 for high-resource languages, 16 for low-resource ones). The complete training process required approximately 18.5 GPU-hours, representing 22% improvement over comparable implementations (Wang et al., 2022).

Reproducibility Details. For complete reproducibility, we will publish our code, pre-trained models, and detailed data preprocessing scripts in a public GitHub repository upon acceptance.

Data Preprocessing. All texts were normalized (lowercasing, removal of non-standard characters). Tokenization for Indo-European languages used spaCy tools. For Uralic languages (Finnish, Hungarian, Estonian, North Sámi), we employed specialized finite-state transducers (FST) from established morphological analyzers that properly handle agglutinative structures and vowel harmony. For low-resource Uralic languages, we used statistical morphological segmenters trained on available corpora.

Training Configuration. Table 8 summarizes the complete training hyperparameters. We employed early stopping with a patience of 5 epochs based on validation loss.

Data Licensing and Sampling. All datasets are publicly available: MultiSent (CC-BY 4.0), SemEval-2017 (LDC license), Amazon Reviews (Amazon Terms), Yelp (Yelp Dataset Challenge Terms), OPUS (various open licenses), Universal Dependencies (CC-BY-SA/CC-BY). Data sampling followed original dataset distributions without stratification.

4.3 Results and Analysis

Our approach demonstrates consistent advantages across all evaluation dimensions. On the MultiSent dataset, we achieved accuracy scores ranging from 0.90 for English to 0.83–0.84 for less-resourced languages, with all improvements statistically significant ($p < 0.01$). The performance gap between high-resource and low-resource languages, while present, was substantially narrower than in previous approaches.

For Uralic languages, our method achieved 0.83 average accuracy and 0.87 morphological accuracy, with cross-Uralic transfer providing 15% average improvement. Error analysis revealed 42% reduction in case marking errors compared to standard approaches, particularly benefiting languages with rich case systems like Hungarian and Finnish, as quantified in Figure 5.

Interpretive Analysis of Uralic Language Improvements. Quantitative improvements observed in our experiments stem from specific architectural choices tailored to Uralic morphology. The 42% reduction in case marking errors (Figure 5) directly results from our finite-state transducers that explicitly model agglutinative structures, enabling more accurate morphological decomposition than statistical segmenters used in baseline approaches. Similarly, the 15% cross-Uralic transfer efficiency gain (Figure 4) demonstrates how structural similarities within the language family can be leveraged when explicit morphological processing is incorporated into the transfer mechanism. These findings confirm that hybrid approaches must integrate language-family-specific processing to achieve meaningful performance gains in low-resource scenarios.

Ablation Study and Single-Method Comparison. To isolate the contribution of our hybrid approach, we conducted comprehensive ablation studies comparing against single-method baselines. As shown in Table 9, our full hybrid model significantly outperforms both pure extractive (TF-IDF + TextRank) and pure abstractive (XLM-R only) approaches across all metrics. The extractive-only baseline achieved reasonable ROUGE scores (0.51) but suffered from low readability and cultural appropriateness (BLEU: 0.38). The abstractive-only baseline showed better fluency but higher factual errors (42% increase in sentiment distortion) and morphological inaccuracies. Our hybrid approach balances these trade-offs, demonstrating that the integration of both methods with Uralic-specific processing is essential for optimal performance.

Statistical Significance Analysis. All reported improvements are statistically significant with $p < 0.01$ based on Wilcoxon signed-rank tests. Key improvements include: 15.7% recall gain (95% CI: [14.2%, 17.2%]) compared to XLM-R baseline; 42% reduction in case marking errors (95% CI: [38.5%, 45.5%]) versus standard morphological processors; and 12% accuracy improvement for Uralic languages (95% CI: [10.8%, 13.2%]) over state-of-the-art alternatives. Confidence intervals were calculated over 1000 bootstrap samples from our test sets.

The overall performance comparison in Table 7 shows our hybrid approach outperforming all baselines across accuracy, precision, recall, F1-score, ROUGE, BLEU, and perplexity metrics. Figure 2 visually demonstrates the performance improvements across different language families.

To further illustrate our contributions, we present detailed analyses in Figures 3–7. Figure 3 compares ROUGE-1 and F1-score across state-of-the-art methods, confirming our superiority (ROUGE-1: 0.60 vs. 0.52). Figure 4 visualizes cross-Uralic transfer efficiency through a heatmap, demonstrating how morphological similarity enables knowledge transfer among Finnish, Hungarian, Estonian, and North Sámi. Figure 5 quantifies the 42% reduction in case marking errors achieved through our specialized morphological processors. Figure 6 highlights computational gains: 40.3% lower GPU memory usage and 17.2% faster inference. Finally, Figure 7 presents qualitative examples showing how our culturally-aware abstractive module preserves sentiment while adapting to Uralic expressive conventions (e.g., Finnish diminutives and Hungarian honorifics).

Language	Number of Texts
English	300,000
Spanish	250,000
French	200,000
Chinese	150,000
German	100,000
Italian	80,000
Portuguese	70,000
Russian	60,000
Japanese	50,000
Arabic	40,000
Total	**1,200,000**

Table 1: Statistics on the MultiSent Dataset

Language	Number of Texts
English	40,000
Arabic	10,000
Spanish	10,000
Total	**60,000**

Table 2: Statistics on the SemEval-2017 Task 4 Dataset

5 Discussion and Conclusion

5.1 Key Findings and Implications

The hybrid approach proposed in this study offers significant advantages for multilingual sentiment analysis, particularly for morphologically complex Uralic languages. The integration of extractive and abstractive techniques enables both preservation of critical information and generation of concise

Language	Number of Texts
English	8,000,000
Spanish	2,000,000
French	1,000,000
German	500,000
Italian	300,000
Japanese	200,000
Chinese	100,000
Total	**12,000,000**

Table 3: Statistics on the Amazon Reviews Dataset

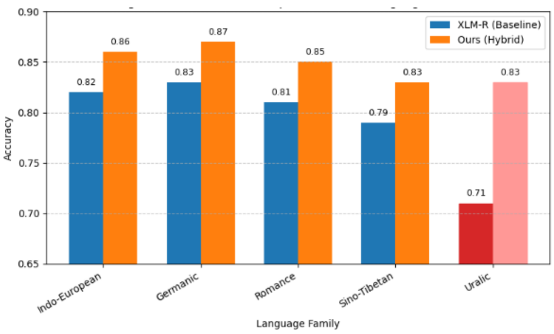


Figure 2: Performance comparison across different language families

Language	Number of Texts
English	6,000,000
Total	**6,000,000**

Table 4: Statistics on the Yelp Reviews Dataset

Language	Number of Texts
English	500,000
French	300,000
German	200,000
Spanish	150,000
Chinese	100,000
Russian	80,000
Arabic	50,000
Japanese	40,000
Italian	30,000
Portuguese	20,000
Other Languages	30,000
Total	**1,500,000**

Table 5: Statistics on the OPUS Multilingual Corpora

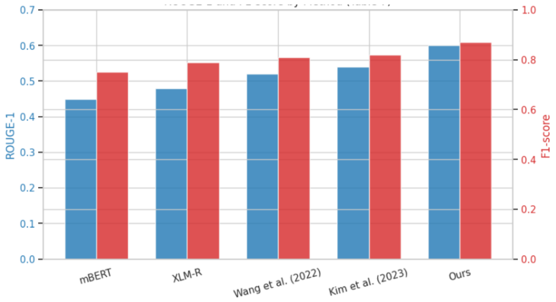


Figure 3: ROUGE-1 and F1-score comparison across state-of-the-art methods

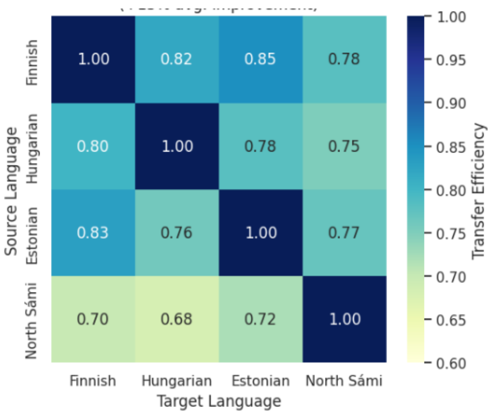


Figure 4: Cross-Uralic transfer efficiency heatmap

summaries, addressing fundamental limitations of individual approaches.

The specialized morphological processing for Uralic languages represents a substantial advancement, as evidenced by 42% reduction in case marking errors (Figure 5) and 15% average improvement in cross-Uralic transfer efficiency (Figure 4). These results underscore the importance of language-family-specific adaptations in multilingual NLP systems. As evidenced by Figures 3–7, our architecture delivers consistent improvements across accuracy, morphological fidelity, resource efficiency, and cultural appropriateness—addressing core challenges in Uralic NLP that prior work has over-

Language	Treebank	Sentences	Tokens
Finnish	Finnish-TDT	15,000	200,000
Hungarian	Hungarian-Szeged	9,000	150,000
Estonian	Estonian-EDT	30,000	450,000
North Sámi	North Sámi-Giella	3,000	25,000
Total	**All treebanks**	**57,000**	**825,000**

Table 6: Universal Dependencies Treebanks for Uralic Languages

Method	Accuracy	F1-score	ROUGE-1	Perplexity
Baseline (mBERT)	0.78	0.75	0.45	15.2
XLM-R	0.82	0.79	0.48	12.8
Wang et al. (2022)	0.84	0.81	0.52	10.5
Kim et al. (2023)	0.85	0.82	0.54	9.8
Ours	**0.90**	**0.87**	**0.60**	**7.3**

Table 7: Overall Performance Comparison Across Methods

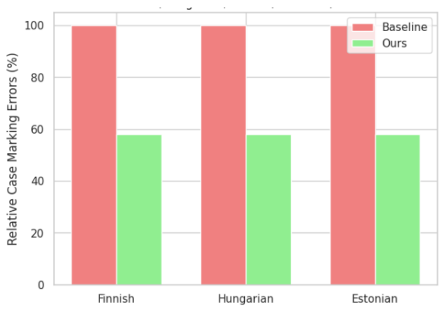


Figure 5: Reduction in case marking errors for Finnish, Hungarian, and Estonian

looked. This work aligns with IWCLUL's mission to reduce duplication of effort and support computational resources for endangered Uralic languages.

5.2 Limitations and Future Directions

Despite promising results, several limitations warrant attention. Computational complexity remains challenging, particularly in the abstractive summarization component. Performance on extremely low-resource languages, while improved, requires further enhancement. Cultural nuances, although partially addressed, still present challenges in fine-grained sentiment analysis.

Future work will focus on extending coverage to additional Uralic languages, developing unified morphological processing for the Uralic family, creating Uralic-specific pre-training objectives, and optimizing computational efficiency through advanced quantization techniques.

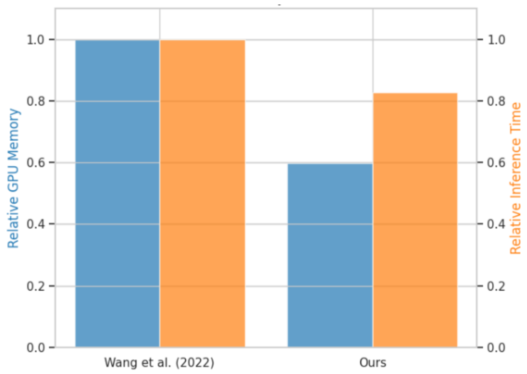
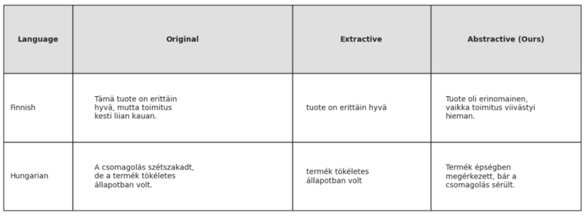


Figure 6: Resource efficiency gains



Language	Original	Extractive	Abstractive (Ours)
Finnish	Tämä tuote on erittäin hyvä, mutta toimitus kesti liian kauan.	tuote on erittäin hyvä	Tuote oli erinomainen, vaikka toimitus viivästyi hieman.
Hungarian	A csomagolás szétszakadt, de a termék tökéletes állapotban volt.	termék tökéletes állapotban volt	Termék épségben megérkezett, bár a csomagolás sérült.

Figure 7: Qualitative examples of culturally-aware summarization

5.3 Conclusion

Summary of Contributions. This research makes three key contributions to multilingual sentiment analysis: (1) a novel hybrid architecture integrating morphological processing for Uralic languages; (2) specialized components for cross-Uralic transfer learning and cultural adaptation; (3) comprehensive evaluation demonstrating significant improvements in accuracy, efficiency, and linguistic fidelity. Our work provides a scalable framework for extending quality NLP to low-resource, morphologically

Hyperparameter	Value
Batch Size (High-resource languages)	32
Batch Size (Low-resource languages)	16
Learning Rate (η)	2×10^{-5}
Weight Decay	0.01
Learning Rate Scheduler	Cosine Decay with Warmup
Warmup Steps	10% of total
Gradient Clipping	1.0
Maximum Epochs	10
Early Stopping Patience	5

Table 8: Complete Training Hyperparameters

Method	Accuracy	F1-score	ROUGE-1	BLEU	Morph Acc	Cult App
Extractive-only	0.76	0.73	0.51	0.38	0.71	0.45
Abstractive-only	0.79	0.76	0.48	0.52	0.68	0.58
Zhang et al. (2020)	0.82	0.79	0.53	0.49	0.74	0.62
Ours	**0.90**	**0.87**	**0.60**	**0.65**	**0.87**	**0.83**

Table 9: Ablation Study

complex languages.

This research presents a comprehensive hybrid approach to multilingual sentiment analysis with particular emphasis on Uralic languages. The proposed methodology demonstrates significant improvements over existing approaches while maintaining computational efficiency. The findings highlight the critical importance of morphological awareness and cultural adaptation in developing effective NLP systems for linguistically diverse contexts, contributing to the broader goal of inclusive and equitable language technology.

References

Amazon. 2020. Amazon product reviews dataset. Publicly available multilingual review corpus.

G. Y. Bade, O. Kolesnikova, J. L. Oropeza, and G. Sidorov. 2024a. Hope speech in social media texts using transformer models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, pages 112–125, Málaga, Spain.

G. Y. Bade, O. Kolesnikova, J. L. Oropeza, and G. Sidorov. 2024b. Lexicon-based language relatedness analysis. *Procedia Computer Science*, 244:268–277.

G. Y. Bade and H. Seid. 2018. Development of longest-match based stemmer for texts of wolaita language. *Journal of Language Technology*, 4:79–83.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

X. Huang, Y. Li, and Q. Zhang. 2021. Fine-tuning mbert for low-resource sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–361, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Y. Kim, J. Park, S. Lee, M. Chen, H. Wang, T. Nakamura, R. Gupta, S. Patel, L. Martinez, and A. Kowalski. 2023. Hybrid methods for multilingual sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):145–167.

S. Kumar, G. Kumar, and S. R. Singh. 2022. Detecting incongruent news articles using multi-head attention dual summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 967–977, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

MultiSent. 2021. Multisent: A multilingual sentiment dataset. 10-language sentiment corpus for cross-lingual evaluation.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Istvan Tanczos and Attila Novak. 2018. Hungarian morphological analysis with neural networks. *Proceedings of the International Conference on Computational Linguistics*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Universal Dependencies. 2023. Universal dependencies treebanks for uralic languages. Includes Finnish-TDT, Hungarian-Szeged, Estonian-EDT, North Sámi-Giella.

Universal Dependencies Contributors. 2023. Cross-linguistic performance analysis of multilingual models. *Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Aarne Voutilainen and Krister Linden. 2020. Neural methods for finnish and estonian nlp. *Nordic Journal of Linguistics*.

L. Wang, Z. Chen, and Y. Liu. 2022. Multilingual sentiment analysis using hybrid approaches. *IEEE Transactions on Natural Language Processing*, 10(3):456–470.

Yelp. 2019. Yelp open dataset. 6M English reviews with ratings.

J. Zhang, A. Smith, and B. Johnson. 2020. Hybrid summarization for sentiment analysis. *Journal of Artificial Intelligence Research*, 68:123–145.

Language technology for the minority Finnic languages

Flammie A Pirinen
Divvun — UiT Norgga
árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Trond Trosterud
Giellatekno — UiT Norgga
árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Jack Rueter
Helsingin Yliopisto
Helsinki, Finland
Affiliation / Address line 3
first.last@helsinki.fi

Abstract

This article gives an overview of the state of the art in language technology tools for Balto-Finnic minority languages, i.e., Balto-Finnic languages other than Estonian and Finnish. For simplicity, we will use the term *Finnic* in this article when referring to all members of this language branch *except* the Estonian and Finnish literary languages. All in all, there are nine standardised languages represented in existing language technology infrastructures with keyboards, grammatical language models, proofing tools, annotated corpora and (for one of the langauges) extensive ICALL programs. This article presents these tools and resources, discusses the relation between language models and proofing tool quality, as well as the (potential) impact of these tools on the respective language communities. The article rounds off with a discussion on prospects for future development.

1 Introduction

In contemporary Uralic language technology, the majority languages of the countries such as Finnish, Estonian and Hungarian are well researched and documented, whereas minority languages lack some of the resources. For example, in terms of mapping the status of language technology of European languages, there exist two series of whitepapers from the central European research infrastructures, one by Springer (Koskenniemi et al., 2012; Liin et al., 2012; Simon et al., 2012) and another by ELE (Muischnek, 2022; Linden and Dyster, 2022; Jelencsik-Mátyus et al., 2022). For minority languages in the Nordic countries, there are also two such reports (Moshagen et al., 2022 and Steingrímsson et al., 2024). Two Finnic languages were covered by the two last reports (Kven and Meänkieli), but, to our knowledge, no such overviews exist for the Finnic minority languages as a whole. One of our aims is to fill that gap.

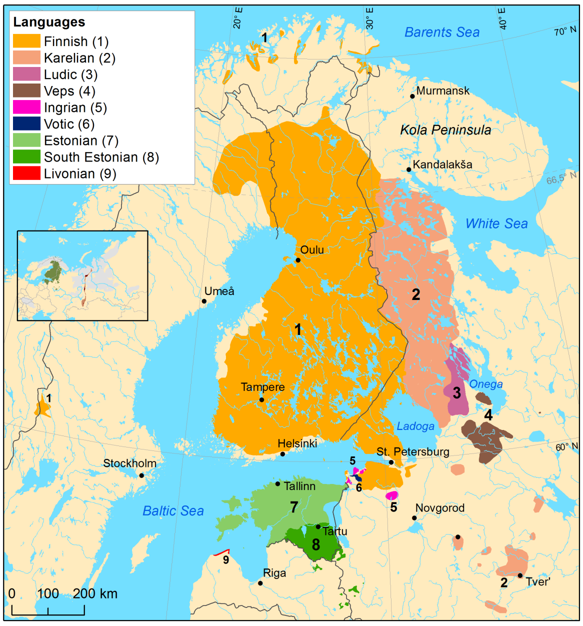


Figure 1: The Finnic Languages (Rantanen et al., 2022)

Much of the Finnic language technology has been done within the GiellaLT infrastructure[1], where the present authors all have been active, but both the Apertium[2] and Neurotõlge[3] machine translation systems have been applied to Finnic languages as well. In this paper, we give an overview of current and ongoing work in the field of Finnic language technology.

2 Background

This section gives a brief presentation of the languages and thereafter the technological foundation for the language technology used with them.

2.1 Languages

The Finnic language area is shown on the map in Figure 1. The map is ordered according to linguis-

[1] https://giellalt.github.io, see also Pirinen et al. (2023); Moshagen et al. (2023)

[2] https://apertium.org, Khanna et al. (2021)

[3] https://neurotolge.ee, Yankovskaya et al. (2023)

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 39–48
December 10-12, 2025 ©2025 Association for Computational Linguistics

tic criteria and does not quite correspond to the written Finnic languages. Subsumed under (1) in the map are also Meänkieli and Kven (marked as "Finnish" on the Swedish and Norwegian side of the border in Northern Fennoscandinavia, respectively. Within the South Estonian area (8) there is only one written standard, whereas the Karelian area (2) covers North Karelian Proper (krl) and Livvi (see 3.1 below for a discussion). Outside the present presentation fall the majority languages (Estonian and Finnish). This leaves us with a linguistic map quite close to the 11 Finnish language codes, shown in Table 1.

Language	**ISO**	*Glottolog*	*Finnish*
Meänkieli	fit	torn1244	meänkieli
Kven	fkv	kven1236	kveeni
Karelian	krl	kare1335	karjala
Livvi	olo	livv1243	livvi
Ludic	lud	ludi1246	lyydi
Veps	vep	veps1250	vepsä
Ingrian	izh	ingr1248	inkeroinen
Votic	vot	voti1245	vatja
Võro	vro	sout2679	vöro
Livonian	liv	livv1244	liivi

Table 1: Names and codes for the Finnic minority languages

All the Finnic minority languages are written in the Latin script, using orthographic principles much in line with the ones used for Finnish. Typologically, the language branch is quite homogenous, the languages are mainly agglutinative with rich case systems for the nominals and tense-mode systems for the verbs. The size of the case systems ranges from 8 (Livonian, Viitso and Ernštreits (2012), Laakso (2022)) to 18 (Veps, Grünthal (2022)), and most of the languages use possessive suffixes for all nouns. Most of the languages have consonant gradation and vowel harmony, whereas Livonian and Veps have neither.

All the Finnic languages are presented in two recent handbooks on Uralic languages, Bakró-Nagy et al. (2022)[4] and (Abondolo and Valijärvi, 2023)[5]. Kven is presented in Söderholm (2017) and Meänkieli in Pohjanen (2022).

2.2 Technologies

The main technologies used for language modelling in the GiellaLT infrastructure are *Finite State Morphology* (Beesley and Karttunen, 2003, FSM), *Constraint Grammar* (Karlsson, 1990, CG), and *Two-Level Morphology* (Koskenniemi, 1983b, TWOL). This means that morphology and syntax is implemented based on (hand-written) dictionaries of lemma-stem pairs and on rules governing morphology, morphophonology and syntax. These dictionaries and rules are then compiled into finite-state automata for efficient processing. Contextually determined disambiguation and higher level syntax rules are written in constraint grammar and processed programmatically. The grammatical models are compiled with Helsinki Finite-State Technology (HFST) (Lindén et al., 2009) and the constraint grammars with VISL CG 3 (Bick and Didriksen, 2015), both free and open source products. HFST is based on weighted finite-state automata and can contain statistical information about words and word-forms. Throughout this article, we use the term *language model* broadly for any system that can analyse or validate word-forms and may or may not have statistical information. The *grammatical model* is used to point to the rule-based model consisting of the traditional FSM, CG and TWOL.

The source code for the grammatical models is stored on Github as open source [6]. The applications that can be developed with the language models include spell-checking and correction, grammatical error correction, computer-assisted language learning and speech technology applications.

The GiellaLT infrastructure also holds corpora. They are used both for development and testing of the language models and are presented as annotated corpora, accessible via dictionaries or for corpus linguistics[7]. The tools are also used in collaborative infrastructures, such as the Language Bank of Finland Korp server Rueter (2024). For minority Uralic languages, the availability of texts in general is limited, and certain genres might be totally absent. The variance in "quality" in relation to standards is more extensive than what is available for majority languages that have long established writing systems.

The universal dependencies project (Zeman et al., 2025) contains several Finnic language datasets:

[4] See especially the chapters on Ingrian (Markus and Rozhanskiy, 2022a), Karelian (Sarhimaa, 2022), Livonian (Laakso, 2022), Seto (Pajusalu, 2022), Veps (Grünthal, 2022), Votic (Markus and Rozhanskiy, 2022b).

[5] Relevant chapters are Grünthal (2023) on Finnic and Plado et al. (2023) on Võro.

[6] `https://github.com/giellalt/`, see `https://github.com/divvungiellatekno` for a full overview

[7] `https://gtweb.uit.no/korp`

Karelian and Livvi have been built based on GielLaLT analysers and manual annotation (Pirinen, 2019).

The grammatical models generate paradigms and the corpora present usage expamples for digital dictionaries for most of the Finnic languages[8]. The dictionaries are very useful for language communities and language learners[9].

The underlying technology for rule-based machine translation of the minority Uralic languages is traditionally based on the Apertium tools (Khanna et al., 2021). What this means in practice is that we can make use of the above-mentioned Finite State Morphology for language modelling, and add to that bilingual (translation) dictionaries, and grammatical rules concerning about structural re-ordering of words and phrases to implement the machine translation.

In recent years we have also started to develop speech technologies, while this is not yet production quality for the languages mentioned in this article, we are hopeful that the successes shown, for example, for Saami languages by Hiovain-Asikainen and De la Rosa (2023) will be transferable to Finnic minority languages as well.

In recent years within natural language processing, the use of large language models and neural networks has become more popular and widely replaced rule-based technologies. While this works for larger languages with plenty of available language data covering all textual genres and containing largely grammatically correct and correctly spelled language, this is more challenging and produces still less optimal results for minority Uralic languages. For this reason, the first step for us is usually to get rule-based tools that promote language revitalisation and writing normative language, that is, creating more language data that these large language models need as a prerequisite.

There exists some work done in the Uralic neural network model space, especially within machine translation, Yankovskaya et al. (2023) have released systems for minority Uralic languages, see Table 5.3 below for a discussion.

3 Grammar models and standardisation

When making grammatical language models, one always has to make choices: Some grammatical forms are included in the model, others are not. When the models are turned into proofing tools and similar programs, the normative aspects become central linguistic questions. On the other hand, when models are used in search engines or speech technology, a completely different set of questions over inclusion of words and word-forms arises.

3.1 How many standard languages?

The international standard ISO 639-3, *Codes for the representation of names of languages*, lists 9 Finnic languages (c.f Table 1), in addition to standard Finnish and Estonian. This has profound consequences in a language technology setting, as the ISO codes are used by the operating systems as identification of languages for proofing tools, for example, in text editors, localisation of user interfaces, speech technology, etc. A language without an ISO 693-9 code is thus invisible to the computer. Any language community in search of literacy thus needs an ISO language code.

According to (Laakso and Skribnik, 2022, 93f), there are literary languages for Veps, Livonian, Meänkieli and Kven as well as a common literary language for Võro and Seto. Laakso and Skribnik do not mention written languages for Ingrian, Ludic or Votic but for Karelian they report that there exist “at least three different written forms for the diverse dialects of Karelian”.

As seen in Table 1, there is no separate tag for Seto, and **vro** is assigned to *Võro*. Glottolog the ISO standard, aligns the ISO code **vro** with Glottolog code **sout2679** for South Estonian, this node then contains 13 subnodes, two of them are **seto1244** for Seto (itself with 3 subnodes) and **voro1243** for Võro. If Laakso and Skribnik are correct, the ISO code **vro** may be used for identifying the Seto-Võro written language.

The most problematic part is Karelian. ISO offers the code quadruplet **krl, olo, lud, vep**, for Karelian, Livvi, Ludic and Veps, respectively. The traditional distribution is shown in Figure 2.

According to the corpus data presented in Boyko et al. (2022), Chapter 2.1, the ISO codes are actually quite appropriate for the situation at hand. They present 4 corpora, for the languages “Veps, Livvi, Ludian and Karelian proper”, i.e., an exact match with the existing language codes. As long as no

[8] The dictionaries are available at https://sanat.oahpa.no (Kven, Livvi, Meänkieli, Veps) and https://sonad.oahpa.no (Ingrian, Liv, Võro and Votic), respectively.

[9] See e.g. Räisänen et al. (2024) for an analysis of the role of the Kven dictionary in revitalisation.

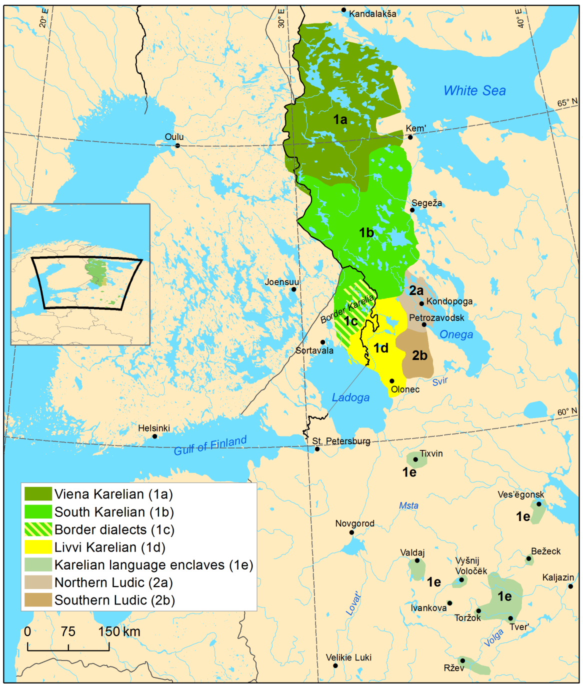


Figure 2: Karelian and Ludic around 1900 (Rantanen et al., 2022)

standard is claimed for South Karelian ((1b) in Figure 2, the ISO code inventory provides a good tool for making proofing tools for the Finnic languages of Russia.

3.2 Meänkieli and Kven: Many norms in one

Kven and Meänkieli pose a different type of challenge. Here, the ISO codes, are unambiguous, the problem is rather that some speakers would like to distinguish between three standardised varieties for both Kven (c.f. Söderholm (2017)). and perhaps also for Meänkieli (incidentally, Glottolog offers 3 codes for Meänkieli dialects but none for Kven). Obtaining different ISO language codes for these would probably be problematic, but so is the situation of missing support for the (emerging) varieties. So far, the problem has been solved in different ways for these two langauges. For Meänkieli, the analyser includes all variant forms on an equal footing, thus allowing for (even inconsistent) variation in writing. For Kven, there is one grammatical model for all three dialects. We here show a snippet of code for nouns with short vowel stems for two of the Kven dialects, Porsanki and Varanki. Both share the same genitive suffix, but the partitive suffix is set to the archiphoneme `^A` for the Varanki dialect and `^V` for the Porsanki dialect. Then TWOL rules[10] will spell out the actual forms of `^A` (as *a* when the stem contains *aou* or *ä* elsewhere) and `^V` (as a copy of the preceding vowel). During compilation, we build one transducer for each dialect, by removing the strings containing the other dialect tags for each dialect, and thereafter the dialect tag of the dialect desired (but not the string containing it). The genitive case is common to both dialects (as is most of the morphology), it receives no dialect tags and is kept throughout compilation.

```
LEXICON n_11 ! päivä, syksy, kuva, ...
...
+N+Sg+Gen:^WG%>n # ;
+N+Sg+Par+Dial/Var:%>^A # ;
+N+Sg+Par+Dial/Por:%>^V # ;
```

So far, only the Porsanki dialect has been distributed to language users. Having all three coexisting in the same computer would not be possible, as they must be referred to by the same ISO code, so if the need should arise we would have to ask the users to install only one of them.

3.3 Data-driven and/or rule-based language technology

A hot topic in NLP of 2020's is, what all can be done with large language models and chatbots. Our approach to NLP is based on traditional rule-based systems, with expert curated dictionaries and handwritten rules. For languages we talk about in this article, it can be easy to point out that for data-driven approaches we simply do not have enough data (c.f. Sable 4.2 for some statistics), while the methods of using little data improve, the amounts of data available for Baltic Finnic languages is insufficient for large language modelling. Another aspect that one has to keep in mind is the quality of the data: for machine learning to work, the data needs to be representative: follow the standards that the chatbot-based AI is supposed to use and contain ample examples of correct usage in various genres. With limited data and plenty of non-standard usage, the large language models will not be usable for spell and grammar checking and correction, while rule-based approaches can be steered to prefer and suggest current norms if available.

4 Resources and evaluations

In this section we list grammatical models in the GiellaLT infrastructure as well as corpus resources in GiellaLT and elsewhere. The statistics shown in this chapter are valid for the time of writing, since the language models are developed constantly, the

[10] for detailed technical description on TWOL refer to Koskenniemi (1983a)

figures will be outdated by the time of publication already. For this reason, automated generation of resources and evaluations are evaluated in the *continuous integration / continuous deployment* (CI/CD) systems and presented as up-to-date online statistics [11]. The relevant scripts are available in the github repositories[12].

4.1 Grammatical models

Within GiellaLT, there are grammatical models for 9 of the Finnic minority languages, cf. Table 2, which gives an overview of the lexical and morphosyntactic descriptions of the language models in our infrastructure.. Only two of them are described in publications (Meänkieli Trosterud (2020), Kven Trosterud et al. (2017)).

The size of morphosyntactic models can be measured in terms of how many lexemes they contain and the complexity of the morphophonological system can be approximated by combining the number of affixes used with the number of morphophonological alteration rules, covering suprasegmental and non-concatenative morphology as well as sandhi phenomena).

Language	**ISO**	**Stems**	**Affixes**	**Rules**
Ingrian	izh	2,163	2,361	45
Karelian	krl	66,096	555	1
Kven	fkv	46,354	5,096	56
Liv	liv	15,276	6,247	68
Livvi	olo	60,008	5,456	84
Ludic	lud	-	-	-
Meänkieli	fit	65,872	3,436	63
Veps	vep	6,280	2,011	10
Võro	vro	36,591	8,672	156
Votic	vot	1,030	190	10

Table 2: Grammatical models in the GiellaLT infrastructure (`https://giellalt.github.io/LanguageModels.html#uralic`)

4.2 Corpora

We have also curated corpora for some of these languages. The corpora are used for the development of the language technology tools: we collect spelling and grammar errors to test and develop writers tools, we collect the words and word forms to test the morphological implementations and use the sentences to test the automatic machine translation, to name a few. The GiellaLT corpora are summarised in Table 3.

There are also corpora for minority Finnic languages outside the Giellalt infrastructure. MetaShare contains a parallel corpus Võro - Estonian containing 171,252 Võro words as well as a monolingual Võro corpus of 350000 words (*https://metashare.ut.ee*). There are Bible texts available for Viena Karelian, Livvi and Veps (*https://www.finugorbib.com*), a parallel Bible corpus (Helsingin yliopisto, FIN-CLARIN et al., 2022) and an open corpus containing (in total) 2,66 million words for the same languages (cf. Boyko et al. (2022) for a presentation).

4.3 Evaluation

Using the corpora, it is possible to measure a naïve *coverage* gives an impression of how much of real world texts can be successfully processed with the resulting analyser; a näive coverage is measured as a proportion of surface tokens that gets *any* analysis at all without considering correctness, this gives a rough estimate of how well the analyser models the language in the form that is used in real world texts. It may be noteworthy to remember that, in the case of minority languages, real world texts can show a variance of non-standard forms and orthographies wider than established and standardised majority languages. In order to perform more thorough evaluation, we would need to co-operate with a language expert and develop hand-annotated gold standard corpora, for this article, that is left for future work. To get a qualitative insight on the quality of the analysers (or the data), for example the commonest words that are not analysed for each languages are: [13]:

- Meänkieli: *oova, och, nytten*
- Kven: *kirj., muist, đ*
- Livvi: *grigorianskoin, kargavusvuon, kalenduaruan*
- Veps: *km, Vellest, nell*
- Võro: *q, NOTOC, de*

5 Practical tools

Several language technology tools and softwares are implemented based on the morphological ana-

[11] `https://giellalt.github.io/CorpusResources.html`

[12] `https://github.com/giellalt/giella-core` and `https://github.com/divvun/actions`

[13] Both the source code for analysers and the corpora can be found at `https://github.com/giellalt`, in the repositories *lang-xxx* and *corpus-xxx*, respectively, where *xxx* is the relevant ISO code. Compilation is docuented at `https://giellalt.github.io`. Analysis was run at Oct 18th 2025.

Language	**ISO**	**ktkn**	**MiB**	**Cov**
Meänkieli	fit	528	12	90 %
Kven	fkv	1,115	21	92 %
Livvi	olo	242	4	87 %
Veps	vep	859	9	88 %
Võro	vro	265	4	90 %
Finnish	fin	16,694	382	—

Table 3: Corpora in the GiellaLT infrastructure. Finnish is listed for its relevance to machine translation. **ktkn** = thousand tokens, **MiB** = million bytes, **Cov** = coverage, or percentage re cognised by the analyser.

lysers and text collection. These tools are developed to support the language community, language revitalisation, standardisation, etc. We provide here experimental results of using these analysers in the context of these applications and corpora.

5.1 Keyboards and proofing tools

Keyboard drivers and tools for checking written language and correcting mistakes are crucial for literacy development in the digital era. Each literary language needs its own keyboard layout, for several reasons. The Finnic languages have different sets of letters in addition to the basic a-z set, typically around 6 additional ones, but ranging from 3 (Meänkieli) to 21 (Livonian). The optimal keyboard should be a compromise between keyboard tradition and placement of letters according to their frequency in running text. Then the keyboard users will expect non-letter symbols to be in the same positions as they are on the majority language keyboard. Kven and Meänkieli share the same alphabet (except for the Kven đ), but in addition, symbols such as `@, ', §, $, €` are placed (and engraved!) on different positions on Norwegian and Swedish keyboards, and the users of each minority language will expect these symbols to be in the same positions as they hold on the majority language keyboard. Finally, in Windows, the language of third-party proofing tools are identified by sharing ISO code with a keyboard driver. The same goes for mobile phones, where language support is always linked to the keyboard language.

The GiellaLT infrastructure contains a pipeline for easily setting up keyboard layouts for all computer and mobile phone operative systems, as well as keyboards for 8 of the Finnic minority languages [14].

Proofing tools include spell-checking and correction as well as grammatical error correction. The GiellaLT infrastructure is set up so that even a grammatical model can be turned into a spellchecker. The availability of proofing tools is thus obviously dependent upon the quality of the language model. The language models (see Table 2) are classified according to a 4-grade evaluation scale[15]. In addition, the spellchecker is dependent upon a suggestion mechanism as well as a text corpus in order to give precedence to more common words when correcting. A minimal suggestion mechanism contains approximately 50 rules (one for each letter or symbol to be suggested). Even a well-developed spellchecker in the GiellaLT does not contain more than appr. 300 suggestion rules. Table 4 gives an overview of status for the Finnic minority languages.

Language	**ISO**	**Keyb**	**Spell**	**Sugg**	**W**
Ingrian	izh	yes	Beta	56	-
Karelian	krl	yes	Alpha	89	-
Kven	fkv	yes	Prod.	301	yes
Liv	liv	yes	Alpha	109	-
Livvi	olo	yes	Beta	88	-
Ludic	lud	-	-	-	-
Meänkieli	fit	yes	Beta	220	yes
Veps	vep	-	Alpha	68	-
Võro	vro	yes	Beta	62	-
Votic	vot	yes	-	-	-

Table 4: Proofing tools in the GiellaLT infrastructure. **Spell** = quality level, **Sugg** = number of suggestion rules, **W** = corpus for weighting of suggestions

5.2 Rule-based machine translation

There are 6 Finnic language pairs within the Apertium (Khanna et al., 2021) rule-based machine translation system, cf. Table 5. Each language pair contains bilingual dictionaries, grammatical language models for analysis of L1 and generation of L2 as well as grammars for lexical selection and grammatical differences. As can be seen from the number of lexical entries, the language pairs range from usable machine translators to early stage projects.

5.3 Neural machine translation

The neural machine translation project *Neurotõlge* (*neurotolge.ee*, see Yankovskaya et al. (2023)) offers

[14] For an overview and links to the keyboards, see *https://giellalt.github.io/KeyboardLayouts.html#uralic-languages*

[15] For a definition of the various grades, see *https://giellalt.github.io/MaturityClassification.html*

Pair	**Entries**
Finnish—Livvi	30,212
Karelian—Livvi	6,419
Finnish—Kven	4,624
Karelian—Finnish	2,297
Vorõ—Estonian	161
Livonian—Finnish	37

Table 5: Machine translation models

Language	**Paradigm info**
Kven	10,557
Livonian	5,693
Livvi	3,538
Meänkieli	1,526
Veps	392
Võro	4,023

Table 6: Paradigm info

machine translation between (among other Uralic languages) the Finnic minority languages Livvi Karelian, Viena Karelian, Lude, Veps, Livonian and Võro and the majority languages Finnish, Swedish, Norwegian Bokmål and Russian. The monolingual corpora presented in Yankovskaya et al. (2023, 765) range from 5,000 (Ludic) to 115,300 and 162,000 (Veps and Võro) sentences. The amount of parallel sentences for the languages in Russia with Russian are 10,000 – 27,000, with the Bible dominating for all languages except Ludic.

Compared to their result for Finnish to Inari Saami and Norwegian to South Saami (which boast the quite good BLEU scores of 67.34 and 60.79, respectively), their results for the Finnic languages (op.cit. p. 768) are far worse (BLEU 24.17 for Estonian to Livonian and 30.63 for Estonian to Võro, the latter even worse than their previous result of 34.11). As shown by Yankovskaya et al. (2023), the main reason for this is the paucity of text, and the lack of balance for the parallel text, for the Finnic languages.

There are some existing critical evaluations of Neurotõlge for Sámi languages, c.f. Wiechetek et al. (2024, 2023), but these evaluations concentrate upon key semantic and grammatical elements of the translated texts rather than the overall closeness between translation and reference, as Yankovskaya et al. (2023) do.

6 Possibilities and perspectives

There are grammatical models for most Finnic minority languages, they show a coverage for running text on around or slightly 90 % (cf. Table 3). This is typical result achieved by rewriting formal grammars as grammar models. Grammars are seldom comprehensive, they typically sketch main patterns and obvious exceptions. In order to go the time-consuming work of getting a coverage of, say, 98 %, one has to include native speakers with knowledge of the norm in the team, so that they can add the description not included in the grammars. It is thus important that language researchers, teachers and learners are included in the process.

One way that the teachers and learners might help, is to simply provide paradigmatic information on word inflection. Providing simple information on a single word *häkki+N+Sg+Ade: häkil*, for example, provides the coder with information on gradation, and an adjacent plural form *häkki+N+Pl+Ade: häkkilöil*. These bits of information can be generated in a class environment where each student is given nouns, verbs or adjectives to describe in paradigms. The teacher checks to see that the forms are correct and the paradigmatic information is added to the infrastructure testing.

The GiellaLT infrastructure provides two different kinds of testing: One is impressionistic testing: Tools that generate parts of the model for the developer to inspect (e.g. generating all forms of a certain case). Another type is regression testing. Here, the linguist has set up for example model paradigms for parts of the morphology, and the model is tested continuously in order to ensure that it does not get worse.

There are test paradigms for the grammatical models of the Finnich minority languages to a various degree. Table 6 gives an overview of paradigm cells in the testing setup for the different languages. The figures might provide us with a picture of the time allocated to developing the different models. One could, of course, also add language-form information to the paradigmatic information, which could help solve problems in Veps, for example, where the Veps magazine *Kodima*[16] and the Veps edition of *Wikipedia*[17] are written in two different orthographies.

There is always a continuum of dialects and languages and standards within these minority lan-

[16] https://omamedia.ru/fi/publication/kodima

[17] https://vep.wikipedia.org

guages, one benefit of rule-based approaches is that they offer good control over the variation: It is possible to implement morphophonological rules and lexical analyses that concern specific variants. When this language technology is combined with a tool like spell-checking and correction, it is a powerful tool for language normativisation and support of writing culture. Experience with Kven has shown that the same lexica and morphological tagging structures can be used for describing language variants by river valley. Applied to Karelian languages, this might allow us to share mutual word stems, on the one hand, but distinguish morphological branches on the other. When it comes to sharing mutual lexica, it should be noted that the shared lexica are set off as their own groups. In work with Saami languages, proper noun lexica are shared. Even here, however, not all proper nouns can be shared. In work with the Permyak-Komi and Zyrian-Komi, additional sharing of lexica has been included for 100% matches in Russian loan words. For the Karelian languages using shared lexica is dependent on the use of parallel phonematic writing practices.

For future work, there is a lot that can be done in curating more lexical data and corpora for these languages. There is also a potential of developing speech technology applications based on the example of existing systems in Sámi languages. All of this requires collaboration, of course, between language communities and computational linguists. An important and ever more relevant issue in collaboration of language communities and computational linguists is ethical issues related to ownership of the language data and language itself, there has been a lot of research on this topic by us and others and we want to point towards (Wiechetek et al., 2024, 2022) for further references.

7 Conclusion

In this article, we have summarised the state of the art in minority Finnic language technology. We have shown that there exist some resources and have compared them to related languages to highlight the potential future possibilities these languages already have available.

The main part of the language technology work on Finnic so far has been concentrated on language models and proofing tools. For 5 of the 9 languages, we have developed grammatical models showing a coverage on running text extending 85 % (for three of them, 90 %).

The situation for available corpora is rather limited. Only for Kven and Meänkieli are there text collections available other than text from (Incubator) Wikipedias. To what extent the content of the corpora follow established standards is unclear. The corpora referred to here do not include all published text, but it is clear that the basis for data-driven language technology is shaky. In this perspective, we note on the positive side that despite this, there is neural-based MT for 5 of the languages presented here.

References

Daniel Abondolo and Riitta-Liisa Valijärvi, editors. 2023. *The Uralic Languages*.

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors. 2022. *The Oxford Guide to the Uralic Languages*, 1 edition. Oxford Guides to the World's Languages. Oxford University Press, Incorporated, Oxford.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39.

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*.

Riho Grünthal. 2022. Veps. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 291–307. Oxford.

Riho Grünthal. 2023. The Finnic languages. In *The Uralic Languages*.

Helsingin yliopisto, FIN-CLARIN, Jack Rueter, and Erik Axelson. 2022. Raamatun jakeita uralilaisille kielille, rinnakkaiskorpus, Korp.

Katri Hiovain-Asikainen and Javier De la Rosa. 2023. Developing TTS and ASR for Lule and North Sámi languages. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 48–52.

Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, and Tamás Váradi. 2022. *Report on the Hungarian Language*. European Language Equality (ELE).

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

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.

Kimmo Koskenniemi. 1983a. Two-level morphology: A general computational model for word-form recognition and production.

Kimmo Koskenniemi. 1983b. *Two-level Morphology. A General Computational Model for Word-forms Production and Generation*, volume 11 of *Publications of the Department of General Linguistics*. University of Helsinki.

Kimmo Koskenniemi, Krister Lindén, Lauri Carlsson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirja Hyvärinen, Imre Bartis, Pirkko Nuolijärvi, and Aino Piehl. 2012. *The Finnish Language in the Digital Age*. Springer.

Johanna Laakso. 2022. Livonian. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 380–391. Oxford.

Johanna Laakso and Elena Skribnik. 2022. Graphization and orthographies of Uralic minority languages. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 91–100. Oxford.

Krista Liin, Kadri Muischnek, Kaili Müürisep, and Kadri Vider. 2012. *The Estonian Language in the Digital Age*. Springer.

Krister Linden and Wilhelmina Dyster. 2022. *Report on the Finnish Language*. European Language Equality (ELE).

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Elena Markus and Fedor Rozhanskiy. 2022a. Ingrian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Elena Markus and Fedor Rozhanskiy. 2022b. Votic. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Sjur Nørstebø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. *The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP*, volume 2 of *NEALT Monograph Series*, pages 70–94. NEALT.

Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård. 2022. *Report on the Nordic Minority Languages*. D1.38. European Language Equality (ELE).

Kadri Muischnek. 2022. *Report on the Estonian Language*, volume D1.12. European Language Equality (ELE).

Karl Pajusalu. 2022. Seto South Estonian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

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. 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.

Helen Plado, Liina Lindström, and Sulev Iva. 2023. Võro South Estonian. In *The Uralic Languages*.

Bengt Pohjanen. 2022. *Meänkieli – Grammatik, lärobok, historik, texter*. Barents Publisher, Överkalix.

T. Rantanen, H. Tolvanen, M. Roose, J. Ylikoski, and O. Vesakoski. 2022. Best practices for spatial language data harmonization, sharing and map creation – a case study of uralic. *PLoS ONE*, 17(6).

Jack Rueter. 2024. Testing and enhancement of language models (transducers) from GiellaLT (scientific blog). *HAL*. 23 pages.

Anna-Kaisa Räisänen, Aili Eriksen, Thomas Brevik Kjærstad, and Trond Trosterud. 2024. Kvensk revitalisering, normering og leksikografi. *LexicoNordica*, 1(31).

Anneli Sarhimaa. 2022. Karelian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Eszter Simon, Piroska Lendvai, Géza Németh, Gábor Olaszy, and Klára Vicsi. 2012. *The Hungarian Language in the Digital Age*. Springer.

Steinþór Steingrímsson, Iben Nyholm Debess, Kimmo Granqvist, Per Langgård, and Trond Trosterud. 2024. *Language Technology for Less-Resourced Languages in the Nordics. Current Development and Collaborative Opportunities*. Stjórnaráð Íslands.

Eira Söderholm. 2017. *Kvensk grammatikk*. Cappelen Damm.

Sindre Reino Trosterud, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto, and Kaisa Maliniemi. 2017. A morphological analyser for Kven.

Trond Trosterud. 2020. Språkteknologi for meänkieli.

Tiit-Rein Viitso and Valts Ernštreits. 2012. *Līvõkīel-ēstikīel-leţkīel sõnārõntõz: = Liivi-eesti-läti sõnaraamat = Lībiešu-igauņu-latviešu vārdnīca*. Tartu Ülikool, and Latviešu valodas aģentūra.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. Unmasking the myth of effortless big data-making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1167–1177.

Linda Wiechetek, Flammie Pirinen, and Per Kummervold. 2023. A manual evaluation method of neural MT for indigenous languages. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 1–10.

Linda Wiechetek, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, and 741 others. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Kildin Saami-Russian-(English) Parallel Corpus Building

Evan Hansen
University of Eastern Finland
Joensuu, Finland
evan.hansen@uef.fi

Abstract

This paper presents two parallel corpora of written Kildin Saami and the process of their compilation. The first, a dictionary corpus, contains 101,889 Kildin Saami tokens of example phrases/sentences from three Russian-Kildin Saami dictionaries and the glossary of the nonfiction book *Saami ornaments*, accompanied by the examples' respective headwords and translations into up to four other languages. Headwords where possible are paired with their underived base, making it a suitable resource for investigating questions surrounding morphological derivation in Kildin Saami. The second corpus comprises 23,884 Kildin Saami tokens and was compiled from *Saami ornaments*, a trilingual (Russian-Kildin Saami-English) book introducing various Saami handicrafts and their creators from across Russian Sápmi.

1 Dictionary Corpus

This first corpus was originally built to facilitate the study of morphological derivation in Kildin Saami in my master's thesis on the subject of the reflexivizer -Эдт- (*-edt-*). A description of the corpus is found in Hansen (2025, 38-43). The work described here aimed at enriching the corpus and comprises the example phrases and sentences featured in three bilingual Kildin Saami-Russian dictionaries: Antonova (2014), Afanas'eva et al. (1985), and Kert (1986), as well as the trilingual Kildin Saami-Russian-English glossary entries from Mozolevskaja and Mechkina (2015).

In all, the corpus contains 21,800 Kildin Saami examples, composed of 101,889 tokens at roughly 4.7 tokens per example. Due to as of yet unnormalized orthographic variation across the source material, it is not yet possible to reliably estimate the number of types or unique headwords, nor can the number of headwords unique to each source be given.

The corpus is currently stored in a private GitHub repository[1].

Table 1: Dictionary corpus contents

Source	Examples
Afanas'eva et al. (1985)	10,160
Antonova (2014)	10,641
Kert (1986)	230
Mozolevskaja and Mechkina (2015)	849

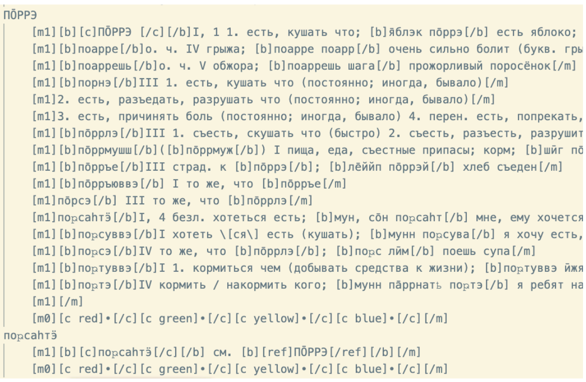
1.1 Corpus structure

The corpus is in .csv format in UTF-8 encoding and contains 15 columns of data. The first five keep track of the source text, the associated headword, the headword in normalized orthography, the headword string in reverse, and the underived base headword of the aforementioned headword. Following these are two columns for the Kildin Saami (normalized orthography and original), then columns for Russian, Finnish, English, and German translations of the Kildin Saami example phrases. Lastly, there are columns for notes, inflectional information of the headword, and English and Russian definitions of the headword.

1.2 Data source selection

The three source dictionaries were chosen for a few reasons, the first being that they were accessible in .dsl format, a type of structured text file that can be used to compile dictionaries. The data being structured in this way made it straightforward to query and extract relevant data when corpus building, and it was additionally possible to load and visualize them simultaneously in the dictionary application Alpus. It was later a simple task to incorporate the data from Mozolevskaja and Mechk-

[1] Inquiries for accessing this resource should be directed to the data admins of the GitHub organization *langdoc*. The corpus is housed in the *sjd-parallel-corpus* repository (https://github.com/langdoc/sjd-parallel-corpus), which is currently set to "private" as it is a work in progress.



```
ПО̄РРЭ
    [m1][b][c]ПО̄РРЭ [/c][/b]I, 1 1. есть, кушать что; [b]я̄блэк по̄ррэ[/b] есть яблоко;
    [m1][b]поарре[/b]о. ч. IV грыжа; [b]поарре поарр[/b] очень сильно болит (букв. гры
    [m1][b]поаррешь[/b]о. ч. V обжора; [b]поаррешь шага[/b] прожорливый поросёнок[/m]
    [m1][b]порнэ[/b]III 1. есть, кушать что (постоянно; иногда, бывало)[/m]
    [m1]2. есть, разъедать, разрушать что (постоянно; иногда, бывало)[/m]
    [m1]3. есть, причинять боль (постоянно; иногда, бывало) 4. перен. есть, попрекать,
    [m1][b]по̄ррлэ[/b]III 1. съесть, скушать что (быстро) 2. съесть, разъесть, разрушит
    [m1][b]по̄ррмушш[/b]([b]по̄ррмуж[/b]) I пища, еда, съестные припасы; корм; [b]шӣг по̄
    [m1][b]по̄ррье[/b]III страд. к [b]по̄ррэ[/b]; [b]ле̄йӣп по̄ррэй[/b] хлеб съеден[/m]
    [m1][b]по̄ррьюввэ[/b] I то же, что [b]по̄ррье[/m]
    [m1]по̄рсэ[/b] III то же, что [b]по̄ррлэ[/m]
    [m1]поҏсаhтӭ[/b]I, 4 безл. хотеться есть; [b]мун, со̄н поҏсаhт[/b] мне, ему хочется
    [m1][b]поҏсуввэ[/b]I хотеть \[ся\] есть (кушать); [b]мунн поҏсува[/b] я хочу есть,
    [m1][b]поҏсэ[/b]IV то же, что [b]по̄ррлэ[/b]; [b]поҏс лӣм[/b] поешь супа[/m]
    [m1][b]поҏтуввэ[/b]I 1. кормиться чем (добывать средства к жизни); [b]поҏтуввэ ӣжя
    [m1][b]поҏтэ[/b]IV кормить / накормить кого; [b]мунн па̄ррнатҍ поҏтэ[/b] я ребят на
    [m1][/m]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
поҏсаhтӭ
    [m1][b][c]поҏсаhтӭ[/c][/b] см. [b][ref]ПО̄РРЭ[/ref][/b][/m]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
```

Figure 1: A screenshot from the .dsl file of Afanas'eva et al. (1985). The upper nest of entries contains по̄ррэ 'to eat' and its derivations. Below is the individual entry for порсаhтӭ 'to be hungry', which includes a reference back to по̄ррэ.

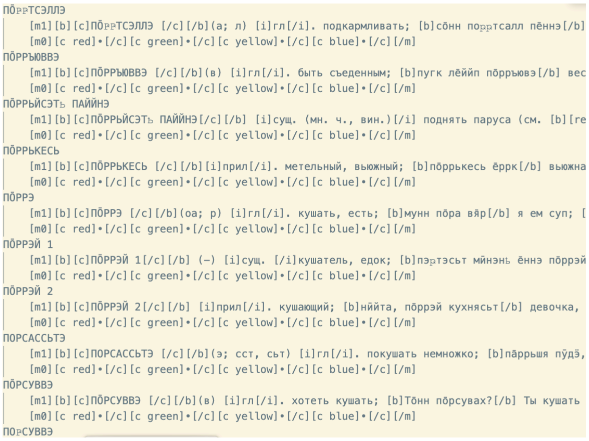
ina (2015), following its inclusion in the corpus described in Section (2). This subset of data was added due to its wealth of example sentences and accompanying headwords.

In the case of Afanas'eva et al. (1985), another benefit is the nested structure of entries, where each headword houses daughter entries (if applicable) for headwords derived from it. Outside of the nests, daughter entries are also each afforded individual entries in alphabetical order, with a reference tag to the underived base form. Figure (1)

This greatly simplified the task of identifying derivational pairs, especially given that the regular morphophonological processes in Kildin Saami that manifest alongside derivational word formation can obscure quite drastically these links. This is because the orthography of words can be affected not just word-finally, which occurs for instance due to consonant gradation, but also at the beginning of a word where vowels often undergo the change of ablaut. We can consider the following base word and its derivations: я̄л (*eal*) 'life', е̄льсуввэ (*jel'suvve*) 'to want to live', ӣллсэ (*iillse*) 'to get on (in life).' The result of such orthographic shifts renders typical alphabetic organization of dictionary entries rather unsuitable for Kildin Saami and for investigations into derivation. Afanas'eva et al. (1985) was further instrumental in identifying derivational pairs for data points from the other three sources, which do not possess a nested structure or reference tags for derived headword entries.

1.3 More about the data sources

In Afanas'eva et al. (1985), the headwords are based on everyday spoken language and cover the many facets of Saami society, culture, and day-to-day living. The dictionary comprises around 8,000 headwords, among which according to the foreword are a certain number of Russian loanwords (Afanas'eva et al., 1985, 9).



```
ПО̄ҏҏТСЭЛЛЭ
    [m1][b][c]ПО̄ҏҏТСЭЛЛЭ [/c][/b](а; л) [i]гл[/i]. подкармливать; [b]со̄нн поҏҏтсалл пе̄ннэ[/b]
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЪЮВВЭ
    [m1][b][c]ПО̄РРЪЮВВЭ [/c][/b](в) [i]гл[/i]. быть съеденным; [b]пугк ле̄йӣп по̄ррьювэ[/b] вес
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЫЙСЭТҍ ПАЙЙНЭ
    [m1][b][c]ПО̄РРЫЙСЭТҍ ПАЙЙНЭ[/c][/b] [i]сущ. (мн. ч., вин.)[/i] поднять паруса (см. [b][re
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЬКЕСЬ
    [m1][b][c]ПО̄РРЬКЕСЬ [/c][/b][i]прил[/i]. метельный, вьюжный; [b]по̄ррькесь е̄ррк[/b] вьюжна
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЭ
    [m1][b][c]ПО̄РРЭ [/c][/b](оа; р) [i]гл[/i]. кушать, есть; [b]мунн по̄ра вя̄р[/b] я ем суп; [
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЭЙ 1
    [m1][b][c]ПО̄РРЭЙ 1[/c][/b] (–) [i]сущ. [/i]кушатель, едок; [b]пэҏтэсьт мӣнэнҍ е̄ннэ по̄ррэй
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РРЭЙ 2
    [m1][b][c]ПО̄РРЭЙ 2[/c][/b] [i]прил[/i]. кушающий; [b]нӣйта, по̄ррэй кухнясьт[/b] девочка,
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПОРСАССЬТЭ
    [m1][b][c]ПОРСАССЬТЭ [/c][/b](э; сст, сьт) [i]гл[/i]. покушать немножко; [b]па̄ррьшя пӯдӭ,
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПО̄РСУВВЭ
    [m1][b][c]ПО̄РСУВВЭ [/c][/b](в) [i]гл[/i]. хотеть кушать; [b]То̄нн по̄рсувах?[/b] Ты кушать
    [m0][c red]•[/c][c green]•[/c][c yellow]•[/c][c blue]•[/c][/m]
ПОҏСУВВЭ
```

Figure 2: A screenshot from the .dsl file of Antonova (2014), showing individual entries for по̄ррэ 'to eat' and others in alphabetical order. Compare with Figure (1), in which there is a noticable nested entry structure. Also important are POS tags (e.g. гл. (rus. 'verb') and information indicating morphophonological changes occurring during inflection (e.g. for по̄ррэ, ablaut through ō (oo) ⇒ оа (å) and consonant gradation with pp (rr) ⇒ p (r)).

In the preface of Antonova (2014, 5), it is stated that the dictionary was compiled for a few reasons. The work was initially to be a simple word list to accompany the 2013 Kildin Saami translation of *Pippi Longstocking* entitled Та̄рьенч Кукесьсуххк (*Taar'jenjč Kukessuhk*), translated by Aleksandra Antonova from the Russian version (Lindgren, 2014). There was a need for this word list due to the fact that a significant portion of the words featured in the translation were absent from the dictionaries available at the time. These dictionaries were further asserted to be not widely accessible. It was then decided that a new dictionary was necessary, not only for making the Kildin Saami translation of *Pippi Longstocking* more accessible, but also to provide the language community with a more comprehensive dictionary that includes practical vocabulary for everyday communication. The lexicography team additionally sought to incorporate terms that are highly relevant to Saami culture. The dictionary comprises around 8,000 headwords.

The later dictionary Antonova and Scheller (2021–) building upon Antonova (2014) features many spelling corrections, as well as general spelling convention normalizations based on Afanas'eva et al. (1985) (e.g. ли (*li*) ⇒ лӣ (*lii*) '[3SG] is' and āммьсе (*aamm'c'e*) ⇒ āммьсэ (*aamm'ce*) 'to yawn'). The dictionaries appear to share example sentences/phrases. In any case, data from this later work are not included in the corpus as no API or other tool are available to facilitate their extraction.

Kert (1986, 5) in its turn is a bidirectional bilingual Kildin Saami-Russian dictionary, composed of 4,000 headwords. It is described as being based on the vocabulary used in primary school textbooks, also including cultural terms relevant to Saami life. Russian loan words, with some exceptions, are excluded.

From Mozolevskaja and Mechkina (2015), 674 headwords and 849 examples were extracted. See Section (2) for more information about the text.

1.4 Dictionaries as data sources

When considering the data, the prescriptive nature of dictionaries should be kept in mind. The examples and phrases in the data were originally constructed by the authors (an exception being proverbs, though these do exhibit a certain degree of variation based on preliminary observations) and provide the reader with examples of how the words can be used in context. The source texts assert other information as well, such as how a given word declines. As such, questions of language variation and frequency are not feasible.

In addition to the above, a further consideration is that with the latest source publication (Mozolevskaja and Mechkina, 2015) being from 2015, there is roughly a decade of language change that is not featured.

1.5 Corpus compilation tools

The corpus building process largely took place in the Visual Studio Code (VSCode) application. Python scripts facilitated the work with the data, at times incorporating the libraries pandas, csv, deepl, and re. Some of VSCode's build-in functions like the find-and-replace tool also streamlined the process. The dictionary file reader Alpus was further utilized to cross-check the intended structure of the source dictionary entries as needed. Alpus was useful also in the way that it could query the dictionaries simultaneously, and certain orthographic discrepancies between the works would be ignored.

1.6 Data preprocessing

To extract the required data from the .dsl[2] dictionary files, it was mainly necessary to modify the organizational structures used to house the data. These were by-and-large strings that resemble HTML tags (for instance '[b]' and '[/b]') and were perhaps inserted by the original author for formatting the output. At times these disrupted the extraction process, sometimes cutting through data strings, or there were errors where tags were left open. A further complication was the fact that tags enveloping data also surrounded items like headwords or their inflected/declined forms. These difficulties aside, the tags also helped define boundaries for what needed to be extracted from the files.

Slight modifications to the code were needed as well due to some variations in structure between the dictionaries.

1.7 Data Extraction

When it came to the stage of extracting data, a list of headwords and their possible underived forms from Afanas'eva et al. (1985) was created. Using this list of pairs as a starting point, the examples from Antonova (2014) and Kert (1986) were extracted; as entries in these dictionary do not have references to their underived base words, it was possible to supplement these from pairs collected from Afanas'eva et al. (1985). It is however the case that the dictionaries use variations of the contemporary Cyrillic writing script employed for Kildin Saami, and it is possible that a few incorrect pairings were created across the dictionaries. Among the orthographic discrepancies are those relating to vowel length (e.g. т|оā|гктэдтэ (Antonova, 2014) ⇒ т|оа|гктэдтэ (Afanas'eva et al., 1985) and preaspiration (e.g. суэ|хх|птэдтэ (Antonova, 2014) ⇒ суэ|h|птэдтэ (Afanas'eva et al., 1985).

These and other orthographic differences appear to be systematic for the most part but for the moment still have yet to be resolved.

Having established these headword-underived base pairs, they were matched up with their associated example phrases and sentences. The Kildin Saami examples were collected along with their Russian counterparts. The dictionary of origin was also recorded.

Later when revisiting the corpus for making im-

[2] These are freely available for download from https://slovari.saami.su/.

provements, all headwords *without* an associated example were added and paired with their underived base where possible.

1.8 Secondary translations

In an effort to make the data more accessible, machine translation was carried out on the data into English, Finnish, and German, as many publications related to Kildin Saami have been written in these languages. These translations were effectuated based on the Russian translations of the Kildin Saami entries given in the source dictionaries. The translations were made by means of the DeepL API Pro service and deepl Python library.

Though DeepL is one of the most robust machine translation services available, the quality and consistency of the translations appears to be not extremely reliable; this is regrettable, though it can be a helpful supplement to have access to when trying to quickly scan through the data.

As Kert (1986) and Mozolevskaja and Mechkina (2015) were not added to the corpus until later, they do not feature secondary translations.

1.9 Data postprocessing

At this stage, many errors were quite visible for the dictionary data in the way that for instance a tabulation error would generate an incorrect number of .csv columns across a block of data rows, and this could easily be detected both through even a brief glance and with the help of VSCode's built-in .csv error identifying features. Some errors also took the form of duplicate lines. The majority of large errors such as these were resolvable simply by using the find-and-replace tool in VSCode. Others required a bit more time and attention, where for example Russian data had somehow gotten into the Kildin Saami example column. Aside from these errors, some last escape characters and structural strings were removed.

One last postprocessing step concerned the secondary translations. With these certain translations that had for some reason been left blank by the API tool were redone. Another issue was related to grammatical and other tags in Afanas'eva et al. (1985) using abbreviated terminology. There are several dozen of these, such as страдательный залог 'passive voice' which appears as страд. in the dictionary entries. The Russian example data had been preprocessed by replacing the abbreviations with the corresponding expanded forms and enclosing them in asterisks, though this for many examples interfered with the translation done by the DeepL API. In total however, only around 500 of almost 21,000 examples contain one of these abbreviations, though not all translations have been addressed yet.

1.10 Future improvements

As mentioned in the above subsections, a notable issue with the corpus data is the lack of normalization of typos and of differences in orthography that can be found throughout the four source texts. This can perhaps first be fairly easily remedied for most cases of variation by targeting the *headwords* column, thereby making it possible to identify more underived roots for the headwords of those data points not originating from Afanas'eva et al. (1985). It would be most ideal however to eventually normalize the tokens in the examples as well, which would make it possible to explore derivations through corpus searches that include inflected forms.

A second improvement would largely affect the data originating from Antonova (2014). Among the headwords in this dictionary are entries for inflected forms of other headwords. These examples would be reassociated with the uninflected form.

A third key change for the future would be to assign at least one example sentence/phrase to those headwords in the data without one. These examples could in principle be copied from elsewhere in the data, as some headwords without examples do appear in those of others.

Next, the corpus will be continually enriched with new headwords as they are found. Kert (1988, 91) for instance includes the word рӯппсесьсиагаш (*ruupps'es'silagaš*) 'reddish' (from рӯппсесь (*ruupps'es'*) 'red'), which is not found in the corpus, nor is its adjectivizing suffix -сиагаш (-*silagaš*) '-ish' in combination with any other base. This is not the only attested derivational suffix missing from the corpus data, and it is crucial that such forms be represented for the corpus to be useful as a resource for exploring word-forming morphology in Kildin Saami.

Even within the corpus itself are words that are not given their own entries as headwords. The Russian loanword бутерброд (*but'erbrod*) 'sandwich' (same spelling and meaning in Russian) for example is used 10 times in the data but has no dedicated entry.

Finally, it should be mentioned that the data in the example sentences/phrases columns do not feature morphosyntactic tags. The data can already be queried fairly easily through RegEx searches, though inflected forms, in particular of underived words, remain a large hurdle. One possible avenue for data tagging would be to utilize GiellaLT's FST-analyser, although it is unclear whether this tool is available for research outside GiellaLT.[3]

2 *Saami ornaments* corpus

The second corpus at hand draws its contents from the 2015 publication entitled *Saami ornaments* (rus. Саамские узоры; sjd.[4] Са̄мь кырьйнэз), a trilingual book that showcases Saami handicrafts from across 15 Saami siidas (sjd. сыййт, *syjjt*) within Russian Sápmi. In all, the corpus features 23,884 Kildin Saami tokens aligned with the Russian and English parallel texts. According to the front matter of the book, the original text is the Russian, which was subsequently translated into the Kildin Saami and English (Mozolevskaja and Mechkina, 2015, 2). The corpus is currently stored in a private GitHub repository (see [1]).

A significant portion of the data comes from the book's glossary, from which a total of 674 headwords and 849 examples were extracted. Of the examples, 55 are proverbs. What is more, entries appear to follow the inflection categorization system used in Afanas'eva et al. (1985), also borrowing certain example phrases. The glossary does however diverge at times from the aforementioned dictionary in orthography (in particular when it comes to long and short vowels), as well as with the inclusion of certain terms not found within Afanas'eva et al. (1985) or in the other two dictionaries referenced in Section (1).

2.1 Data structure

The corpus comprises 3,640 rows of data in .csv format, in UTF-8 encoding. There are two columns for the Kildin Saami data, one for the original data modified and the other for normalized orthography. Aligned with these are columns for the Russian and English parallel textual data. Further columns include one for notes, which mainly records the notes left in the margins of the book by those involved in the revising and editing process, as well as two columns for the sections and subsections to which the data belong. The sections reflect the overarching sections of the book, for instance the various Saami siidas and the "About the authors" portion. Subsections largely follow the headers for the many handicraft items in the book. Duplicate subsections within the same section are numbered. The final column is for I.D. numbers, which were added in order to keep track of the original internal structuring of the book.

2.2 Corpus building

Thanks to the availability of an OCR'ed copy of the book, as well as the book's highly structured contents with overall consistent use of (sub)headers and bulleted lists, the process of compiling this corpus was relatively straightforward.

The raw data first were copied and pasted into a .txt file section by section, keeping parallel language sections together. The data were then preprocessed using a Python script in order to have one sentence of data per line in the .txt file. Frequent problem strings such as '1.' and abbreviations like 'мн.' (rus. 'plural') were accounted for in the script to simplify the process. The data were then aligned in VSCode using the *Edit CSV* extension created by *janisdd*. The extension provides an Excel-like interface for reading and writing .csv files, allowing for simple data manipulation.

Concerning the alignment of the three versions, sentential alignment was prioritized. Instances where multiple sentences in one language mapped to one sentence in another were aligned as one data point, avoiding any divisions of single sentences. Rarely, and as necessary, sentences were rearranged for alignment.

Image captions and "Master of Sami Handicraft" boxes (sections providing the name of the craftsperson) were excluded for the most part, given their repetitive structure and redundant information.

2.3 Postprocessing

Once all data had been compiled, the data were checked for lookalike character replacements (e.g. the Latin <ä> (U+00E4) for the Cyrillic <ӓ> (U+04D3)) and other nonstandard characters. The biggest modification replaced the combining overlines (U+0305; e.g. <я̅>) used in the book to indicate a long vowel, this in place of the standard orthography's combining macron (U+0304; e.g. <я̄>).

[3] The imprint of GiellaLT's digital Kildin Saami dictionary mentions the existence of an automaton for paradigm generation, see https://sanj.oahpa.no/about/. But the GiellaLT infrastructure offers only an embryonic version, see https://github.com/giellalt/lang-sjd.

[4] Kildin Saami.

2.4 Preliminary observations

A notable aspect within the glossary is the inclusion of Varzino (Arsjogg) dialectal varieties for certain headwords. Some of these are given in Table (2). Note that the “standard”[5] variants are taken from Antonova (2014).

A further point of interest within the glossary are certain terms which are not listed in the three dictionaries included in the dictionary corpus outlined in Section (1). Among these are Е̄кесь Та̄ссьт (*iekes’ taass’t*) ‘Mars’, га̄рэс *(gaares)* ‘worsted yarn’, and быдтъесь *(bydtjes’)* ‘necessary’.

2.5 Future improvements

As of yet, the text in the normalized orthography data column has not undergone any normalization for orthographic variation or for possible typos, though this certainly is planned for the near future to allow for more streamlined corpus queries across multiple corpora; in this same vein, the corpus will soon be converted to XML format, as this has been preferred by the Kola Saami Documentation Project for such materials [6].

Table 2: Comparison of Standard and Arsjogg variants (NOM.SG)

Standard	Varzino (Arsjogg)	Translation
нэ̄дт	нэ̄ввт	’handle’
ка̄ӆӆт	ка̄ӆт	’pocket’
кыффкэмкарь	кыјјм-ка̄рръ	’mirror’
(or: кыйхмкарь)	”	
коалль	коалльт	’gold’

Certain segments of the data need to be revisited for alignment modifications. Though overall it was possible to identify correspondences across the three language versions, certain sentences appear to have been skipped over during the translation process. To some extent these unclear sections may be due to the corpus compiler’s competencies in Russian and Kildin Saami, which include structural knowledge and a novice L2 proficiency level in both languages.

3 Corpora applications

Turning first to the dictionary corpus, the structure makes it ideal for investigating questions related to derivation in Kildin Saami.[7] As derivation in the language is almost exclusively accomplished through suffixation, a key strength of the corpus is its column for headwords spelled in reverse order. Querying the data in a spreadsheet viewer, the rows of data can be sorted with this column to quickly identify all instances of a given suffix when it is the final derivation. This reversed spelling column being based on the normalized headword column further makes it possible to gather headwords with suffixes that vary orthographically by source (e.g. -аһтӭ and -аххьтэ; -нэх(х)ьк and -нэһкь).

Additionally, the ‘root’ column has been populated for many data points from Antonova (2014); Mozolevskaja and Mechkina (2015); Kert (1986) using derivational pairs created from Afanas’eva et al. (1985). This establishment of derivational relations for these data from other sources then saves the corpus user the time of having to identify the connections manually, not only for the parent-descendent pairings but also for the links between sibling headwords with a shared root word. The original nested headword structure from Afanas’eva et al. (1985) in effect is broadened, opening up opportunities to examine entire families of headwords side-by-side and pulling from multiple sources at once. This is particularly useful when comparing a headword carrying two or more layers of derivational suffixation with its possible intermediate headwords (e.g. шэ̄ннтэ (*šeennte*) ‘to grow’ ⇒ шэ̄ннтлэ (*šeenntle*) ‘to grow a little, grow quickly’ ⇒ шэ̄ннтлуввэ *šeenntluvve*) ‘to start to grow, get taller’) or when considering aspectual derivations (e.g. шэ̄ннтлэ (*šeenntle*), шэ̄ннтассьтэ (*šeenntass’te*), шэ̄нтсэ (*šeentse*) ‘to grow quickly’; шэнтнэ (*šentne*) ‘to grow continuously’).

Aside from derivation, patterns in Kildin Saami-Russian translation could be explored using the Russian columns. To investigate translation strategies involving desiderativity for instance, RegEx could be employed to return all rows of data with the Russian verb хотеть ‘to want’ and its variations. Important would be to also include the column containing Russian definitions of the Kildin Saami headwords, as not all have an accompanying Kildin Saami-Russian example phrase pair.

Through the incorporation of new headwords over time and more thorough orthographic normal-

[5] “Standard” here refers to the language as it is used in Afanas’eva et al. (1985), Antonova (2014), Kert (1986),and Mozolevskaja and Mechkina (2015). It is not explicitly stated which varieties are used in these texts.

[6] A description of the early stages of this project is found in Rießler and Wilbur (2007); see also Rießler (2024).

[7] My many thanks to the two anonymous reviewers who suggested including this section.

izations, the corpus will become a reference tool of increasing usage potential for those who develop community-facing resources, from spell-checkers to pedagogical materials. Lexicographers especially may benefit from the spelling normalization columns when deciding which forms to include in a dictionary and possibly even list multiple variations for users' ease of access and reduced prescriptivism. We can take as an example how Antonova (2014) features ōннъюввэ (*oonnjuvve*) while Mozolevskaja and Mechkina (2015) additionally uses оаннъюввэ (*ånnjuvve*), both meaning 'to be used' and derived from the verb оаннӭ (ånn'e) 'to use' though using a different stem.

As for the *Saami ornaments* corpus, the source material was selected for compilation with the intention of increasing the amount of multilingual parallel corpus data available for Kildin Saami. Particular to the source publication is its trilingual parallel structure, which makes it quite accessible to English-speaking researchers who have little to no proficiency in Russian and/or Kildin Saami. The corpus with its three parallel versions can serve as a starting point for questions pertaining to translation. Relatedly, quite many Russian loanwords are present in the data, some of which are not found in the dictionary corpus, like этнографическэ (*etnografičeske*) 'ethnographic'; focused studies on Russian borrowings may benefit from incorporating these data points into their research.

What is more, the data are valuable for their representation of contemporary nonfiction written language and their subject area of Saami handicrafts. With these attributes in mind, directions of research using the corpus could include analyses of vocabulary in relevant semantic domains (e.g. colors, materials, crafting tools/techniques) and language change through comparisons with source materials from other time periods. Researchers from the fields of literary studies, history, and anthropology may also have interest in the data.

4 A note on copyright

The two corpora described in this paper derive their contents from source texts that are protected under copyright. In principle, copyright laws within the European Union permit the use of the source materials for use in academic research; this includes converting them to a digital format, storing them, and processing them as is typically done when mining textual data.[8] It is furthermore permissible to conduct research with such materials in collaboration with other researchers.

Relatedly, fragments of the source material may be published in order to illustrate the data in teaching and scientific publication contexts. However, data extracted from material protected under copyright may not be made freely available, and for this reason the corpora are stored in a private repository.

Ideally, the two corpora would eventually be made more freely accessible, which could be accomplished with permission from the legal owners of the data.

References

Nina E. Afanas'eva, Aleksandra A. Antonova, Boris A. Gluchov, Lazar' D. Jakovlev, and Ekaterina I. Mečkina. 1985. *Saamsko-russkij slovar' = [Sām'-rūšš soagknehk'] = [Saami-Russian dictionary].* Russkij jazyk.

Aleksandra A. Antonova. 2014. *Saamsko-russkij slovar' =[Saami-Russian dictionary].* ANO Arktičeskij centr naučnich issledovanij i ėkspertiz.

Aleksandra A. Antonova and Elisabeth Scheller. 2021–. *Saamsko-russkij i Russko-saamskij slovar' =[Saami-Russian and Russian-Saami dictionary].* UiT The Arctic University of Norway.

Evan Hansen. 2025. Kildin Saami *-edt-* Reflexivized Verbs. Master's thesis, University of Eastern Finland.

Georgij Martynovič Kert. 1988. Slovoobrazovanie imen v saamskom jazyke =[Word formation of nouns in the Saami language]. In Georgij Martynovič Kert, editor, *Pribaltijsko-finskoe jazykoznanie. Voprosy leksikologii i grammatiki*, Trudy Karel'skogo Filiala Akademii Nauk SSSR, pages 84–91. Karelskij filial AN SSSR.

Georgij Martynovič Kert. 1986. *Slovar' saamsko-russkij i russko-saamskij =[Saami-Russian and Russian-Saami dictionary].* Prosveščenie.

Astrid Lindgren. 2014. *Taar jenjč Kukessuhk [Pippi Longstocking].*

Anastasija E. Mozolevskaja and Ekaterina I. Mechkina. 2015. *Saamskie uzory =[Saam' kyr'jnez] =[Sami ornaments].* Drozdov-na-Murmane.

[8] The EU Directive 2019/790 on Copyright in the Digital Single Market outlines copyright exceptions that allow for text and data mining for scientific researcher purposes. This Directive is reflected in national laws within the EU.

Michael Rießler. 2024. Kola Saami Christian Text Corpus. In Mika Hämäläinen, Flammie Pirinen, Melany Macias, and Mario Crespo Avila, editors, *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 138–144. ACL.

Michael Rießler and Joshua Wilbur. 2007. Documenting the endangered Kola Saami languages. In *Språk og språkforhold i Sápmi*, pages 39–82.

SampoNLP: A Self-Referential Toolkit for Morphological Analysis of Subword Tokenizers

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

Ekaterina Chelombitko
DataSpike
Dubai, UAE
ekaterina@dataspike.io

Aleksey Komissarov
aglabx
Paphos, Cyprus
ad3002@gmail.com

Abstract

The quality of subword tokenization is critical for Large Language Models, yet evaluating tokenizers for morphologically rich Uralic languages is hampered by the lack of clean morpheme lexicons.

We introduce SampoNLP, a corpus-free toolkit for morphological lexicon creation using MDL-inspired Self-Referential Atomicity Scoring, which filters composite forms through internal structural cues - suited for low-resource settings.

Using the high-purity lexicons generated by SampoNLP for Finnish, Hungarian, and Estonian, we conduct a systematic evaluation of BPE tokenizers across a range of vocabulary sizes (8k–256k). We propose a unified metric, the Integrated Performance Score (IPS), to navigate the trade-off between morpheme coverage and over-splitting. By analyzing the IPS curves, we identify the "elbow points" of diminishing returns and provide the first empirically grounded recommendations for optimal vocabulary sizes (k) in these languages. Our study not only offers practical guidance but also quantitatively demonstrates the limitations of standard BPE for highly agglutinative languages. The SampoNLP library and all generated resources are made publicly available[1].

1 Introduction

The performance of subword tokenization algorithms like Byte-Pair Encoding (BPE) (Sennrich et al., 2016) is a cornerstone of modern Natural Language Processing (NLP). While highly effective for many languages, their purely statistical nature poses a significant challenge for morphologically rich, agglutinative languages (Bostrom and Durrett, 2020; Rust et al., 2021). In the Uralic family, a group of languages known for its complex morphology and diverse linguistic phenomena (Hämäläinen, 2019), words are often long concatenations of morphemes (e.g., Finnish talo-i-ssa-ni-ko-kaan - "not in my houses either?"). For such languages, the quality of tokenization is not just an engineering detail but a critical factor that determines a model's ability to grasp grammatical structure and generalize effectively (Hämäläinen et al., 2021; Gerz et al., 2018). This raises a pressing, yet under-explored, practical question, known to be a challenge in Uralic NLP: What is the optimal tokenizer vocabulary size (k) to achieve robust morphological representation? The importance of this question was highlighted by recent work demonstrating the benefits of specialized tokenizers for these languages (Chelombitko and Komissarov, 2024).

Addressing this question reveals a more fundamental problem: the scarcity of high-purity morphological resources for evaluation. While lexical data is available in spell-checking dictionaries, their raw combination of stems and affixes results in a noisy candidate list. Manual curation is not scalable, and established corpus-based methods like Morfessor (Creutz and Lagus, 2007) are ill-suited for the many low-resource Uralic languages (Arkhangelskiy, 2019).

To address this challenge, we present SampoNLP, a toolkit based on a corpus-free and self-referential pipeline for refining morphological lexicons. The proposed method, "MDL-inspired Self-Referential Atomicity Scoring," draws its theoretical motivation from the Minimum Description Length principle (Rissanen, 1978), but adapts it to a type-only setting. The core algorithm iteratively estimates the atomicity of each candidate, distinguishing between simple and composite forms by analyzing internal structural patterns within the dataset itself. This lightweight and reproducible approach offers a practical way to produce cleaner morphological resources, a recognized need for data-scarce environments where traditional corpus-based methods are not viable (Hämäläinen, 2019).

[1] https://github.com/AragonerUA/SampoNLP

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 57–67
December 10-12, 2025 ©2025 Association for Computational Linguistics

Having established a robust methodology for resource creation, we leverage our generated lexicons to address the core problem of this paper: the vocabulary-morphology trade-off inherent in BPE tokenization (Bostrom and Durrett, 2020). We conducted a systematic evaluation of BPE tokenizers for Finnish, Hungarian, and Estonian across vocabulary sizes from 8k to 256k. The development of novel evaluation frameworks that go beyond downstream performance is a growing area of research (Chelombitko et al., 2024). In line with this, to precisely navigate the aforementioned trade-off, we introduce the Integrated Performance Score (IPS), a single metric that balances Lexical Morpheme Coverage (LMC) against the Over-Split Rate (OSR). This allows us to model the performance curve and identify the optimal vocabulary range, providing a principled answer to our central research question.

Our contributions are thus twofold and equally significant:

1. **A Corpus-Free Morphological Method:** We introduce a fully automatic and reproducible pipeline for refining morphological lexicons without relying on corpus frequencies or external resources, released as an open-source toolkit, *SampoNLP*.

2. **A Quantitative Evaluation:** We conduct a systematic analysis of BPE tokenizers for Finnish, Estonian, and Hungarian, examining how vocabulary size affects morphological granularity through newly defined metrics of coverage and over-segmentation.

2 Related Work

The evaluation and optimization of subword tokenization for morphologically rich languages intersects several research areas: subword tokenization algorithms, unsupervised morphological analysis, rule-based analyzers, and language-specific NLP for Uralic languages.

2.1 Subword Tokenization and Morphology

Byte-Pair Encoding (BPE) (Sennrich et al., 2016) has become the de facto standard for subword tokenization in modern NLP. Alongside it, methods like the Unigram Language Model (Kudo, 2018) have been proposed, but the purely statistical nature of these approaches presents well-documented challenges for morphologically complex languages. The work of (Bostrom and Durrett, 2020) demonstrated that BPE tokenizers often fail to align with linguistic morpheme boundaries. Interestingly, parallel challenges in identifying meaningful subsequence units have been explored in domains beyond NLP, such as the tokenization of biological sequences like primate genomes (Popova et al., 2025).

The question of optimal vocabulary size has often been guided by heuristics or evaluated indirectly via downstream task performance (Mielke et al., 2021). Our work directly addresses this gap by proposing a methodology for intrinsic, morphologically-grounded evaluation to provide data-driven recommendations for Uralic languages.

2.2 Unsupervised Morphological Analysis

The unsupervised discovery of morphological structure has a rich history. One major family of approaches relies on statistical cues from corpora to identify boundaries. Classic methods such as Branching Entropy and Accessor Variety (Chen et al., 2004) analyze the predictability of subsequent characters to hypothesize morpheme breaks. Another prominent family of methods is based on the Minimum Description Length (MDL) principle. Morfessor (Creutz and Lagus, 2007) and its variants represent the canonical probabilistic approach, finding a lexicon that best compresses a text corpus. While successful, these methods are fundamentally corpus-based, requiring token frequency information that may not be available in low-resource settings.

Our approach, while MDL-inspired, operates in a corpus-free, type-only regime. It represents a different paradigm: self-referential filtering of a candidate list. By operating purely on the internal structure of a candidate set, we provide a lightweight method suited to resource-scarce scenarios, a persistent challenge in Uralic NLP (Arkhangelskiy, 2019).

2.3 Rule-Based Analyzers and Tokenization for Uralic Languages

For Uralic languages, rule-based morphological analyzers built on Finite-State Transducers (FSTs) like Omorfi (Pirinen, 2015) and the GiellaLT[2] infrastructure (Jauhiainen et al., 2020) are invaluable resources. While their generative outputs are linguistically comprehensive, they are not directly optimized for use as a minimal reference morphemes lexicon. Our IMDP pipeline offers a contrasting

[2] https://giellalt.github.io/

approach: a data-driven methodology for distilling such a lexicon from a type-only candidate list, as can be extracted from dictionary-based resources like Hunspell, without requiring token frequencies from a corpus.

The challenge of effective tokenization for this language family has recently gained significant attention. Broader findings have established that language-specific modeling is crucial for morphologically rich languages, with studies on Finnish demonstrating clear benefits of monolingual models like FinBERT over multilingual ones (Virtanen et al., 2019). Building on this principle, a recent study by (Chelombitko and Komissarov, 2024) specifically addressed the severe underrepresentation of Uralic languages in large multilingual models. They demonstrated that training specialized, large-vocabulary monolingual tokenizers yields substantial improvements in compression efficiency. However, while establishing the need for specialized resources, their work left the question of how to determine an optimal vocabulary size open for future investigation.

Concurrently, the need for better evaluation metrics has become a prominent research topic. The Qtok framework (Chelombitko et al., 2024), for instance, proposed a comprehensive approach to evaluating multilingual tokenizer quality, while other studies have also advocated for moving beyond downstream task performance towards more intrinsic, linguistically-informed measures (Beinborn and Pinter, 2023). Our Integrated Performance Score (IPS) directly addresses this call from the community for more morphologically-grounded metrics.

Our current work builds on these foundations. It utilizes similar high-quality data sources as those in (Chelombitko and Komissarov, 2024) to train the tokenizers being evaluated. Furthermore, by proposing a concrete methodology, it answers the call for better evaluation and finds the optimal vocabulary sizes that the former study alluded to, thus providing a logical next step in this line of research.

3 Methodology. The IMDP Pipeline

To create a high-purity morpheme lexicon from a noisy, raw list of candidate forms, we propose the Iterative Morphological Decomposition Pipeline (IMDP). Our approach is designed to be fully automatic and operates in a corpus-free, type-only regime, requiring only the candidate list as input. The core of the pipeline is a method we term "MDL-inspired Self-Referential Atomicity Scoring," which iteratively evaluates how "fundamental" each candidate is relative to the entire set. The entire process is visualized in Figure 1.

The pipeline consists of three main stages: (1) Prefiltering and Initial Scoring, (2) Iterative Score Refinement, and (3) Final Filtering via Automated Thresholding.

3.1 Stage 1: Candidate Pre-filtering and Initial Scoring

This initial stage aims to drastically reduce nonlinguistic noise and establish a baseline score for each plausible candidate.

3.1.1 Hard Pre-filtering

First, we apply a series of deterministic filters to the raw input list C_{raw}. A token $t \in C_{raw}$ is discarded if it:

1. Contains symbols from a non-target script (e.g., Cyrillic in a Latin-based list). We define a valid character set Σ for each language (e.g., [a-záéíóöőúüű] for Hungarian).

2. Contains any non-alphabetic characters (e.g., numbers, punctuation, URLs), excluding initial/final hyphens used to mark affixes.

3. Is a proper noun or acronym (heuristic: starts with a capital letter or consists of multiple uppercase letters).

4. Is excessively long ($|t| > 30$) or too short ($|t| < min\_length$), unless t is a single character present in a language-specific whitelist of valid one-character morphemes W.

3.1.2 Type-support Filtering

To filter out typographical errors and other singleton noise, we apply a "type-support" criterion to the remaining set of candidates C'. A candidate $t \in C'$ is kept only if it appears as a substring in at least m other unique candidates in C'. This ensures that we only consider patterns that are structurally recurrent within the dataset itself. $support(t) = |\{c \in C' | t$ is a substring of $c\}|$ We retain t if $support(t) \geq m$ (we use $m = 3$). The resulting set is our final candidate pool C.

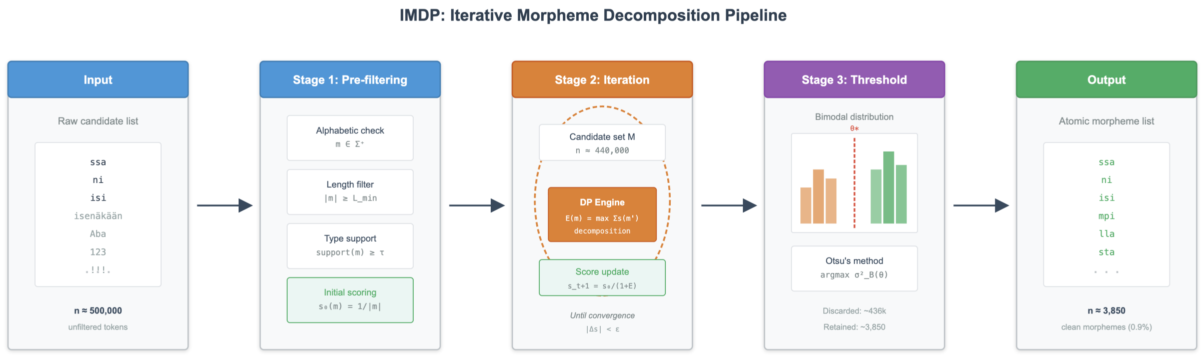


Figure 1: An overview of the Iterative Morphological Decomposition Pipeline (IMDP).

3.1.3 Initial Atomicity Scoring

Each surviving candidate $t \in C$ is assigned an initial Atomicity Score $S_0(t)$. This score is based on the MDL-inspired principle that, all else being equal, shorter forms are more likely to be fundamental morphemic units. The score is defined as the inverse of the token's length: $S_0(t) = \frac{1}{|t|}$, where $|t|$ is the number of characters in t.

3.2 Stage 2: Iterative Score Refinement

This is the core of our method. We iteratively refine the Atomicity Scores until they converge. In each iteration $k + 1$, the score of every token $t \in C$ is re-calculated based on its "explainability" by other tokens in the set.

3.2.1 Optimal Decomposition and Best Explanation Power (BEP)

For each token t, we find its optimal decomposition into a sequence of smaller tokens $(m_1, m_2, ..., m_n)$ where each $m_i \in C$. The optimal decomposition is the one that maximizes the sum of the scores of its constituents (taken from the previous iteration, S_k). We find this maximum sum using a dynamic programming algorithm and term it the Best Explanation Power, $BEP_k(t)$.

$$\mathrm{BEP}_k(t) = \max_{\substack{t=m_1\cdots m_n \\ n\geq 2}} \sum_{i=1}^{n} S_k(m_i).$$

The search space for decompositions is constrained by two rules:

1. **Multi-component:** The algorithm considers segmentations into any number of parts, not just two.

2. **Degeneracy Prevention:** Segments of length 1 are only considered if they are in the whitelist W.

3.2.2 Score Update Rule

The new score $S_{k+1}(t)$ is calculated by comparing the token's own score with its explainability. A token is penalized only if the "evidence" for it being composite $(BEP_k(t))$ is stronger than the evidence for it being an atom $(S_k(t))$.

$$S_{k+1}(t) = \begin{cases} S_k(t), & \text{if } \mathrm{BEP}_k(t) \leq S_k(t), \\ \dfrac{S_0(t)}{1 + \mathrm{BEP}_k(t)}, & \text{if } \mathrm{BEP}_k(t) > S_k(t). \end{cases}$$

This update rule creates a competitive dynamic where atomic morphemes retain high scores, while composite words are iteratively penalized towards zero.

3.2.3 Convergence

The iterative process continues until the system reaches a stable state. We define convergence as the point where the maximum absolute change in any token's score between two consecutive iterations falls below a small threshold

$$\max_{t\in C} \left| S_{k+1}(t) - S_k(t) \right| < \varepsilon$$

We use $\varepsilon = 1e - 7$ and a safeguard limit of $max\_iterations = 100$.

3.3 Stage 3: Final Filtering via Automated Thresholding

After the scores converge, the final distribution of scores typically shows a heavy concentration of composite candidates at very low scores, while atomic candidates retain higher scores. To automatically and reproducibly determine a separation threshold between these groups, we employ Otsu's method (Otsu, 1979). Originally developed for image processing to separate foreground from background, this algorithm finds an optimal threshold τ

for a distribution by maximizing the inter-class variance between the two resulting classes (in our case, "atomic" vs. "composite"). This data-driven approach avoids manual parameter tuning and adapts to the specific score distribution of each dataset.

All tokens t with a final score $S_{final}(t) >= \tau$ are classified as atomic and form our final, high-purity morpheme lexicon.

Lang	Initial Cands	Atomic Morphs	Reduct %	Reduct Factor
Fin	499,647	3,850	99.23%	129.8x
Est	281,256	5,705	97.97%	49.3x
Hung	103,317	3,189	96.91%	32.4x

Table 1: Efficiency of the IMDP pipeline in cleaning and reducing morpheme candidate lists.

4 Experimental Setup

To evaluate the impact of vocabulary size on morphological coverage, we conducted a systematic analysis for three Uralic languages: Finnish, Hungarian, and Estonian. Our experimental setup consists of three main stages: creating the reference morphemes, training the tokenizers, and defining the evaluation metrics.

4.1 Data

Our methodology requires two types of data for each language: a raw list of morpheme candidates for cleaning and a large text corpus for tokenizer training.

1. **Morpheme Candidate Lists:** The initial "dirty" lists of candidates were constructed from authoritative, open-source spell-checking dictionaries based on the Hunspell framework[3]. For Hungarian and Estonian, we utilized the comprehensive dictionaries curated by The LibreOffice Project[4]. For Finnish, which requires special handling of compounds, we used the dedicated dictionary from the hunspell-fi project[5]. For each language, the full set of unique stems (from.dicfiles) and affixes (from.afffiles) was merged to create a comprehensive but structurally noisy candidate list, which serves as the input to our IMDP pipeline. This approach of leveraging widely available dictionary resources provides a practical starting point for morphological analysis.

2. **Text Corpora:** For training the BPE tokenizers, we used large, pre-processed corpora derived from Wikipedia snapshots[6]. Our choice of data source and preprocessing methodology aligns with previous work on creating specialized Uralic tokenizers (Chelombitko and Komissarov, 2024), ensuring a comparable basis for our analysis. It is critical to emphasize that these corpora were used exclusively for training the BPE tokenizers and were not used in any stage of our morpheme list refinement pipeline, thus preserving the corpus-free nature of the IMDP method.

4.2 Reference Lexicon Creation

For each of the three languages, we applied our Iterative Morphological Decomposition Pipeline (IMDP), as described in Section 3, to the corresponding raw candidate list. The pipeline was configured with the following parameters: a minimum morpheme length $min\_length = 1$, a minimum type-support $m = 3$, and a convergence threshold $\varepsilon = 1e-7$. The process was run until convergence. The final filtering was performed using the automatically determined Otsu threshold (Otsu, 1979). This procedure yielded three high-purity reference morpheme lexicons (G_{fin}, G_{hun}, G_{est}), the statistics of which are summarized in Table 1.

4.3 Tokenizer Training

Using the tokenizers library[7] and SentencePiece (Kudo and Richardson, 2018) for comparison, we trained a series of Byte-Pair Encoding (BPE) tokenizers for each language from scratch. The tokenizers were trained on the respective Wikipedia corpora. To analyze the effect of vocabulary size, we trained separate models for a range of vocabulary sizes k, starting from 8,000 and up to 256,000 ($k \in \{$8k, 16k, 32k, 40k, 50k, 64k, 80k, 100k, 128k, 150k, 180k, 200k, 220k, 240k, 256k$\}$). All tokenizers were trained with a $min\_frequency$ of 2 for merges.

4.4 Evaluation Metrics

To provide a nuanced and rigorous evaluation of tokenizer quality, we must account for the fundamental trade-off between morphological coverage

[3] https://hunspell.github.io/

[4] https://github.com/LibreOffice/dictionaries

[5] https://github.com/fginter/hunspell-fi

[6] https://dumps.wikimedia.org

[7] https://github.com/huggingface/tokenizers

and over-segmentation. A tokenizer that perfectly represents all morphemes (high coverage) but also excessively splits common words is not optimal. To capture this balance in a single, unified score, we introduce the Integrated Performance Score (IPS).

The IPS models this trade-off geometrically. We consider a 2D space where the ideal tokenizer resides at the point (Coverage=1, OverSplit=0). The IPS of any real tokenizer is its normalized Euclidean distance from this ideal point, scaled to a [0, 1] range where 1 is perfect.

First, we define the two core components:

1. **Lexical Morpheme Coverage (LMC):** The fraction of atomic morphemes from our reference lexicon G that are perfectly represented as a single token in the tokenizer's vocabulary V_k. This measures the tokenizer's lexical "knowledge" of fundamental morphological units.

$$\text{LMC} = \frac{\left|\{\, m \in G \;|\; m \in V_k \,\}\right|}{|G|}.$$

2. **Over-split Rate (OSR):** The fraction of morphemes from G that the tokenizer fails to represent as single tokens, thus always splitting them into multiple pieces.

$$\text{OSR} = \frac{\left|\left\{ m \in M \;\middle|\; \begin{array}{c} m \text{ occurs in } \geq 1 \text{ word} \\ m \text{ never as a single token} \end{array} \right\}\right|}{|\{ m \in M \mid m \text{ in } \geq 1 \text{ word}\}|}.$$

From these, the Integrated Performance Score (IPS) is calculated as:

$$IPS = 1 - \left(\frac{\sqrt{(1-LMC)^2 + OSR^2}}{\sqrt{2}}\right)$$

This single metric allows for a clear and direct comparison of tokenizers across different vocabulary sizes. A higher IPS indicates a better balance between representing morphemes and avoiding excessive fragmentation. Our final analysis of optimal vocabulary sizes is based on identifying the "elbow point" on the IPS vs. vocabulary size curve.

5 Results and Analysis

Our experiment yielded clear and significant patterns regarding the relationship between tokenizer vocabulary size and morphological performance. To capture the fundamental trade-off between coverage and over-segmentation, we analyzed the Integrated Performance Score (IPS) for each language. The resulting IPS curves for Estonian (Figure 5), Finnish (Figure 6), and Hungarian (Figure 4) clearly show the performance profile for each language. Supplementary details on the component metrics (LMC and OSR) available in Figures 2 and 3.

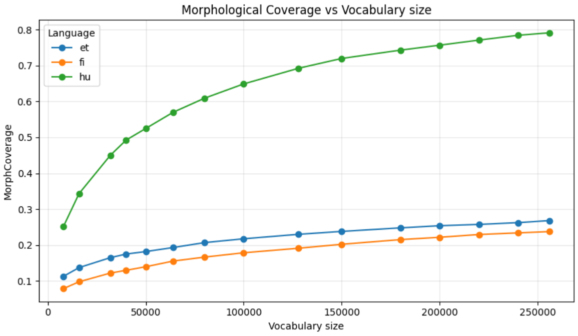


Figure 2: Lexical Morpheme Coverage (LMC) across different vocabulary sizes (k). LMC represents the percentage of reference morphemes found as single, complete tokens in the tokenizer's vocabulary.

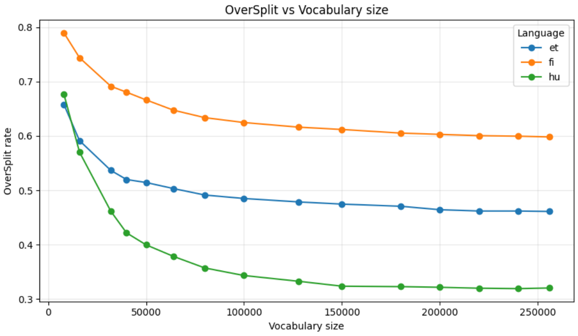


Figure 3: Over-Split Rate (OSR) as a function of vocabulary size (k). OSR denotes the fraction of reference morphemes that occur in words but never appear as a single token in any tokenization.

5.1 General Observation: A Clear Trade-off Profile

The IPS curves for all three languages exhibit a classic logarithmic growth pattern, demonstrating the law of diminishing returns. The score increases rapidly for smaller vocabulary sizes, indicating that initial additions to the vocabulary are highly efficient at capturing morphological structure. However, the rate of improvement progressively slows, showing that ever-larger vocabularies provide only marginal gains at a significant cost to model size. This confirms that a "sweet spot" or an optimal range exists for each language.

5.2 Cross-Linguistic Analysis: Three Distinct Performance Tiers

The results reveal three distinct performance tiers, highlighting the varying degrees to which standard BPE can model the morphology of these languages.

1. **Hungarian (hu):** As shown in Figure 4, Hungarian demonstrates by far the best performance. Its IPS curve starts at 0.29 and rises sharply, reaching a maximum of 0.73. This high score suggests that BPE is reasonably effective at learning the statistical regularities of Hungarian morphology.

2. **Estonian (et):** Estonian occupies the middle tier, with its IPS curve depicted in Figure 5. The score starts at 0.22 and reaches a maximum of 0.39. While better than Finnish, this score indicates that less than 40% of the "ideal" tokenizer performance is achieved, even with a large vocabulary.

3. **Finnish (fi):** Figure 6 illustrates the most challenging profile for Finnish. With a maximum IPS of only 0.31, the results quantitatively demonstrate that standard BPE is fundamentally ill-suited for capturing the complexities of Finnish morphology.

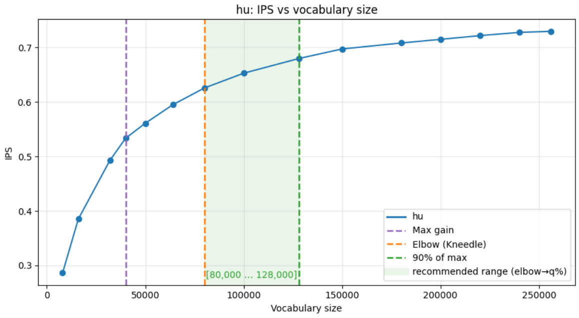


Figure 4: IPS vs. vocabulary size (k) for **Hungarian**. Hungarian shows the most consistent improvement in IPS, reflecting its comparatively transparent agglutinative structure with fewer morphophonological alternations. The elbow point is at 80k, and the 90% quality threshold at 128k, yielding a recommended range of 80k–128k.

5.3 Identifying the Optimal Vocabulary Range (k*)

To determine a practical and effective vocabulary size, we define a recommended range for k*. The lower bound of this range is the "elbow" point (k_elbow), identified by the Kneedle algorithm (Satopää et al., 2011), which marks the point of diminishing returns. The upper bound is the 90% quality point (k_q90), where 90% of the maximum observed IPS is achieved. As shown in Figures 4, 6, 5, and summarized in Table 2, this analysis leads to the following recommendations:

1. **Hungarian (hu):** The IPS curve for Hungarian (Figure 4) shows a clear optimal range between k=80,000 and k=128,000. The elbow is found at 80k, and 90% of the maximum performance is reached at 128k. As visualized on the plot, expanding the vocabulary beyond this range yields only minimal performance gains.

2. **Estonian (et):** For Estonian (Figure 5), the recommended range is also k=80,000 to k=128,000. Similar to Hungarian, the elbow is at 80k and the 90% quality mark is at 128k, establishing this as the zone of best compromise between performance and size.

3. **Finnish (fi):** The analysis for Finnish (Figure 6) indicates a need for a larger vocabulary. The elbow is at k=80,000, but to achieve 90% of the (albeit low) maximum performance, a vocabulary of k=150,000 is required. This suggests that for Finnish, the optimal range is k=80,000 to k=150,000, reflecting the language's high morphological complexity.

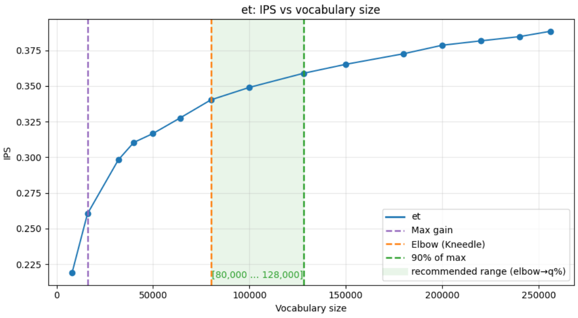


Figure 5: IPS vs. vocabulary size (k) for **Estonian**. While the overall pattern of diminishing returns is similar to Hungarian, the lower IPS plateau indicates reduced learnability due to Estonian's extensive morphophonological alternations, which obscure orthographic morpheme boundaries. The recommended range remains 80k–128k.

These findings provide a quantitative foundation for the critical decision of vocabulary sizing, transforming it from a heuristic-based choice into a principled optimization problem. Complete numerical

Lang	Max Gain Point (k_gain)	Elbow Point (k_elbow)	90% Quality Point (k_q90)	Recommend k* Range
Hung	40,000	80,000	128,000	80k – 128k
Est	16,000	80,000	128,000	80k – 128k
Fin	64,000	80,000	150,000	80k – 150k

Table 2: Key points on the IPS curve for determining the optimal vocabulary range.

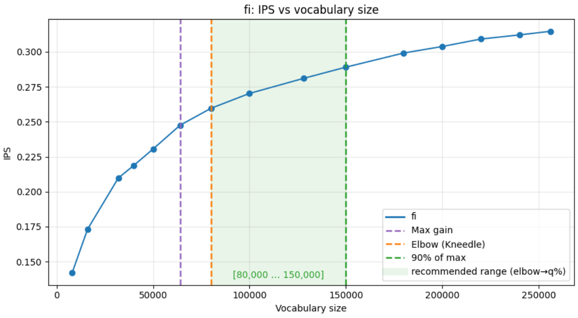


Figure 6: IPS vs. vocabulary size (k) for **Finnish**. Finnish exhibits the lowest IPS plateau, consistent with its rich system of consonant gradation and stem alternations, which make orthographic segmentation less stable for BPE. The elbow is at 80k, while 90% of the maximum IPS is reached at 150k, suggesting a recommended range of 80k–150k.

results for all evaluated vocabulary sizes are provided in Appendix A (Table 3) for reference.

6 Conclusion

In this work, we addressed the dual challenge of creating high-purity morphological resources in a corpus-free setting and using them to evaluate subword tokenizers for Uralic languages. We introduced SampoNLP, a toolkit featuring a novel pipeline based on "MDL-inspired Self-Referential Atomicity Scoring," which successfully refines noisy candidate lists into clean morpheme lexicons.

Applying these lexicons, our systematic evaluation of BPE tokenizers yielded two key findings. First, we provide an empirically-grounded recommendations for optimal vocabulary sizes, identifying a range of 80k-128k for Hungarian and Estonian, and 80k-150k for Finnish, as the most effective trade-off between performance and model size. Second, our results quantitatively demonstrate the severe limitations of standard BPE for highly agglutinative languages like Finnish, where performance plateaus at a strikingly low level.

This study confirms that while vocabulary size optimization is a crucial step, it is not a panacea. We release our SampoNLP library and the generated morpheme lists to the community to facilitate reproducible research and encourage the development of more morphologically-aware tokenization methods for the Uralic language family.

Discussion

Our results yield two key insights. First, the effectiveness of BPE varies dramatically by language: while Hungarian achieves a high IPS (max ∼0.73), the low scores for Finnish (∼0.31) and Estonian (∼0.39) quantitatively demonstrate the algorithm's fundamental limitations for these highly agglutinative languages. Second, for all languages, an empirically identifiable "sweet spot" for vocabulary size exists, beyond which performance gains diminish. Here, “optimality” is understood as morphological sufficiency - the point at which the tokenizer captures the productive structure of a language with minimal redundancy. This notion is intrinsic by design, offering a language-level criterion rather than task-specific optimization.

We acknowledge the limitations of our approach. The IPS metric abstracts away qualitative segmentation differences - a necessary compromise for scalability. Our use of clean, standardized corpora also isolates the variable of vocabulary size but does not reflect the noise of real-world data. These aspects represent clear avenues for future work.

While our method produces a refined set of recurrent sub-lexical units, we do not claim full linguistic morpheme correctness. The IMDP segmentation is orthographic and self-referential in nature, providing a practical approximation rather than a phonologically grounded morphological analysis.

In conclusion, our findings suggest that while optimizing k* is a crucial step, it may be insufficient for languages like Finnish. The low performance ceiling for BPE underscores the need for morphologically-aware tokenization methods. We believe our SampoNLP toolkit and the generated lexicons provide the community with a reproducible benchmark to develop and test such new strategies.

References

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.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Iaroslav Chelombitko and Aleksey Komissarov. 2024. Specialized monolingual BPE tokenizers for Uralic languages representation in large language models. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 89–95, Helsinki, Finland. 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. *Preprint*, arXiv:2410.12989.

Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30:75–93.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1).

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

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

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of Open Source Software*, 4(37).

T. Jauhiainen, Krister Linden, Niko Partanen, and .... 2020. Uralic language identification (uli) 2020 shared task: Wanca 2017 web corpora for uralic languages. *Proceedings of the VarDial Workshop at LREC 2020*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *Preprint*, arXiv:2112.10508.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Tommi A Pirinen. 2015. Omorfi — free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Marina Popova, Iaroslav Chelombitko, and Aleksey Komissarov. 2025. When repeats drive the vocabulary: a byte-pair encoding analysis of t2t primate genomes. *Preprint*, arXiv:2505.08918.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Ville Satopää, Joshua Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings of the 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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. *Preprint*, arXiv:1912.07076.

A Appendix

Language	Vocabulary Size (k)	Total Morphemes	Morpheme Coverage %	Over-Split Rate %
Estonian	8,000	5,705	11.27%	65.79%
Estonian	16,000	5,705	13.71%	59.13%
Estonian	32,000	5,705	16.49%	53.65%
Estonian	40,000	5,705	17.48%	51.98%
Estonian	50,000	5,705	18.18%	51.43%
Estonian	64,000	5,705	19.30%	50.32%
Estonian	80,000	5,705	20.68%	49.13%
Estonian	100,000	5,705	21.74%	48.49%
Estonian	128,000	5,705	22.99%	47.86%
Estonian	150,000	5,705	23.79%	47.46%
Estonian	180,000	5,705	24.78%	47.06%
Estonian	200,000	5,705	25.38%	46.43%
Estonian	220,000	5,705	25.74%	46.19%
Estonian	240,000	5,705	26.23%	46.19%
Estonian	256,000	5,705	26.81%	46.11%
Finnish	8,000	3,850	7.85%	78.96%
Finnish	16,000	3,850	9.76%	74.37%
Finnish	32,000	3,850	12.20%	69.13%
Finnish	40,000	3,850	12.95%	68.04%
Finnish	50,000	3,850	13.95%	66.62%
Finnish	64,000	3,850	15.53%	64.73%
Finnish	80,000	3,850	16.63%	63.36%
Finnish	100,000	3,850	17.84%	62.46%
Finnish	128,000	3,850	19.11%	61.61%
Finnish	150,000	3,850	20.21%	61.18%
Finnish	180,000	3,850	21.51%	60.52%
Finnish	200,000	3,850	22.16%	60.28%
Finnish	220,000	3,850	22.93%	60.05%
Finnish	240,000	3,850	23.38%	59.95%
Finnish	256,000	3,850	23.73%	59.81%
Hungarian	8,000	3,189	25.15%	67.72%
Hungarian	16,000	3,189	34.34%	57.01%
Hungarian	32,000	3,189	45.03%	46.14%
Hungarian	40,000	3,189	49.23%	42.17%
Hungarian	50,000	3,189	52.46%	39.97%
Hungarian	64,000	3,189	56.98%	37.84%
Hungarian	80,000	3,189	60.90%	35.72%
Hungarian	100,000	3,189	64.88%	34.33%
Hungarian	128,000	3,189	69.24%	33.27%
Hungarian	150,000	3,189	71.97%	32.37%
Hungarian	180,000	3,189	74.29%	32.28%
Hungarian	200,000	3,189	75.67%	32.16%
Hungarian	220,000	3,189	77.08%	32.00%
Hungarian	240,000	3,189	78.43%	31.92%
Hungarian	256,000	3,189	79.12%	32.04%

Table 3: Detailed experimental results for BPE tokenizers of varying vocabulary sizes across three Uralic languages. Morpheme Coverage represents the percentage of reference morphemes found in the vocabulary (LMC). Over-Split Rate is the percentage of reference morphemes with support in W that never appear as a single token in any tokenization.

Timur and the Mansi spellchecker

Csilla Horváth
University of Helsinki
csilla.horvath@helsinki.fi

Abstract

The article presents the results of an experiment involving the use of the Mansi FST and spellchecker created by the GiellaLT infrastructure. The Mansi are one of the indigenous peoples of the Russian Federation. The Mansi language is an endangered Uralic language primarily spoken in western Siberia, along the Ob River and its tributaries. The present article discusses the efficiency of the Mansi FST and spellchecker when used for translating Mansi literature from the 1950s.

1 Introduction

The article presents the results of an experiment of using the Mansi FST and spellchecker, created by the GiellaLT infrastructure on a text sample from a genre other than that for which the tool was designed.

Section 2 provides a brief overview of the Mansi language, its orthography and literature. Section 3 describes the experiment and its results. Section 4 presents the additional findings that were discovered during the experiment. Section 5 proposes conclusions.

2 Background

2.1 The Mansi language

The Mansi are one of the Arctic indigenous peoples of the Russian Federation. The Mansi are often still regarded as traditional communities living primarily on fishing, hunting, and gathering, and to some extent on reindeer breeding, however, as a result of industrialisation and urbanisation, taking place in their home region since the 1960s, the majority of the Mansi has been living in urban type of settlements, alienated from their once traditional lifestyle.

The Mansi language is a contested, endangered minority language. It is spoken mainly in Western-Siberia, along the Ob River and its tributaries. Four Mansi dialect groups were documented in the nineteenth century: Northern, Eastern, Southern, and Western Mansi, each of which had several (sub-)dialects. The Southern and Western dialects are already extinct, the Eastern dialect is either extinct or moribund. In this paper, Mansi language henceforth refers to the Northern Mansi variety. Mansi is used in both spoken and written form (cf. Virtanen and Horváth (2023)).

According to the results of the 2021 Federal Census, there were 12,308 Mansis living on the territory of the Russian Federation. 1,008 people claimed to use Mansi, altogether 951 of them were of Mansi ethnicity, while beside the Mansi, 8 Nenets, and 9 Khanties claimed to use the Mansi language. Nowadays, Mansi has its strongest position in the sphere of family language use, but since the turn of the century it has been introduced to new domains of language use as well, such as heritage language education, theatre and popular music, print, broadcast and social media (c.f. Horváth (2020, 2024, 2025)).

2.2 The written Mansi language

The literacy of the Mansi language, similarly to other indigenous languages of the Russian North, has been created in 1931. The Mansi standard orthography originally used Latin-based alphabet, then switched to a Cyrillic-based alphabet in 1937. Since 1937, the Mansi writing system has undergone several significant changes. In the earliest period, the Cyrillic transcription contained no special characters, and vowel length was not marked either. Later, perhaps at the beginning of the 1970s, a special character was introduced to denote the

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 68–71
December 10-12, 2025 ©2025 Association for Computational Linguistics

velar nasal, while marking vowel length became widespread in the 1980s.

Currently, two slightly different variants of Mansi orthography are in use, one, more marginal, used in some of the academic and pedagogical publications (dictionaries, traditional schoolbooks), the other, more general, used in all other work, including print and broadcast media, social media, and schoolbooks designed for heritage language learners.

2.3 Mansi literature

According to a comprehensive catalogue of Mansi publications (Юрьевна and Яковлевна (2007)) and personal experiences, there only have been approximately 170-180 books published in the Mansi language since 1937. More than half of them are primers and other schoolbooks, while the other half consists of folklore collections (mainly tales), contemporary literature, Mansi translations of Soviet literature for children, Mansi translations of the Gospels and other religious texts.

3 Timur and his squad

3.1 A Mansi FST and spellchecker

In 2025, a Mansi spellchecker has been released, created with a small set of morphophonological rules (32 twolc rules) and a lexicon consisting of 12,000 Mansi entries, as well as a larger set of proper nouns. Tested on a newspaper corpus consisting of approximately 700k tokens, the transducer was able to cover 98.9 % The transducer was turned into a spellchecker.

The Mansi grammatical model was reported to contain 825 continuation lexica and 12,063 stems with an additional set of over 145,000 shared lexemes at GiellaLT for the annotation of 100% equivalents of Russian names and toponyms (see Rueter, 2024). For a presentation, see Rueter et al. (2025).

3.2 Timur

The short novel titled Timur and his squad was written by Arkadiy Gaydar was first published in Russian in 1940. The book tells the story of a gang of village kids who sneak around secretly doing good deeds, protecting families whose fathers and husbands are in the Red Army. The novel had a huge impact on young Soviet audiences.

The short novel was translated into several languages, it was published in Mansi in 1955. According to the bibliographical data, the translation was made by E. Rombandeeva and A. Zyrin. As Zyrin's name is unknown in the history of Ob-Ugric Studies, while a certain Aleksandr Alekseevich Zyrin (1923-2003) appears to be a Turkologist-Orientalist-translator, it is safe to conclude that the Mansi translation, similarly to several other translations of the era, was the work of Evdokiya Rombandeeva only.

As the Russian original dates back to 1940, while the Mansi translation comes from 1955, the text of the short novel may be freely used for testing the Mansi FST without violating copyright law.

3.3 The experiment

The Mansi translation of Gaydar's short novel was digitalised. First, a physical copy of the book was scanned, then the documents were converted into text files with the help of Optical Character Recognition. The text file was proofread, and a copy of it was also manually normalised according to the orthography used in the Mansi press and contemporary educational publications. Both the original and the normalised versions of the text file were analysed by the Mansi spellchecker.

3.4 Results

Rombandeeva's original translation was published in an era of Mansi literacy where neither language-specific characters nor diacritics for long vowels have been in use, and the marking of palatalisation differed from present-day orthography as well. As a result of this, the spellchecker could recognise only wordforms that would not contain long vowels or special characters.

The adjusted version of the Mansi text performed much better. Most of the missing forms were analysed incorrectly due to the stems missing from the lexicon. The missing noun stems generally belong to the semantic field of the Socialist era, consisting of Russian loanwords (e.g. military terms, plant names), while the missing verbs are Mansi verbs that have not been mentioned in the newspaper corpus.

4 Other findings

In Mansi the deminutive suffixes -кве and -рищ can be added not only to nouns, but also to any other part of speech except conjunctions (Ромбандеева, 2017: 209). A verbal stem augmented with these suffixes can take tense, mood, or voice markers. (Riese, 2001: 59). The phenomenon is classified by Kálmán as precative mood (Kálmán, 1989: 61), Rombandeeva, and Riese, following her, call it merely a form. The suffix -кве expresses the speaker's respect, tenderness, as well as joy, delight, pride, and pleasant emotional disposition toward the surrounding environment and objects (Ромбандеева, 2017: 268). According to Kálmán, the -ке- variant of the suffix appears in front of other suffixes (Kálmán, 1989: 61), while Rombandeeva regards the form a dialectal alternation (Ромбандеева, 2017: 268). Beside occasional mentioning of the phenomenon, no detailed paradigm of the possible forms is given, also other examples than nouns or verbs are missing. In the translation of Gaydar's novel, an example documents the suffix -кве on the negative particle.

А̄тикве! — лю̄ньщалтахтынэ̄тэ ёл-пувим, э̄л-оим Женя ла̄вы́с.

а̄ти-кве лю̄ньщалтахты-нэ̄-тэ ёл-пув-им, э̄л-о-им Женя ла̄в-ыс

no-Prec cry-PtcpIpfv-Px3Sg hold.back-PtcpPrf, run.away-PtcpPrf Zhenya say-Pst.3Sg

'No! – said Zhenya, holding back her sobs and running away.'

According to Rombandeeva, when the suffix is used in the imperative, it denotes a polite request, a desire rather than an imperative (Ромбандеева, 2017: 269). Despite the Rombandeeva's short description, precative suffixes rather appear to follow mood markers than preceeding them.

Китыт телеграмма о̄с воекелн, юрт о̄йкакве.

китыт телеграмма о̄с во-е-ке-лн, юрт о̄йка-кве.

second telegram too take-Imp-Prec-2Sg.Sg, friend old.man-Prec

'Take the second telegramm too, comrade.'

According to Rombandeeva, the Mansi existential negative particle а̄тим is the variant of the negative verb а̄ти (Ромбандеева, 2017: 30), Murphy describes the situation the other way round (Murphy, 1968: 225), while Wagner-Nagy regards the two forms separate (Wagner-Nagy, 2011: 203). Neither of the authors mention that apparently а̄тим too can take the precative suffix -кве.

Турманыг та ё̄мтыс, э̄тимас, а тав акваг та а̄тимакве.

турман-ыг та ё̄мт-ыс, э̄тим-ас, а тав акваг та а̄тим-акве.

dark-trs thus become-pret-3Sg become.night-pret-3Sg but 3Sg still thus no-prec

'It's already became dark, night has fallen, still she is nowhere to be seen.'

Albeit minor details, these forms of negation and precative conjugation, were previously unattested in academic literature, also they appear more frequently than anticipated, as their use was considered to be marginal since the second half of the 20th century.

5 Conclusion

The Mansi FST and spellchecker, created by GiellaLT infrastructure, was aimed at language learners and language practitioners who would use the contemporary orthography of the Mansi language, and would create or process texts of the Mansi newspapers. The spellchecker provided excellent performance during the task. The experiment presented in this paper shows that the analyser and spellchecker achieve good results with texts from other genres too, although minor modifications to the model are still needed. While texts using older orthography can be adapted to the contemporary rules automatically rather than manually, the language model of the Mansi transducer requires extension with additional grammatical categories, and the lexicon, particularly the list of verbs, needs to be complemented.

Although the literary text corpus seems to be overshadowed by the newspaper corpus due to its size, the translation of youth literature was the second most common literary genre and the third most common genre of Mansi written texts. Moreover, as the experiment has proven, it supplements the language description both in terms of vocabulary and grammar. Processing literary texts with the Mansi FST proves to be beneficial, as such studies can lead to the discovery of previously undocumented grammatical phenomena.

References

Csilla Horváth. 2020. The vitality and revitalisation attempts of the Mansi language in Khanty-

Mansiysk. University of Szeged, Szeged.

Csilla Horváth. 2024. From the kitchen to pop culture: The role of Mansi heritage speakers in language shift and language revitalisation. Faits de langues, 54:197–212.

Csilla Horváth. 2025. The role of Ob-Ugric native speakers and heritage language speakers in creating Khanty and Mansi print, broadcast and social media. In Minority Language Media, pages 239–262. Palgrave Macmillan.

Bela Kálmán. 1989. Chrestomathia Vogulica. Budapest.

Lawrence W Murphy. 1968. Sosva Vogul grammar. Indiana University, Bllomington.

Timothy Riese. 2001. Vogul, volume 158 of Languages of the world Materials. Lincom Europa, München - Newcastle.

Jack Rueter. 2024. Testing and enhancement of language models (transducers) from GiellaLT (scientific blog). Scientific blog.

Jack Rueter, Csilla Horváth, and Trond Trosterud. 2025. A mansi fst and spellchecker. In Proceedings of the 9th Workshop on Constraint Grammar and Finite State NLP, pages 163–182, Tallinn. University of Tartu Library.

Susanna Virtanen and Csilla Horváth. 2023. Mansi. In The Uralic languages, 2nd edition, pages 665–702, London. Routledge.

Beáta Wagner-Nagy. 2011. On the Typology of Negation in Ob-Ugric and Samoyedic Languages, volume 262 of Suomalais-Ugrilaisen Seuran Toimituksia. Suomalais-Ugrilainen Seura, Helsinki.

Евдокия Ивановна Ромбандеева. 2017. Современный мансийский язык: Лексика, фонетика, графика, орфография, морфология, словообразование. Формат, Tymen.

Волженина Светлана Юрьевна and Фетисова Галина Яковлевна. 2007. Издания на языках народов ханты и манси (1879-2006). ООО Баско, Екатеринбург.

ORACLE: Time-Dependent Recursive Summary Graphs for Foresight on News Data Using LLMs

Lev Kharlashkin, Eiaki Morooka, Yehor Tereshchenko and Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
first.last@metropolia.fi

Abstract

ORACLE turns daily news into week-over-week, decision-ready insights for one of the Finnish University of Applied Sciences. The platform crawls and versions news, applies University-specific relevance filtering, embeds content, classifies items into PESTEL dimensions and builds a concise *Time-Dependent Recursive Summary Graph (TRSG)*: two clustering layers summarized by an LLM and recomputed weekly. A lightweight change detector highlights what is *new*, *removed* or *changed*, then groups differences into themes for PESTEL-aware analysis. We detail the pipeline, discuss concrete design choices that make the system stable in production and present a curriculum-intelligence use case with an evaluation plan.

1 Introduction

Foresight is the systematic detection, interpretation, and anticipation of signals of change to inform long-term decision-making (Chandrasekaran et al., 2022; Yüksel, 2012). For our setting, this means turning large, unstructured text streams (news, reports, social media) into structured representations of emerging trends and weak signals that remain interpretable to human analysts. Public institutions and universities in particular need early indicators of technological, policy, or societal change to align curricula, partnerships, and research priorities (Hämäläinen et al., 2024). Traditional approaches (expert workshops and static reports) cannot match the velocity and volume of modern information flows, motivating continuously operating, evidence-traceable foresight systems.

ORACLE addresses this need for a University of Applied Sciences by transforming Finnish news into a weekly *recursive summary graph* that reveals how narratives emerge, merge, and fade over time, grounded in PESTEL dimensions (see Yusop 2018). Unlike generic dashboards, ORACLE is designed for ongoing operation, traceability, and institutional relevance.

Our contributions are:

- A practical pipeline for daily Finnish news ingestion with hashing-based versioning and vectorized storage.
- A two-level Time-Dependent Recursive Summary Graph (TRSG) that hierarchically organizes narratives and updates weekly while accumulating long-term knowledge.
- A week-to-week change-detection mechanism that groups themes and analyzes institutional relevance through PESTEL.
- A real-world use case—curriculum intelligence—demonstrating decision support for academic stakeholders.

2 Related Work

Semantic representations and retrieval. Dense sentence embeddings (e.g., SBERT, SimCSE) enable semantic clustering and temporal tracking over large text streams (Reimers and Gurevych, 2019; Gao et al., 2021). For scalable retrieval, approximate nearest-neighbor search and vector databases such as FAISS and Milvus support efficient similarity queries over continuously ingested corpora (Johnson et al., 2021; Wang et al., 2021).

Clustering and community detection. Graph-based community detection (Louvain, Leiden) is widely used to expose latent structure in text-derived graphs (Blondel et al., 2008; Traag et al., 2019). Hybrid topic models like BERTopic combine transformer embeddings with clustering to yield interpretable topics, though they are typically applied to static snapshots rather than rolling streams (Grootendorst, 2022).

Summarization and abstraction. Neural abstractive models (BART, PEGASUS) produce flu-

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 72–76
December 10-12, 2025 ©2025 Association for Computational Linguistics

ent summaries but face challenges in factual grounding at multi-document scale (Lewis et al., 2020; Zhang et al., 2020). Extractive methods (LexRank, TextRank) remain strong for faithfulness and traceability (Erkan and Radev, 2004; Mihalcea and Tarau, 2004). Streaming and dynamic summarization explores hierarchical, time-aware abstractions over evolving corpora (Huang et al., 2024).

Temporal modeling and change detection. Modeling evolving topics spans burst detection and dynamic topic models, capturing intensity and drift over time (Kleinberg, 2003; Blei and Lafferty, 2006). These lines inform explicit week-over-week change labeling and rolling updates for interpretable monitoring.

LLMs in foresight and strategic analysis. Recent work investigates LLMs for structured foresight, including multi-layer PESTEL-based prompting to scaffold analysis (Alnajjar and Hämäläinen, 2024). Our system operationalizes these ideas as a continuous, interpretable pipeline combining semantic retrieval, dynamic clustering, hierarchical summarization, and explicit change tracking tailored to institutional decision-making.

3 Oracle Platform

3.1 Data Ingestion and Versioning

Sources. Finnish news (e.g., Yle) using open-source RSS feeds are crawled daily. The pipeline extracts canonical URLs and main content (boilerplate-stripped), preserving the original HTML for audit.

Hashing. We compute a stable content hash over normalized HTML. If the hash changes for a known URL, a new version is stored and re-embedded. This suppresses duplicates while tracking edits (e.g., headline updates).

3.2 Relevance Filtering for University

A two-stage filter prioritizes items that matter institutionally:

1. **Lexical stage.** Query expansions over names (*University*, *UAS*, Finnish aliases), domains (education, R&D, local industry) and geography. This fast pass removes obviously unrelated stories.
2. **Semantic stage.** Embedding-based similarity against curated exemplars (e.g., skills funding, curriculum reform, regional innovation). Borderline items are kept if semantically close.

Non-relevant items are cold-stored for later replays (interests can shift).

3.3 Embeddings, Storage and PESTEL

Documents are embedded (OpenAI TextEmbedding-3) and stored in Milvus with metadata: source, , publication date, PESTEL label and version chain. A compact supervised classifier assigns a single PESTEL label per item (multi-label is feasible but not used here). Cluster-level PESTEL distributions are computed by aggregating item labels.

4 Time-Dependent Recursive Summary Graph (TRSG)

Goal. Replace a flat feed with a weekly, two-level structure that is compact, faithful and easy to compare across weeks.

Cumulative Knowledge Base. While the crawling process runs daily and TRSG graphs are materialized weekly, all ingested data is stored in a single, persistent vector knowledge base. Each new crawl extends this base rather than overwriting it. This design allows the system to reason over historical embeddings when constructing new weekly hierarchies, enabling the detection of longer-term phenomena such as cluster drift, gradual topic convergence or the emergence of entirely new abstract groupings. Over time, these evolving structures provide early hints of potential future trends rather than merely weekly snapshots.

4.1 Construction

L0→L1 (sub-clusters). Build an item-level similarity graph for the week using cosine similarity; run Leiden (Traag et al., 2019). Summarize each community with a *factual* prompt (names, dates, figures, relationships) and embed the resulting text as the L1 node.

L1→L2 (meta-clusters). Cluster L1 summaries and produce an *abstract* summary per meta-cluster (themes, trends, implications, minimal specifics). L2 nodes represent the week's landscape.

Stability knobs. Small graphs use direct cosine; large graphs use FAISS with range thresholds. Weekly hierarchies are snapshotted for audit and fast reload.

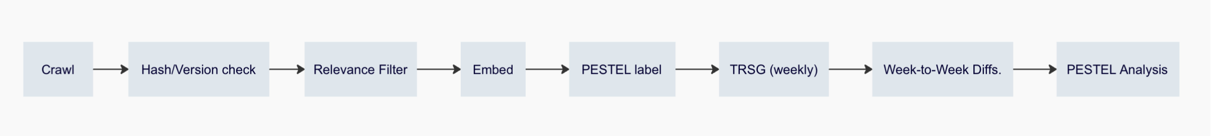
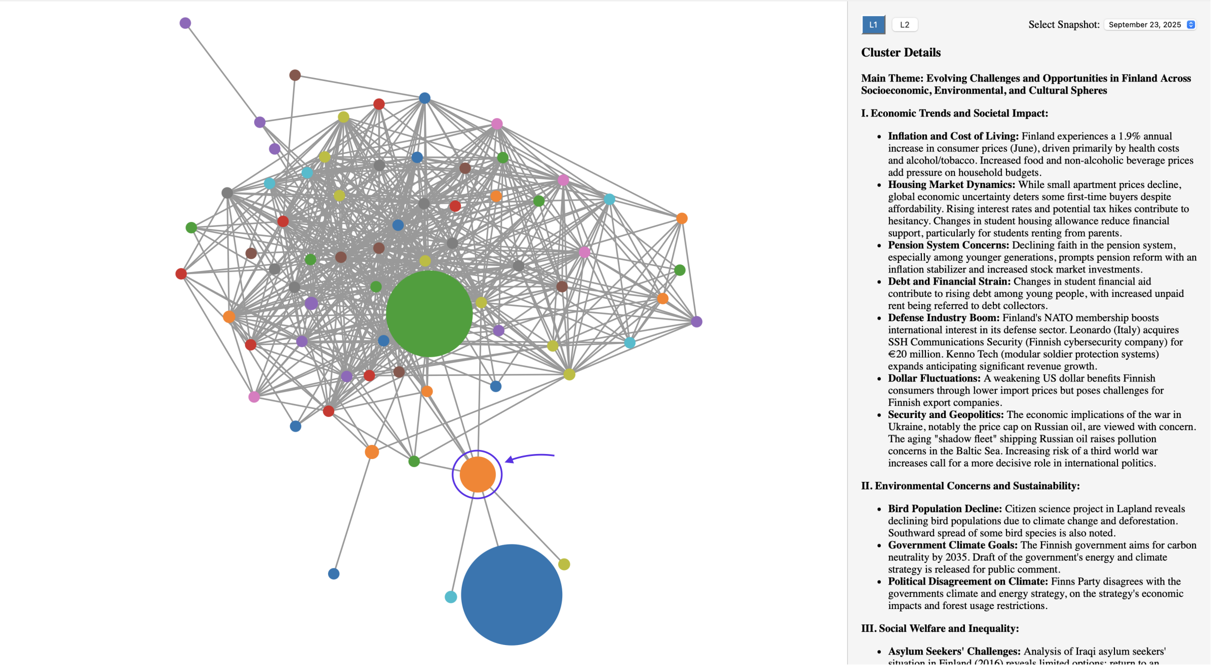


Figure 1: ORACLE workflow overview.



Select Snapshot: September 23, 2025

Cluster Details

Main Theme: Evolving Challenges and Opportunities in Finland Across Socioeconomic, Environmental, and Cultural Spheres

I. Economic Trends and Societal Impact:

- **Inflation and Cost of Living:** Finland experiences a 1.9% annual increase in consumer prices (June), driven primarily by health costs and alcohol/tobacco. Increased food and non-alcoholic beverage prices add pressure on household budgets.
- **Housing Market Dynamics:** While small apartment prices decline, global economic uncertainty deters some first-time buyers despite affordability. Rising interest rates and potential tax hikes contribute to hesitancy. Changes in student housing allowance reduce financial support, particularly for students renting from parents.
- **Pension System Concerns:** Declining faith in the pension system, especially among younger generations, prompts pension reform with an inflation stabilizer and increased stock market investments.
- **Debt and Financial Strain:** Changes in student financial aid contribute to rising debt among young people, with increased unpaid rent being referred to debt collectors.
- **Defense Industry Boom:** Finland's NATO membership boosts international interest in its defense sector. Leonardo (Italy) acquires SSH Communications Security (Finnish cybersecurity company) for €20 million. Kenno Tech (modular soldier protection systems) expands anticipating significant revenue growth.
- **Dollar Fluctuations:** A weakening US dollar benefits Finnish consumers through lower import prices but poses challenges for Finnish export companies.
- **Security and Geopolitics:** The economic implications of the war in Ukraine, notably the price cap on Russian oil, are viewed with concern. The aging "shadow fleet" shipping Russian oil raises pollution concerns in the Baltic Sea. Increasing risk of a third world war increases call for a more decisive role in international politics.

II. Environmental Concerns and Sustainability:

- **Bird Population Decline:** Citizen science project in Lapland reveals declining bird populations due to climate change and deforestation. Southward spread of some bird species is also noted.
- **Government Climate Goals:** The Finnish government aims for carbon neutrality by 2035. Draft of the government's energy and climate strategy is released for public comment.
- **Political Disagreement on Climate:** Finns Party disagrees with the governments climate and energy strategy, on the strategy's economic impacts and forest usage restrictions.

III. Social Welfare and Inequality:

- **Asylum Seekers' Challenges:** Analysis of Iraqi asylum seekers' situation in Finland (2016) reveals limited options: return to an

Figure 2: Example of the L1 layer in the TRSG. Each node represents a cluster of semantically related news items and edges indicate similarity links. The summarization step condenses each cluster into a factual thematic report, forming a connected graph of emerging narratives.

4.2 Prompt Design

To enforce consistent abstraction, TRSG uses level-specific prompts and recursive summarization when input exceeds model limits.

L1 – Thematic summaries. Summarizes factual content across related news items:

> “Create a comprehensive L1 thematic summary in English. Do not evaluate or give suggestions. Identify the main theme, structure logically and include key facts—entities, dates, figures, policy changes and regional details. Write a complete, factual report without introductory remarks.”

This produces grounded, information-rich cluster summaries.

L2 – Strategic synthesis. Aggregates L1 outputs into cross-domain insights:

> “Create a unified strategic L2 intelligence briefing in English. Do not evaluate or compare. Extract overarching patterns and systemic trends, emphasizing transformation forces across domains. Present as a coherent intelligence report without meta-commentary.”

L2 abstracts factual clusters into foresight-level narratives.

Recursive summarization. If a cluster’s combined text exceeds the model’s context limit, the content is split into balanced chunks and summarized recursively. Each batch is first summarized individually with the same prompt and those interim summaries are then re-summarized to produce the final L1 or L2 output. This hierarchical compression maintains completeness and coherence even for large clusters.

5 Week-to-Week Change Detection

Snapshots from consecutive weeks are compared at L1 and L2 using cosine similarity.

Matching. For each new summary, find the best old neighbour. Labels: **Stable** (sim $\geq$0.90), **Changed** (0.70–0.90), **Added** (<0.70). Unmatched old summaries are **Removed**. This yields structured deltas rather than vague impressions.

Component	Default / Behavior
Embedding model	`text-embedding-3-small`
L0→L1 threshold	0.75 (cosine)
L1→L2 threshold	0.55 (cosine)
Small-n fallback	Direct cosine (no FAISS)
Clustering	Leiden (modularity)
Summarization	Gemini 2.0 Flash (L1 factual / L2 abstract)
Persistence	Weekly snapshots (`pickle`)

Table 1: Key TRSG defaults used in production.

Theme grouping. Added/Removed lists are converted to human-readable themes: short micro-labels per text (LLM), then canonicalization via TF–IDF + agglomerative clustering (cosine distance). The result is a deterministic set of {`label`, `added_texts`, `removed_texts`} per level.

PESTEL analysis. Users select perspectives. For each theme, a schema-constrained analysis returns {`title, analysis, level, group, importance [0,1]`}. The results are cached in MySQL, keyed by week pair and perspective, to ensure reproducibility and fast retrieval.

6 Use Case: Curriculum Intelligence

An analyst monitoring Political+Technological developments compares weeks 23 and 28. TRSG highlights two new L2 themes: *EU digital skills funding* and *quantum computing policy momentum*. L1 reveals the concrete facts (program names, funding figures, named institutions). The PESTEL analysis recommends actions: (i) align elective modules with EU skill frameworks, (ii) add a quantum fundamentals stack (concepts + labs) and (iii) explore partnerships with local industry labs. The value is not prediction but *traceable synthesis* tailored to University's remit.

In addition to supporting concrete curriculum adjustments, the TRSG output provides a reusable evidence base for cross-faculty coordination. Because every proposed action links back to the underlying news signals and cluster summaries, departments can justify decisions using a shared source of truth rather than ad-hoc interpretations or anecdotal reports. This auditability also supports institutional learning: when decisions are revisited months later, the exact informational context that motivated them remains inspectable. In practice, the platform enables routine, low-friction foresight workflows (monthly horizon scans, annual strategy cycles or accreditation preparations) without requiring analysts to rebuild situational awareness from scratch.

The same machinery also generalizes to other decision layers, including research prioritization, stakeholder engagement and regional partnership planning. Since TRSG captures both stable background narratives and emerging weak signals, it helps distinguish between noise, structural change and transient bursts of attention. For universities operating in rapidly evolving technological and policy environments, this distinction is essential: long-horizon initiatives (e.g., lab infrastructure, degree redesign) demand evidence of durable trends, while quick interventions (e.g., micro-credential pilots) benefit from timely detection of nascent opportunities. By encoding both views into a single recurring graph, ORACLE makes such multiscale reasoning feasible for non-technical users.

7 Conclusion

ORACLE demonstrates that continuous, traceable foresight over fast-moving news streams is operationally achievable using a combination of embeddings, hierarchical clustering, recursive summarization and week-to-week change detection. By structuring evolving narratives into a two-level Time-Dependent Recursive Summary Graph (TRSG) and grounding interpretation in PESTEL, the platform delivers decision-ready intelligence that remains auditable and aligned with institutional needs. The curriculum-intelligence case shows that such a system can inform real strategic choices without relying on opaque prediction. Future work will refine evaluation, explore multi-label PESTEL, and extend the approach to multilingual sources and policy–science–industry link analysis.

8 Limitations

Our system has several inherent limitations:

- **Coverage bias.** The system depends on which news sources are crawled; underrepresented voices or niche media may be missed, skewing the narrative graph.
- **Summary hallucination.** Although prompts are guarded and ground traces are preserved, LLMs may still introduce inaccuracies or omit subtle but relevant facts.
- **Domain specificity.** The PESTEL classifier and thresholds are tuned for one university's domain;

generalizing to another institution or domain will require re-training or re-tuning.

- **Temporal granularity.** Weekly snapshots may miss rapid developments or sub-week bursts; while daily crawling accumulates data, changes within a week are abstracted.

References

Khalid Alnajjar and Mika Hämäläinen. 2024. Mlpestel: the new era of forecasting change in the operational environment of businesses using llms. *Centria University of Applied Sciences*.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of ICML*, pages 113–120.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. In *Journal of Statistical Mechanics: Theory and Experiment*, volume 2008, page P10008. IOP Publishing.

Arun Chandrasekaran, George Thomas, and Reshma Nair. 2022. Forecasting and foresight: Systematic review of methods in ai-driven strategic analysis. *Technological Forecasting and Social Change*, 185:122095.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. In *Journal of Artificial Intelligence Research*, volume 22, pages 457–479.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Mika Hämäläinen, Marita Huhtaniemi, and Leena Unkari-Virtanen. 2024. Suositus ennakointitiedon johtamiseen. *Tulevaisuuden kudelmia. Ennakointikyvykkyyden kehittäminen Metropoliassa*, pages 50–63.

Han Huang, Tao Li, and Wei Xu. 2024. Streamsum: Streaming summarization of evolving news topics with large language models. In *Proceedings of EMNLP*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547. Initial version arXiv:1702.08734.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of EMNLP*, pages 404–411.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Zilliz Wang and 1 others. 2021. Milvus: A purpose-built vector database for scalable similarity search. *arXiv preprint arXiv:2108.04535*.

Tamer Yüksel. 2012. Developing a scale for measuring the pestel dimensions in strategic management. *Procedia - Social and Behavioral Sciences*, 58:1091–1101.

Zaid Yusop. 2018. Pestel analysis. *Paper persented at COMRAP*, pages 34–39.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of ICML*.

Creating a multi-layer Treebank for Tundra Nenets

Nikolett Mus
ELTE Hungarian Research
Centre for Linguistics
Budapest, Hungary
mus.nikolett@nytud.elte.hu

Bruno Guillaume
Université de Lorraine
CNRS, Inria, LORIA
Nancy, France
Bruno.Guillaume@loria.fr

Sylvain Kahane
Université Paris Nanterre
Modyco
Paris, France
sylvain@kahane.fr

Daniel Zeman
Univerzita Karlova
MFF, ÚFAL
Prague, Czechia
zeman@ufal.mff.cuni.cz

Abstract

This paper presents the development of the Tundra Nenets Universal Dependencies (UD) Treebank, the first syntactically annotated resource for the Samoyedic branch of the Uralic family. The treebank integrates spoken-language data and adopts the morphologically enhanced Surface-Syntactic UD (mSUD) framework to capture inflectional morphology and morphology-based syntactic relations. It further incorporates Information Structure annotation. The methodological workflow includes data selection, transcription conventions, sentence and lexeme segmentation, annotation of spoken-language features, lemmatization, treatment of morpheme status, part-of-speech and morphological tagging, and syntactic annotation based on the functional and distributional properties of syntactic elements. We also outline the principles guiding multi-level annotation and justify the theoretical choices underlying the integration of prosodic, morphological, and syntactic information.

1 Introduction

This paper presents the development of the (Tundra) Nenets (Samoyedic, Uralic) Universal Dependencies (UD) Treebank, including data selection and processing, levels of linguistic analysis, and the theoretical and methodological principles guiding the annotation. Given the early stage of development and the limited size of the corpus, the focus is on the foundational methodological approaches and theoretical decisions underlying the construction of the treebank.

Within the Uralic language family, the Finno-Ugric branch is already represented in UD by several treebanks (e.g., Finnish, Estonian, Hungarian, Komi, Udmurt), whereas the Samoyedic branch has remained absent from the data set. The inclusion of Tundra Nenets, a major Samoyedic language spoken in northwestern Siberia, therefore fills a significant gap and contributes to a more balanced coverage within the Uralic family.

Several digital corpora of Nenets exist, for instance in the Endangered Languages Archive (ELAR) and the INEL Nenets corpus (Budzisch and Wagner-Nagy, 2024), yet these resources have largely remained at the level of morphological annotation and provide limited support for syntactic analysis. Syntactic structure, in particular, remains underexplored in Tundra Nenets as well as in other Siberian Uralic and Siberian Arctic languages. The Tundra Nenets UD treebank addresses this gap by providing a systematically annotated syntactic resource, thereby enabling detailed investigations of (morpho)syntactic patterns. In addition, the treebank functions as a methodological case study for adapting the UD framework to a morphologically rich Uralic language. Its development is expected to inform both the creation of comparable resources for other Samoyedic and Siberian languages and broader discussions concerning the representation of typologically complex languages within the UD framework.

The project introduces several innovations within the UD framework (de Marneffe et al., 2021). First, since the treebank is based on spoken language data, it was necessary to determine (i) the level of transcription detail, specifically, which spoken-language-specific features should be included in the syntactic analysis, and (ii) how these phenomena should be represented, that is, the corresponding annotation principles and technical solutions. Second, the morphologically enhanced version of the Surface-Syntactic Universal Dependencies framework (Gerdes et al., 2018, 2019), the mSUD model (Guillaume et al., 2024), was adopted as the basis for annotation. This framework accommodates the rich morphological structure of Tundra Nenets and allows for an explicit representation of morphology-based syntactic relations, while remaining fully compatible with the

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 77–86
December 10-12, 2025 ©2025 Association for Computational Linguistics

UD standard. Third, the Nenets treebank includes (partial) information-structural annotation as part of a new initiative within the frame of the UniDive COST Action (CA21167), which aims to extend UD with additional layers capturing the discourse-pragmatic functions of clausal constituents.

2 The Tundra Nenets language and data

2.1 The language

Nenets is classified as a member of the Samoyedic branch of the Uralic language family. Prior to the twentieth century, linguistic descriptions generally treated Tundra and Forest Nenets as the primary dialectal varieties of the Nenets language. However, these varieties differ substantially in grammar and lexicon, and are not mutually intelligible, which justifies treating them as separate languages (Hajdú, 1968; Salminen, 1998; Burkova, 2022; Mus, 2023a). Since the current treebank includes only Tundra Nenets data, this paper focuses on that variety; with Forest Nenets materials planned for inclusion in future expansions of the corpus.

The Tundra Nenets language is spoken in the northernmost regions of the Russian Federation, primarily in the autonomous Okrugs of Nenets and Yamalo-Nenets and the Taymyrsky Dolgano-Nenetsky district. It covers an extensive Arctic area, extending from northeastern Europe to northwestern Siberia (maps illustrating these territories can be found online[1]).

The language has c. 20,000 speakers, divided into Western, Central, and Eastern dialect groups, each with local subdialects (Hajdú, 1968; Tereshchenko, 1966; Salminen, 1999; Nikolaeva, 2014; Burkova, 2022; Mus, 2023a).

It is an indigenous Arctic language and is classified as *threatened* (EGIDS 6b) (Ethnologue, 2009). Although still used in everyday oral communication across generations, speaker numbers are declining. Widespread bilingualism with Russian has led to notable lexical and structural influence.

Traditionally an oral language, Tundra Nenets achieved literacy only in the late 1920s, when a Cyrillic-based orthography was introduced (Toulouze, 1999). The writing system remains non-standardized, and several Latin-based transliteration systems are employed in scholarly contexts.

Tundra Nenets is a morphologically rich, agglutinative language. Its grammatical relations are expressed mainly by suffixes attached sequentially to the stems. Despite its agglutinative character, stem and affix alternations introduce fusional features.

The language exhibits nominative–accusative alignment (Nikolaeva, 2014): subjects are marked with the nominative case, while direct objects are typically marked with the accusative case, with limited syncretic exceptions (Hajdú, 1968; Nikolaeva, 2014). Finite verbs agree with their subjects in person and number, and may also mark agreement with objects in number when these are topical (Nikolaeva, 2014). Predicate nouns, adjectives, and certain adverbs also show agreement with their subject in person and number, and can also take the suffix of the past tense without inserting an overt copula in the predicate phrase (Nikolaeva, 2014; Hegedűs et al., 2021).

Syntactically, the language is head-final and predominantly (S)OV (Tereshchenko, 1973; Nikolaeva, 2014; Burkova, 2022; Mus, 2023a), with complements preceding their heads. *Right-dislocated* elements or *afterthoughts* occasionally occur, separated by a prosodic break and distinct intonation (Mus and Surányi, 2021, 2025). Coordination generally lacks conjunctions, while subordination is expressed through non-finite verb forms that precede the main predicate. In certain subordinate clause types, the embedded subject can trigger agreement on the non-finite verb through possessive morphology (Nikolaeva, 2014; Mus, 2023b).

2.2 The data

Written and spoken materials of Tundra Nenets are accessible in several archives and collections, including the Endangered Languages Archive (ELAR)[2], the Online Documentation of Siberian Languages[3] (Nikolaeva and Garrett, 2014), and the INEL Nenets corpus[4] (Budzisch and Wagner-Nagy, 2024). Folklore, collected during fieldwork in the region, is the most commonly represented genre in these resources. The transcription conventions, transliteration schemes, and annotation frameworks employed across these collections vary considerably and are sometimes inconsistent.

An online newspaper from the Nenets Autonomous Okrug regularly publishes articles in

[1] https://nenetsresearch.github.io/thea/tools.html

[2] https://www.elararchive.org/

[3] https://siberianlanguages.surrey.ac.uk/

[4] https://inel.corpora.uni-hamburg.de/NenetsCorpus/search

Tundra Nenets alongside Russian, providing a contemporary source of written materials in the language.[5] Additionally, a digitized text collection of approximately 500,000 tokens has been compiled and normalized, representing the written variety of the language (Mus and Metzger, 2021).

Complementing these written sources, new spoken data were collected during consultations in Moscow in 2017 with a Tundra Nenets speaker from the Yamalo-Nenets Autonomous Okrug. Rather than focusing once again on folklore, the fieldwork employed methods from Language Documentation and experimental syntax to elicit semi-controlled, naturalistic language production through interactive, goal-oriented tasks. These included a modified version of the HCRC Map Task[6], the so-called Pear Story narrative task (Chafe, 1980), and a storytelling video stimulus about reindeer herding.[7] A third elicitation type made use of picture-based story sequences, in which the speaker was asked to narrate a story depicted in a series of cartoon-style illustrations. In addition, a questionnaire was designed to prompt conversation on neutral, everyday topics, like cooking, free-time activities, public transport, and comparisons of two cities, ensuring that no sensitive personal information was collected. Finally, a set of scripted dialogues was read aloud providing controlled data for analysing prosodic phrasing and syntactic structures under comparable conditions.

The narratives were recorded in audio format (.wav), and the native speaker participant transcribed a subset of these materials orthographically using an extended Cyrillic alphabet.

Table 1 provides a summary of the available data and their current processing status from the UD perspective. Tasks and datasets that have already been processed and incorporated into the Tundra Nenets UD treebank are highlighted in green[8], while the remaining materials will be processed and added in subsequent releases of the treebank.

Type of task	Length (sec)	Nr. of sentences
HCRC Map Task 1–4	340	93
video-based storytelling: Pear Story	355	78
video-based storytelling: Arctic reindeer	235	n.d.
picture sequence narration 1–4	576	n.d.
thematic topic guided monologue 1–4	1,403	n.d.
scripted dialogue reading 1–3	232	n.d.

Table 1: Tundra Nenets spoken datasets and their current UD processing status

3 Procession of the data

In the following sections, we outline the data processing workflow, describe the methodology used to establish it, and discuss language- and data-specific annotation decisions.

3.1 Transcription and annotation of spoken language phenomena

As noted above, several texts have already been transcribed by the native speaker participant. These recordings were selected as the starting point for annotation, as sentence segmentation had already been performed by the speaker. This made the data particularly suitable for addressing one of the central theoretical challenges in spoken-language analysis: defining what constitutes a syntactic unit. The analysis of these materials provides the empirical foundation for subsequent data processing and for our working principle, which holds that intonational and semantic criteria should be jointly considered when determining sentence boundaries in spoken data.

While intonation units provide important cues for segmentation, they do not always coincide with syntactically or semantically complete utterances. Accordingly, a prosodic boundary was treated as a sentence boundary only when the preceding unit expressed a complete meaning. When an intonational unit was semantically incomplete, the subsequent material was incorporated into the same sentence. This approach ensures that sentence segmentation reflects both the prosodic organization of speech and the syntactic and semantic coherence required for UD annotation. The audio files and their transcriptions were manually annotated and time-aligned at the sentence level in Praat (Boersma and Weenink, 2025).

In addition to sentence-level prosodic alignment, individual lexemes were time-aligned with their corresponding segments in the recordings. This was done to facilitate morphological and syntactic interpretation and to support future research

[5] `https://nvinder.ru/`

[6] `https://groups.inf.ed.ac.uk/maptask/maptasknxt.html`

[7] The Pear Story task was included purely as an exploratory experiment. Given its culturally foreign context, we anticipated that the task would be challenging and highly open to interpretation, yet the language consultant completed it with remarkable fluency and engagement.

[8] To be included in release 2.17 (November 2025).

on the syntax–prosody interface. This step was also undertaken manually.

At this stage, several decisions were required concerning the treatment of spoken-language phenomena and the desired level of analytical detail. The transcriptions prepared by the native speaker follow a normalized Cyrillic orthography that reflects standardized dialectal forms rather than surface phonetic realizations. Consequently, phonetic transcriptions that capture the morphophonological peculiarities of the language were not produced at this stage. For instance, external sandhi processes – phonological alternations operating across word boundaries – observed in the language were not annotated.[9] The corresponding audio recordings, however, will be made publicly available for reference.

Instead, only those spoken-language phenomena were annotated that directly affect word-level analysis, particularly the identification of word boundaries. This decision reflects both the current lack of established UD guidelines for spoken data and the absence of detailed prosodic or phonetic descriptions of Tundra Nenets.

Rather than adopting an external prosodic annotation framework, an inductive approach was taken: recurrent lexeme-level spoken phenomena were identified directly from the recordings and transcriptions. For each such phenomenon, a dedicated annotation tag was created and, where applicable, linked to a syntactic relation within the UD framework. The conventions were inspired by existing spoken UD treebanks and preliminary annotation guidelines (Kahane et al., 2021; Dobrovoljc, 2025), but in several cases, the labeling strategy was adapted to accommodate the specific structural and typological features of Tundra Nenets. Once defined, the conventions were applied consistently across the corpus. This approach ensures that, despite the early stage of UD-based spoken-language analysis, the current representation is internally coherent and flexible enough to accommodate future standardization efforts.

Since the available recordings consist primarily of narrative monologues, certain interactional features typical of spontaneous dialogue – such as overlapping speech – are not attested.

To achieve a detailed lexical representation, two groups of spoken-language items were annotated: non-lexical and lexical items. Non-lexical items were included primarily to ensure the precise identification of word boundaries. As they do not constitute syntactic units, they were uniformly assigned the `discourse` dependency relation. This category includes:

- Noises <n> (both speaker-generated, e.g., cough, laugh, sigh, and environmental, e.g., background chatter, traffic, microphone bumps);

- Pauses <p> occurring within smaller syntactic units or between unrelated constituents;

- Audible disfluencies <d>, such as hesitation markers ("uh", "erm").

Lexical interruptions, by contrast, directly affect syntactic interpretation. These include:

- Unfinished lexemes <un>, which may leave grammatical relations incomplete (e.g., missing a required case marker);

- False starts <f> and repetitions (exact or partial) <er> and <pr>, which may alter expected word order;

- Incorrect word selections <iw>, where a semantically or morphologically related but unintended form is produced.

Finally, pauses coinciding with syntactic boundaries were annotated analogously to punctuation in written texts (POS `PUNCT` and relation `punct`), as they play a key role in delimiting sentence boundaries, while pauses corresponding to hesitations where analyzed analogously to disfluencies (POS `INTJ` and relation `discourse`).

The full inventory of annotated spoken-language features, together with their corresponding tags and syntactic encodings, is summarized in Table 2.

The treebank is encoded in the CoNLL-U standard format, with lexeme-level time-alignment information, indicating the onset and offset of each annotated item, stored in the MISC column.

3.2 Lemmatization, POS tagging, and morphological analysis

Building on the transcription, segmentation and time-aligned annotation of the spoken data, the subsequent stage of corpus development involved

[9] In connected speech, for example, the phrase *тына’ хадамбива / tinaʔ xadamb́iwa* ‘we killed the reindeer’ reindeer.acc.1pl kill-1pl) surfaces as *тына_кадамбива / tina_kadamb́iwa*.

Category	Tag	POS	DEPREL
Noise	<n>	INTJ	`discourse`
Pauses (hesitation)	<p>	INTJ	`discourse`
Audible disfluencies	<d>	INTJ	`discourse`
Unfinished lexemes	<un>	intended lexeme	`reparandum`
False Starts	<f>	intended lexeme	`reparandum`
Repetitions	<er> <pr>	intended lexeme	`reparandum`
Incorrect word	<iw>	intended lexeme	treated as if correct
Pauses (boundary)	<p>	PUNCT	`punct`

Table 2: Annotated spoken-language phenomena in the Tundra Nenets UD Treebank

the lemmatization, part-of-speech tagging, and morphological analysis of the data. Since no automated tools currently exist for Tundra Nenets, these steps were performed manually.

The process of lemmatization was guided by several theoretical decisions. First, in the absence of a unified written standard, dialectal variation was preserved in the forms of the lemma.[10] Second, during segmentation, only inflectional morphemes were detached from the stems, while the derivational morphology was left intact. Inflectional suffixes were retained in their attested surface forms and do not receive normalized lemmas. Third, linking vowels appearing at the boundary between stems and suffixes were treated as integral parts of the stem.

Part-of-speech tagging and morphological analysis were likewise carried out manually. Morphological features were segmented from the stems and glossed (cf. feature `Gloss`). In the segmentation and annotation process, only inflectional morphology was included, as inflectional markers directly contribute to syntactic relations, whereas derivational morphology was treated as part of the lexical stem. This treatment of morphology ensures consistency with the syntactic representation model adopted in the subsequent section, which integrates morphological and syntactic dependencies. The distinction between inflectional and derivational morphology was determined on the basis of descriptive and grammatical traditions established in Hajdú (1968); Nikolaeva (2014); Burkova (2022); Mus (2023a,b).

A distinction from these sources concerns the analysis of the verbal paradigm, specifically the treatment of the verbal linking suffix *-ӈа* (*-ŋa-*) that is added to certain verbal stems before agreement suffixes. Since the status of this suffix is not clear in the literature, this element was segmented from the verb stem and assigned the `AUX` POS, reflecting its auxiliary-like syntactic behavior within the clause.

In the nominal paradigm, the so-called predestinative suffix – *-да* (*-da*) – was also segmented, though it was commonly regarded a derivational element in the descriptive tradition. We propose that this morpheme may in fact participate in a syntactic relation with the noun it modifies, functioning similarly to a determiner. Accordingly, it was segmented and assigned the `det` dependency relation.

The annotation of morphosyntactic features and category labels follows the conventions of the aforementioned descriptive sources. The POS and morphological tagset was adapted from the Leipzig Glossing Rules and Abbreviations framework, with necessary modifications introduced to accommodate the specific structural and typological properties of Tundra Nenets.

To ensure consistency and uniformity across the corpus, all analyzed word forms were compiled into a reference TSV file containing the surface form, lemma, POS tag, morphological information, and translation (see Table 3). This file serves as a master inventory of annotated forms. New raw data are automatically compared against this reference using a Python script: whenever a match is found, the corresponding lemma, POS tag, and gloss information are automatically inserted into the new annotation. This procedure not only preserves consistency between previously annotated and newly added data, but also considerably accelerates the annotation process while retaining manual verification for forms not yet included in the reference file.

Form	**Lemma**	**POS**	**Gloss**
маря	мар”	NOUN	fence
-д’	_	ADP	-poss.gen.2sg

Table 3: Example excerpt from the reference TSV

3.3 mSUD annotation

To account for the complex morphological structure of Tundra Nenets and its role in expressing syntactic relations, the morphologically enhanced Surface-Syntactic Universal Dependencies (mSUD) framework (Guillaume et al., 2024) was adopted as the foundation for annotation.

[10] For example, the numeral ‘three’ occurs as *няр / ńar* in the Western dialect and as *няхар / ńaxar* in the Central and Eastern dialects.

The mSUD framework builds on the principles of the Surface-Syntactic UD model but extends it to explicitly represent morphology-based syntactic relations. It prioritizes functional heads within phrases, i.e. those constituents that determine the syntactic and distributional properties of the entire phrase, while defining dependency relations on functional and distributional grounds (Gerdes et al., 2019).

As noted, within this framework, both independent words and inflectional morphemes are systematically linked to their corresponding syntactic relations. Derivational morphology, by contrast, is not analyzed as directly contributing to syntactic dependencies. The main annotation choices adopted for Tundra Nenets suffixes are summarized in Table 4, which illustrates how distinct types of inflectional morphemes are represented and how their syntactic dependents are encoded within the mSUD relation set.

Inflection	**POS**	**mSUD DEPREL**
Number	DET	−det→ •
Case	ADP	• −comp:obj→
Possessive suffix	DET	−det:poss→ •
Predestinative suffix	DET	−det→ •
Tense suffix	AUX	• −comp:aux→
Mood suffix	AUX	• −comp:aux→
Subject agreement suffix	PRON	−subj→ •
Double agreement suffix	PRON	−subj:obj→ •
Non-finite verb suffix	AUX	• −comp:aux→

Table 4: mSUD annotation for Tundra Nenets

In the table above, the `subj:obj` relation in the verbal paradigm may require further explanation. As will be discussed in greater detail below, in Tundra Nenets transitive verbs can agree not only with their subject but also simultaneously with both their subject and object when the object is topical. Such agreement markers are typically unanalyzable portmanteau morphemes that cannot be segmented into separate units. Consequently, they were treated as a single unit and assigned the `subj:obj` dependency relation.

Because mSUD provides a more fine-grained representation of morphological and syntactic structure than standard UD, it offered a logical starting point for the development of the Tundra Nenets treebank. The annotation process therefore begins in mSUD and is subsequently converted to UD format. This direction of conversion is unidirectional: while mSUD can be reliably reduced to UD through structural simplification, the reverse conversion – from UD to mSUD – would require morphological information not encoded in UD and thus cannot be reconstructed automatically.

Annotation was carried out in a semi-automatic way using ArboratorGrew[11] (Guibon et al., 2020), which allows the creation and application of reusable rules to automate certain aspects of dependency annotation. While our corpus consists of approximately 200 sentences, the syntactic rules we employ are not probabilistic generalizations derived solely from this sample, but stable and well-established structural properties of the language. These rules are categorical (e.g., the agreement morphology on verbs that our annotation framework analyzes as subject marking) and are not subject to variation across larger datasets. Therefore, although the current rule set may be incomplete in the sense that additional rules could be added when annotating a larger corpus, the rules already formulated remain fully applicable and reliable regardless of corpus size. In other words, expanding the dataset would increase coverage but would not invalidate or contradict any existing rules, since they reflect structural facts of the language rather than artifacts of a small sample. For example, the following Grew rule was developed to attach modifying adjectives to their governing nouns and to add the dependency relation `mod` between them:

```
rule r1 {
  pattern { X[upos=ADJ]; Y[upos=NOUN];
    X < Y } without { * -> X }
  commands { add_edge Y -[mod]-> X }
}
```

This semi-automatic workflow ensures consistency across the corpus while allowing manual intervention for complex or ambiguous constructions.

3.4 Production of the UD treebank

Conversion from mSUD to UD is performed in two stages: first, mSUD is converted to SUD, and then SUD is converted to UD. Both conversions are encoded as a set of Grew (Guillaume, 2021) rules that are applied iteratively to make the necessary annotation changes. Figure 2 illustrates the process on one sentence of the treebank.

The first conversion (from mSUD to SUD) involves merging inflectional suffixes with the root word to which they are attached. Several rules are

[11] https://arborator.grew.fr/

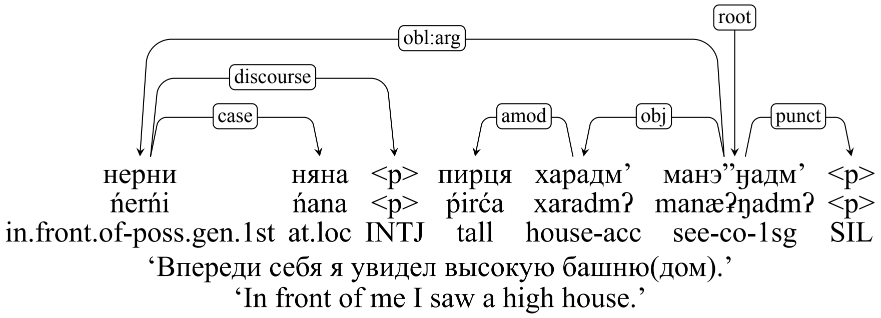

root
obl:arg
discourse
case
amod
obj
punct
нерни няна <p> пирця харадм’ манэ”ŋадм’ <p>
ńerńi ńana <p> ṕirća xaradmʔ manæʔŋadmʔ <p>
in.front.of-poss.gen.1st at.loc INTJ tall house-acc see-co-1sg SIL
‘Впереди себя я увидел высокую башню(дом).’
‘In front of me I saw a high house.’

Figure 1: Example tree from the Nenets UD treebank with Latin transliteration (2nd line), English glosses (3rd line) and fluent translation to Russian and English.

designed to ensure consistent glosses, lemmas and sound alignment after merging. After this first step, the word tokenization is as expected by the UD framework.

The second step relies on the general SUD to UD conversion that is described in (Gerdes et al., 2021). The syntactic structure is modified on the one hand to take into account the different choices in UD and SUD for selecting the head of a phrase in the dependency structure.[12] On the other hand, rules are used to map the SUD dependency relation tagset to the equivalent UD tagset.

The UD-annotated data produced is then validated using the process provided by the UD infrastructure.[13] These validation steps helped to identify inconsistencies in the original mSUD annotation and adapt the conversion rules to annotation choices specific to the Nenets treebank.

3.5 Information-structural roles annotation

A new initiative within the UniDive COST Action (CA21167) seeks to extend the UD framework by incorporating a layer for Information Structure (IS) annotation, drawing inspiration from the Prague Dependency Treebank 2.0.[14] In this approach, IS is treated as a functional phenomenon grounded in meaning, reflecting how speakers organize and interpret content within discourse rather than how it is formally encoded, therefore, we aim to tag IS roles in the treebank to support further formal and functional typological research.

The explicitness of IS annotation will be ensured through a detailed guideline currently under development. This guideline provides clear definitions, diagnostics, and instructions for annotators, allowing IS categories to be assigned systematically and reproducibly rather than impressionistically. Although the framework is still a work in progress and not the focus of the present paper, it reflects established best practices demonstrating that semantic and discourse-level annotation can be made reliable through well-formulated operational criteria.

In Tundra Nenets, certain IS roles are partially encoded morphologically: transitive verbs can carry suffixes marking both the person and number of the subject as well as the number of the object. Object agreement, in particular, indicates the topicality of third-person objects (Dalrymple and Nikolaeva, 2011; Nikolaeva, 2014), compare (1), where the verb agrees only with the subject, with (2), where the object is topical and the verb cross-refers its number (in addition to subject agreement).

(1) a. What did Pavel do?
Whom did Pavel see?
b. Павел Ирина-м’ манэ”ŋа-сь.
Pavel Irina-acc see.3sg-pst
‘Pavel saw Irina.’

(2) a. What did Pavel do to Irina?
b. Павел Ирина-м’ манэ”ŋа-да-сь.
Pavel Irina-acc see.3sg-sg.o-pst
‘Pavel saw Irina.’
or ‘As for Irina, Pavel saw her.’

This project adapts these insights to systematically annotate IS roles at the morpho-syntactic level, providing both a practical framework for the treebank and a model for cross-linguistic comparison.

Building on this foundation, we initiated Information Structure (IS) annotation in the Tundra Nenets UD treebank using a simple, broadly semantic scheme that captures the most fundamen-

[12] For example, the ADP is the head of the prepositional phrase it introduces in SUD, whereas in UD, this ADP depends on the main noun.

[13] https://universaldependencies.org/contributing/validation.html

[14] https://ufal.mff.cuni.cz/prague-dependency-treebank

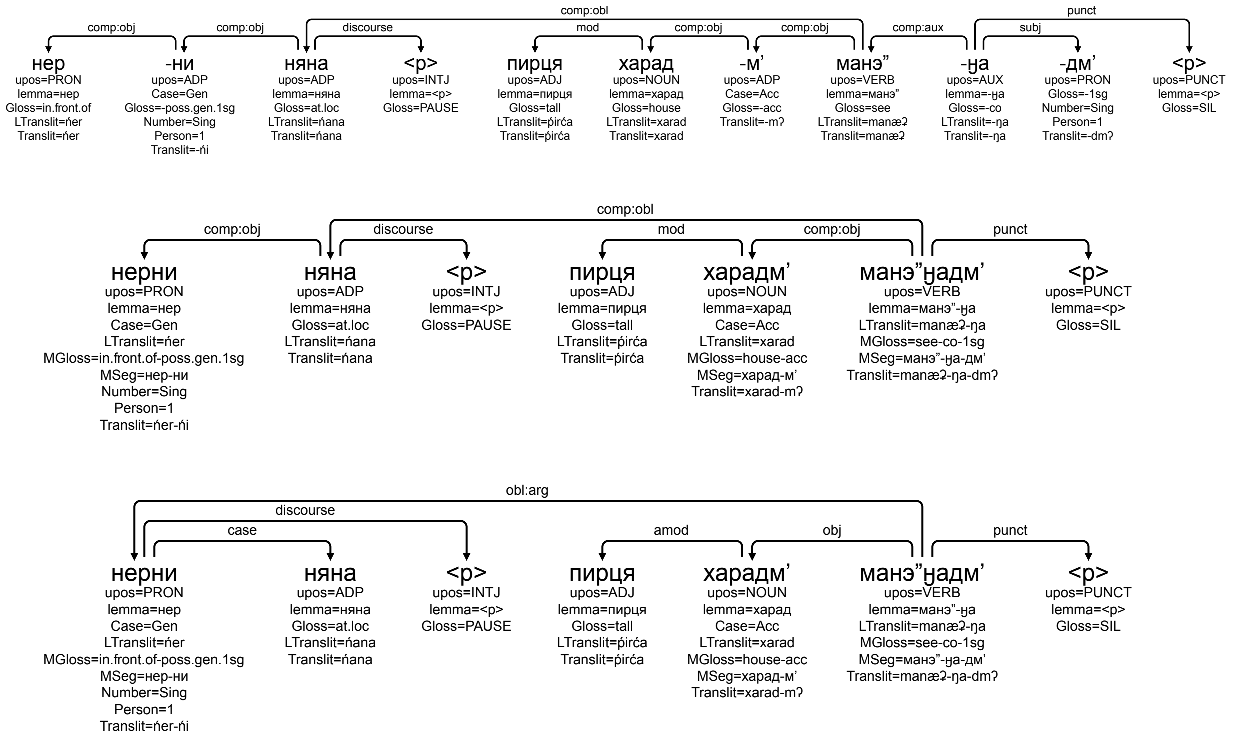


Figure 2: mSUD, SUD and UD annotations of the sentence from Figure 1.

tal distinctions observable in the corpus. We assume that IS distinctions are a universal aspect of language: all languages can differentiate contextual uses of utterances. Since these distinctions are not always overtly encoded, IS is treated primarily as a semantic phenomenon, reflecting how speakers structure and interpret information in discourse rather than as a property directly observable in form.

As a starting point, the annotation will eventually focus on topical third-person objects that trigger agreement, which are intended to be marked in the `MISC` column.[15] However, this annotation is not included in the current release and will be added in a future version of the treebank, once the practical framework for coding and placement in the tree is finalized.

3.6 Transliteration and translation

In addition to the Cyrillic transcription, which makes the data comparable with written Tundra Nenets resources, all annotated texts are also accompanied with a Latin-based transliteration in order to make them accessible to researchers who are not familiar with the Cyrillic script. The transliteration is generated automatically using the *translit* Perl toolkit.[16] The scheme mostly follows a 1-1 mapping between Cyrillic and Latin letters (including some characters that are not used in English, such as ŋ, æ, or the two glottal stops, ʔ and ʡ). Palatalized consonants are an exception: In the Cyrillic writing system, palatalization is often encoded in the following vowel, while in transliteration, we indicate it with an accute accent over the consonant. For example, *няна → ńana*.

Besides word-level English glosses, sentence-level manual translations into both English and Russian are also included. See the example tree in Figure 1.

4 Conclusions and prospects for automation

This study has presented the theoretical and methodological basis of the (Tundra) Nenets Universal Dependencies (UD) Treebank. By adapting the UD and morphologically enhanced Surface-

[15] Although one reviewer suggests that our annotation of topical objects relies only on formal features, this is not the case. In this language, object agreement is a grammatically encoded and semantically motivated marker of topicality. We use it because it directly expresses an IS value, not as a formal shortcut; the morphology itself reflects the discourse status of the argument.

[16] https://github.com/dan-zeman/translit

Syntactic UD (mSUD) frameworks to a morphologically complex Uralic language, the project established reproducible procedures for the syntactic annotation of spoken data. A particular emphasis was placed on spoken data annotation, the treatment of morphology-based syntactic relations, and the annotation of Information Structure roles. The objective of this focus was to develop a consistent and extensible annotation model.

At this stage, the Tundra Nenets treebank remains a manually annotated, small-scale resource. Subsequent endeavors will concentrate on the development of semi-automatic and fully automatic tools for lemmatization, part-of-speech tagging, and morphological analysis to facilitate corpus expansion while maintaining internal consistency and analytical precision.

Beyond its immediate scope, the project offers broader methodological insights for representing spoken and morphologically rich languages within the UD framework. The procedures and conventions developed for Tundra Nenets can be extended to other Samoyedic and Siberian Uralic languages. This contributes to a more balanced typological coverage in UD and advances the treatment of underrepresented language types in computational annotation.

Acknowledgements

This work has been supported by COST Action (CA21167) through two STSM grants: to the Faculty of Mathematics and Physics, Charles University (Beyond Syntax: Integrating Information Structure into UD Treebanks – Tundra Nenets) and to the Modyco lab, Paris Nanterre University (The syntax–prosody interface and harmonization of spoken data's syntactic annotation – Tundra Nenets). The development of the Tundra Nenets treebank was also supported by the projects Autogramm: Induction of Descriptive Grammar from Annotated Corpora (ANR-21-CE38-0017) and ThEA: Theoretical and Experimental Approaches to Dialectal Variation and Contact-Induced Change (NKFIH FK 129235), which contributed to data collection and treebank creation. Daniel Zeman's work has been supported by *LINDAT/CLARIAH-CZ*, project No. LM2023062 of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

Paul Boersma and David Weenink. 2025. Praat: Doing phonetics by computer [computer program]. `https://praat.org`. Version 6.4.45, retrieved 12 October 2025.

Josefina Budzisch and Beáta Wagner-Nagy. 2024. INEL Nenets corpus. version 1.0.

Svetlana Burkova. 2022. Nenets. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford guide to the Uralic languages*, pages 674–708. Oxford University Press, Oxford.

Wallace L. Chafe, editor. 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Advances in Discourse Processes, vol. III. Ablex, Norwood, NJ, USA.

Mary Dalrymple and Irina Nikolaeva. 2011. *Objects and Information Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kaja Dobrovoljc. 2025. Counting trees: A treebank-driven exploration of syntactic variation in speech and writing across languages. *arXiv preprint arXiv:2505.22774*.

Ethnologue. 2009. Ethnologue: Languages of the world.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *TLT 2019 - 18th International Workshop on Treebanks and Linguistic Theories*, pages 126–132, Paris, France. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank? go SUD! In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 35–46, Sofia, Bulgaria. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia. ELRA and ICCL.

Péter Hajdú. 1968. *The Samoyed peoples and languages*, volume 14 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University, Bloomington. 2nd edition.

Veronika Hegedűs, Nikolett Mus, and Balázs Surányi. 2021. Tense, agreement and copula drop in Tundra Nenets copular clauses.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages pp–35. Association for Computational Linguistics.

Nikolett Mus. 2023a. Nenets. In Daniel Abondolo and Riitta-Liisa Valijärvi, editors, *The Uralic Languages: Second Edition*, pages 853–896. Routledge, London.

Nikolett Mus. 2023b. Tundra Nenets. In Anja Behnke and Beáta Wagner-Nagy, editors, *Clause Linkage in the Languages of the Ob-Yenisei Area. Asyndetic Constructions.*, pages 133–174. Brill.

Nikolett Mus and Réka Metzger. 2021. Toward a corpus of tundra nenets: stages and challenges in building a corpus. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2, pages 4–9.

Nikolett Mus and Balázs Surányi. 2021. Post-verbal phrases and their correlates in Tundra Nenets.

Nikolett Mus and Balázs Surányi. 2025. Postverbal phrases in Tundra Nenets: An empirical study of rigid verb finality. Under review at *Folia Linguistica*.

Irina Nikolaeva. 2014. *A Grammar of Tundra Nenets*, volume 65 of *Mouton Grammar Library*. De Gruyter Mouton, Berlin / Boston.

Irina Nikolaeva and Edward Garrett. 2014. Online documentation of siberian languages. Audio resource.

Tapani Salminen. 1998. Nenets. In Daniel Abondolo, editor, *The Uralic Languages*, pages 516–547. Routledge, London.

Tapani Salminen. 1999. Tundra Nenets. online.

Natal'ja M. Tereshchenko. 1966. Neneckij jazyk. In V. I. Lytkin and K. E. Majtinskaja, editors, *Jazyki narodov SSSR. Volume 3: Finno-ugorskie i samodijskie jazyki*, pages 376–395. Nauka, Moscow / Leningrad.

Natal'ja M. Tereshchenko. 1973. *Sintaksis samodijskix jazykov: prostoe predloženie [The syntax of Samoyedic languages: The simple clause]*. Nauka, Leningrad.

Eva Toulouze. 1999. The beginning of a written culture by the Uralic peoples of the north. *Pro Ethnologia*, 7:52–85.

Benchmarking Large Language Models for Lemmatization and Translation of Finnic Runosongs

Lidia Pivovarova
University of Helsinki

Kati Kallio
Finnish Literature Society

Antti Kanner
University of Turku

Jakob Lindström
University of Helsinki

Eetu Mäkelä
University of Helsinki

Liina Saarlo
Estonian Literary Museum

Kaarel Veskis
Estonian Literary Museum

Mari Väina
Estonian Literary Museum

Abstract

We investigate the use of large language models (LLMs) for translation and annotation of Finnic runosongs—a highly variable multilingual poetic corpus with limited linguistic or NLP resources. We manually annotated a corpus of about 200 runosongs in a variety of languages, dialects and genres with lemmas and English translations. Using this manually annotated test set, we benchmark several large language models. We tested several prompt types and developed a collective prompt-writing methodology involving specialists from different backgrounds. Our results highlight both the potential and the limitations of current LLMs for cultural heritage NLP, and point towards strategies for prompt design, evaluation, and integration with linguistic expertise.

1 Introduction

Runosongs are a versatile oral tradition common to most Finnic languages, including South and North Estonian, Votic, Ingrian, Karelian, Ludic, and Finnish. The recently combined corpus of approximately 250,000 texts, recorded between 1564 and 1971 (Janicki et al., 2024b) offers an unprecedented opportunity for computational study. However, the corpus exhibits substantial linguistic, orthographic, and poetic variation, including more than one million distinct word forms.

The runosong corpus covers multiple languages in their non-standard dialectal variants, with blurry borders and multilingual overlap. The texts often use archaic vocabulary and word forms and exhibit considerable poetic parallelism. Karelian, Ludic, Ingrian, and Votic developed written standards only in the late 20th century, and dialects may be written in several ways, leaving much of the corpus in various non-standard orthographies. The orthography of the Estonian part of the corpus has been normalized manually, while all dialectal features have been retained.

Thus, unlike mainstream NLP benchmarks, runosongs involve low-resourced languages, dialectal variation, archaic or poetic morphology, and non-standard orthography. The data exhibit high morphological variation in both suffixes and stems, archaic word forms not attested in contemporary usage, and orthographic inconsistencies across centuries and regions. No dictionaries, parsers, or NLP tools cover the entire corpus.

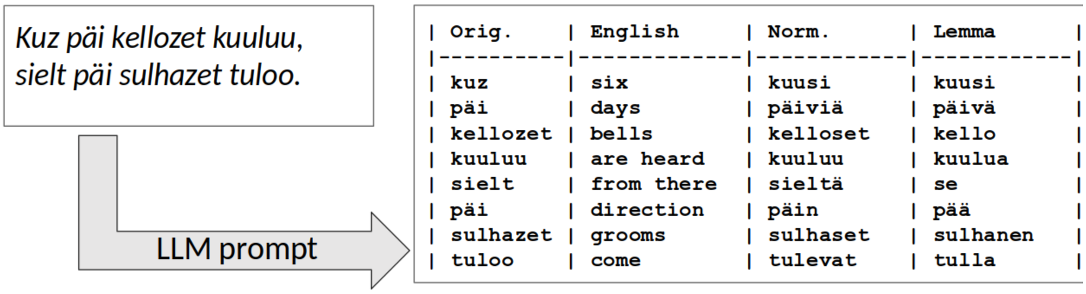
Recent progress in large language models (LLMs) raises the question of whether such models can support the analysis and translation of these kinds of challenging texts. This paper addresses the methodological challenges of applying LLMs to this material, with a focus on translation, orthography normalization, lemmatization, and etymological annotation.

Our approach is illustrated in Figure 1: as an input, an LLM gets a runosong text and a prompt and outputs a structured table, where each word is lemmatized and translated into English. From the very beginning, we observed that modern state-of-the-art LLMs have an impressive ability to understand runosongs, so we chose the way of prompt-engineering and using the largest available models, rather than training or finetuning smaller specialized models. At the same time, we also noticed that models have a high sensitivity to small changes in the prompt, a tendency to hallucinate analyses for unknown words, and inconsistency in outputs across a variety of inputs. Thus, this work is focused on (i) building a representative manually annotated evaluation dataset and (ii) creating extended linguistically motivated prompts that constrain model behavior to get consistent results.

The specific contributions of this work are

- A manually annotated dataset of about 200 runosongs, drawn through stratified random sampling to ensure maximal variety of Finnic languages, dialects, and orthographies;

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 87–105
December 10-12, 2025 ©2025 Association for Computational Linguistics



```
| Orig.    | English      | Norm.       | Lemma       |
|----------|--------------|-------------|-------------|
| kuz      | six          | kuusi       | kuusi       |
| päi      | days         | päiviä      | päivä       |
| kellozet | bells        | kelloset    | kello       |
| kuuluu   | are heard    | kuuluu      | kuulua      |
| sielt    | from there   | sieltä      | se          |
| päi      | direction    | päin        | pää         |
| sulhazet | grooms       | sulhaset    | sulhanen    |
| tuloo    | come         | tulevat     | tulla       |
```

Figure 1: An illustration of our pipeline: on the left is a runosong in Livvi Karelian, on the right the same text processed with an LLM (DeepSeek-R1-BF16 in this case): for each input word, the model returns its English translation, the same word in standard orthography, and lemma. First two words are misinterpreted, and there are issues in normalizing.

- A set of linguistically and culturally informed prompts, developed by linguists and folklorists to most efficiently process runosong data;

- A series of benchmarking experiments with 5 open and 1 proprietary language models, and a variety of prompt pipelines that allows to grasp the challenges of working with non-standard dialects, archaic forms, and complex poetic structures[1].

2 Related Work

The question of how LLMs (mis)represent or (un)master small, minority, indigenous, or endangered languages, and whether they may be useful for scholarly analysis, everyday use, or language revitalization, is a broad one. These languages are often low or ultra low resourced in terms of NLP tools, language description, dictionaries, or available digital texts needed for manual analysis, model training or fine-tuning existing models. Further, available texts may be of non-standard, sensible or historical character, and explanations for their cultural and contextual characteristics may not be available (Aepli, 2024; Lamb et al., 2025; McGiff and Nikolov, 2025; Moshagen et al., 2024). Wiechetek et al. (2024) point out that representing a language or content in an indigenous language via LLMs incorrectly is neither beneficial nor ethical—but neither is digital marginalisation (Paul et al., 2024).

Along recent rapid development of large language models, researchers have started to test their usability for languages and language variants with low resources and poor digital representation (Joshi et al., 2024; Shu et al., 2024; Uzun, 2025), also with harmonising, lemmatizing or translating (Natale et al., 2025; Vidal-Gorène et al., 2025; Alam and Anastasopoulos, 2025; Riemenschneider, 2025). Some experiments, adaptations and evaluations have already been conducted for smaller Finno-Ugric languages (Kuulmets et al., 2025; Partanen, 2024; Pirinen, 2024; Purason et al., 2025) and historical, dialectal, non-standard, or poetic folklore materials (Meaney et al., 2024; Lamb et al., 2025; Burda-Lassen, 2023; Rodriguez and Bernardes, 2025; Tsutsumi and Jinnai, 2025; Xu et al., 2024); for rule-based linguistic analysis vs. LLMs, see Pirinen (2024). The performances vary in terms of task, target language, and text genre, as models have different language combinations in their training data.

Although previous runosong research has employed various computational approaches, tasks requiring linguistic annotation — particularly lemmatization — have still relied on manual work. Harvilahti (1992), Ross (2015) and Saarinen (2018) lemmatized by hand their areal or singer based corpora for subsequent analysis. Computational folkloristics has explored with combining automatic translation and domain specific word lists for multilingual data (Meder et al., 2023) and various NLP options (Al-Laith et al., 2024; Tangherlini and Chen, 2024).

Our earlier and ongoing computational projects on runosongs have applied corpus-based methods, especially line, passage and text similarity recognition based on clustering based on cosine similarity of character bigram vectors (Janicki et al., 2023; Seláf et al., 2025), and alignment similarity (Janicki, 2022, 2023), to analyse e.g. oral intertextuality (Sarv et al., 2024), oral-literary relationships (Mäkelä et al., 2024), dispersion of frequent lines (Janicki et al., 2024a) and regional variation (Kallio, 2024). Previous computational studies

[1] The dataset, the code and the prompt are freely available at https://github.com/hsci-r/filter-llm-lemmatization

have also included analyses on the basis of various data queries (Harend, 2024; Kallio et al., 2024; Veskis, 2025), analysis of metadata (Kallio et al., 2023), verse structures (Sarv, 2015, 2019; Sarv et al., 2021), topic modelling (Sarv, 2020), stylometry, and network analysis (Sarv and Järv, 2023).

3 Data Sampling and Annotation

As our first contribution, we created a manually annotated evaluation set for the linguistic analysis of Finnic runosong texts. As the material is both extremely heterogeneous and highly skewed—pre-19th-century texts are scarce, local genre distributions vary, recorders preferred particular regions and topics—we sought to sample the diversity of the data in a stratified manner instead of through pure random sampling at the level of the whole corpus. Ideally, we would have liked to sample several examples of each dialect and orthographic variety. However, the corpus metadata does not have direct information about dialects. Thus, we made a mapping between the parish where the text was collected and the most probable languages and dialects in which the text could have been performed.

To establish the initial correspondence between spoken dialects and parishes, the available dialectal sources were first compared to determine which aligned most closely with the temporal framework and categorization of the runosong data in *Eesti regilaulude andmebaas* ERAB (Oras et al., 2003) and *Suomen Kansan Vanhat Runot* SKVR(Saarinen and Krikmann, 2004). For Finnish material, the open-source dialect map produced by the Institute for the Languages of Finland was adopted, as it constitutes a broadly accepted compromise to represent dialectal boundaries (Institute for the Languages of Finland, 2020). For the Estonian data, *Eesti murded ja kohanimed* was selected on equivalent grounds (Pajusalu et al., 2018). These were adopted for the slightly different parish division of the Northern Finnic source data, and added with information from other sources for Karelian, Ludic, Ingrian and Votic languages. Most dialect references only record the main variety in each area, creating an overly homogeneous understanding of the situation and, thus, may obscure the presence of minority languages in multilingual areas. Languages and dialects are not evenly distributed in our corpus: e.g. Votic is represented by less than 400 texts mostly from one parish. We took into account Karelian dividing into three, Ingrian into two or three, Estonian into nine, and Finnish into eight dialectal areas, and also checked for the genre distribution in our sample.

We then grouped the data according to the most probable dialect in which the texts could have been written. From each group, we randomly sampled 7 texts. For the Finnish part of the data, we further constrained the sample so that 3 of the texts were collected before 1800—if less than 3 were collected in a parish group before 1800 we selected all available texts. This resulted in a collection of 280 texts, 216 of which were later annotated. In addition to word-by-word annotation, the dialect and genre of each text were determined. Basic statistics grouped by broader dialect area and time of the resulting evaluation set are described in Table 1. Note, that a distribution of languages and dialects in the manually annotated set is not representative of the overall distribution in the whole corpus; it was deliberately skewed to incorporate more difficult instances.

	#texts	#verses	#words
North Estonian	62	1113	4213
South Estonian	30	530	2093
<1800 Finnish	63	982	2973
>=1800 Finnish	25	235	764
Karelian	17	636	2030
Ingrian	8	272	893
Votic	6	144	418
Ludic	4	30	109
Swedish	1	4	20
Total	**216**	**3946**	**13513**

Table 1: Manually annotated corpus statistics

The annotations were carried out by three specialists in Finnic folklore (Kallio, Saarlo, Väina). For each input word, the annotators were asked to provide the following fields:

- `normalized`: the word in modern orthography,
- `local`: the lemma in original local language variant, based on the recorded text,
- `standard`: the lemma in modern Finnish or Estonian (depending on part of the corpus), if it corresponds to the original in stem; otherwise, the lemma in original language variant,
- `root`: the etymological root, a modern word that can serve as a key in an etymological dictionary,
- `translation`: the literal translation into English, as a semantic layer.

The annotators were allowed to use any available resources, e.g. dictionaries, grammars and descriptions of the relevant dialects. However, usage of any LLM during annotation was forbidden. At the beginning, a few texts were annotated collectively to establish common guidelines, which are presented in Appendix 7. The rest of the data were annotated by a single annotator most familiar with the corresponding language. Using only one annotator per text was a practical issue: we preferred a larger annotation set to a smaller one with two annotators. We also opted for relatively quick lemmatization rather than the thorough scholarly analysis that our most difficult texts often require. Difficult cases were discussed throughout the work. Finally, members not involved in the annotations performed a spot checks for the lemmas.

4 Prompt Implementation

Prompt engineering for this project was carried out collaboratively by experts in folkloristics, linguistics, and data science. Early on, we had noticed that giving a model a detailed prompt, explaining, e.g., morphological peculiarities of runosongs or some cultural context improves outputs. Such prompts obviously should be written by domain specialists. At the same time, we noticed that output consistency can be improved by including certain constraints, e.g., very specific output table formats or lists of input words. These parts are easier to write and correct by those who directly implement the pipeline and run the dataset processing automatically. In addition, some texts are too long to be processed by an LLM in one run, so they need to be chunked and the chunking also mentioned in the prompt. Finally, we also want to experiment whether some additional steps—e.g. prompting to translate the whole poem into English before processing it word by word—improve the analysis result.

To do this, we created the prompts to follow a modular system, making it easy to make different combinations. Different parts were created by different team members.

The specific prompt engineering, especially the development of the largest domain-specific parts, was implemented as a creative process where team members played with different models—mostly with Claude, some experiments with ChatGPT—via their web interfaces, trying to analyze a small set of texts and qualitatively assess the results. Prompts were iteratively refined to address systematic errors, e.g. specifying dialect, archaic case forms, and poetic context. Overall, this was a creative process where different ideas were tried and refined. The goal of this stage was to come up with the most promising ideas of what should and should not be included in the prompt.

The resulting prompts were collected and organized into smaller text files, e.g. ”cultural context”, ”phonological variation”, ”output format”, etc. Some were prepared in two versions, one for the Northern—Finnish, Karelian, Ingrian and Votic—and one for the Southern—North and South Estonian—part of the corpus. The prompts were organized in *pipelines*, which specify what text files, in which order, should be included into the prompt. Pipelines are organized in stages, for the cases when a text can be processed sequentially—e.g. first translate then make a table. An example pipeline is shown in Listing 1 and the corresponding prompt is shown in Appendix B.

```
{    "system": "system/main_system.txt",
    "steps": [
        {
            "name": "table_only",
            "task_prompts": [
                "context/general_{lang}.txt",
                "context/cultural.txt",
                "context/poetic.txt",
                "context/linguistic_{lang}.txt",
                "context/phono_{lang}.txt",
                "format/table_format.txt",
                "task/table_{lang}.txt",
                "input/input.txt"
            ],
            "chunking": {
                "chunk_notice": "connectors/
                chunk_notice.txt"
            },
            "validation": {
                "enforce_first_column": true,
                "min_table_cols": 7
            }
        }
    ]
}
```

Listing 1: Modular prompt pipeline (JSON)

5 Experiments

5.1 Setup

The processing setup is shown schematically in Figure 2. Since all LLMs have limitations for the number of intput tokens and output tokens, and many runosongs are too long to produce a single output table, they are split into chunks, each chunk containing k verses, 4-6 tokens per verse. Then a model is prompted with the task-describing prompt, a runosong text and a list of words that should be analysed for each chunk. When we get a model output we check whether the result table is well-formed, i.e. there is a row for each word, and all

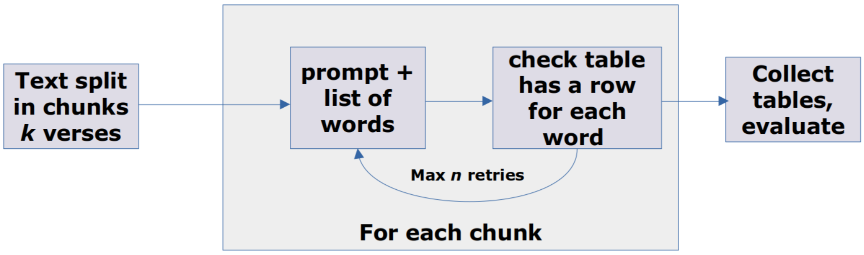


Figure 2: Data processing pipeline.

Label	HuggingFace path / source	number of parameters	chunk size	maximum output	Description
copy-word	-	-	-	-	A baseline that just copies the input word into each target
poro	LumiOpen/Llama-Poro-2-70B-Instruct	70B	25	4000	Open Finnish instruction-tuned model; strong Finnish-centric baseline for dialectal and low-resource varieties.
llama	meta-llama/Llama-3.3-70B-Instruct	70B	100	16000	High-quality multilingual dense model; main reference baseline.
databricks	databricks/dbrx-instruct	132B	25	8000	Large open Mixture-of-Experts model from industry; efficient large-scale architecture representative of current best practice.
mixtral	mistralai/Mixtral-8x22B-Instruct-v0.1	141B	100	8000	Open MoE model with strong reasoning and translation ability; good trade-off between quality and cost.
deepseek	unsloth/DeepSeek-R1-BF16	671B	100	16000	Massive reasoning-oriented MoE model; tests benefits of very high parameter capacity and long-context inference.
claude	Claude-3.7-Sonnet-20250219	unknown	25	8000	Closed commercial model accessed via Anthropic API; included for comparison with state-of-the-art proprietary systems in reasoning and translation quality.

Table 2: Models and hyperparameters used for benchmarking.

columns are filled. If this is not the case, we add an additional retrial note to the prompt and process the same chunk again, up to n times. We found that a model quite often outputs a more consistent result in the retry. However, if this does not happen at the first or second retrial, this indicates some major difficulty with this specific poem. Thus, we set $n = 2$ in all our experiments.

As for the chunk size, it was set separately for each model, together with the maximum number of output tokens. Both parameters definitely affect model output, in addition to its efficiency. E.g., setting too long output token limit may trigger hallucinations and yield worse results than a more constrained output. On the other hand, too low output limit may result in failure to process some texts, due to their peculiarities. Nevertheless, for this paper we fix hyperparameters for each model and focus on a comparison of prompts and pipelines. The model and the hyperparameters used are shown in Table 2.

The initial impression from our manual experiments was that the Anthropic model Claude 3.7 yields significantly better results than ChatGPT. Thus, we use the former as our proprietary model benchmark.

We also added a ”copy-word” baseline, that copies an input word for each output column.

5.2 Pipelines

Based on our initial experiments of different ways to affect model performance, for numerical experi-

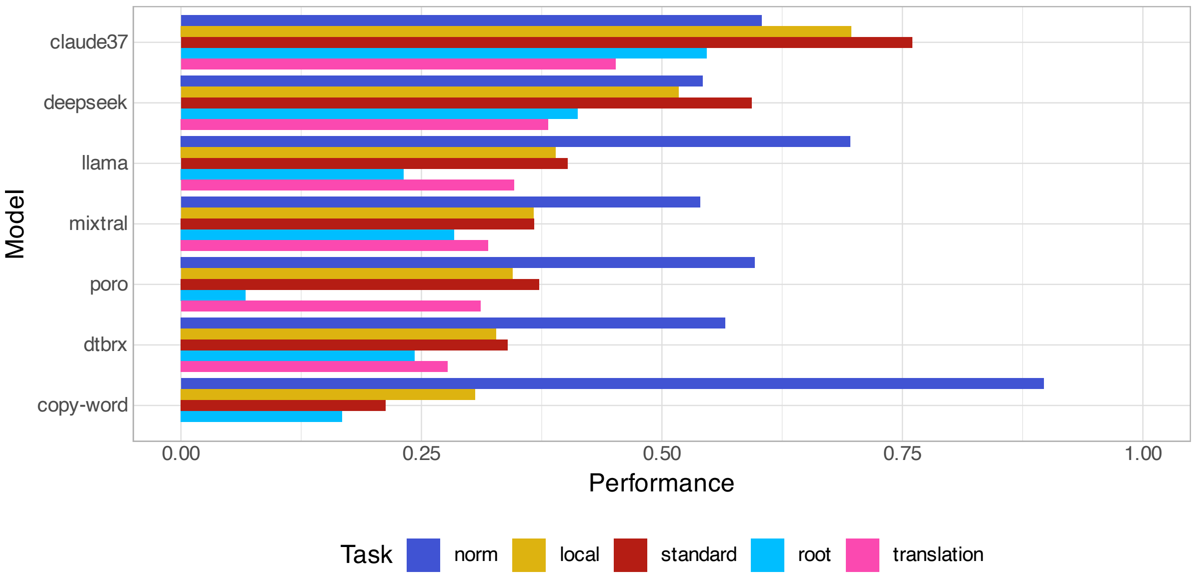


Figure 3: Overall results, averaged across all texts in the collection. We show exact string match for all fields except for English translation, where we show cosine distance between ground truth and model output embeddings. We show the best performing pipeline for each model-task pair.

ments, we chose the 5 following pipelines:

- `table only`: a model is asked to directly run the main task, i.e. word-by-word table analysis;
- `translation and table merged`: a model is prompted to produce a verse-level English translation of the runosong and then output the table; the hypothesis here is that translating the full text first would lead to a better understanding of the text's semantics, which can improve the quality of subsequent analysis;
- `translation and table sequentially`: the difference with the previous approach is that here we perform two model calls and use output of the first stage—i.e., translation—as part of the input prompt for the second stage;
- `translation -> fix -> table`: here we add one intermediate stage and prompt model to analyze the verse-level translation and correct it where necessary;
- `translation -> fix -> table -> fix`: we add one more self-correction step, prompting the model to correct the table produced in the previous step.

Each pipeline we test in two variants: with linguistic information (as exemplified in Appendix B) and without such information, i.e. relying only on the internal knowledge a model may possess.

5.3 Evaluation

Most of the fields in the output table—normalized word, lemma in a modern language, etc.—are suitable for exact comparison. For these fields, we use accuracy, i.e. a percentage of cases where a model output is exactly the same as a manual annotation.

The only exception is an English translation field, where semantic similarity is more appropriate than an exact match. For this field we use *cosine similarity* between embeddings for a manual translation and a model output, using an English model from the Spacy library[2].

6 Results and Discussion

Even though they seemed to be working in our preliminary experiments, in the end, we did not find any benefit to adding translation or fix stages to the pipeline, neither given in a sequence nor as part of a merged prompt. For the best-performing models, there was essentially no difference in numerical results, and for the smaller, more poorly-performing models, the adding of steps actually usually hindered performance. We also observed that in some cases the full-text translation was missing from the model's output, despite being explicitly prompted. Thus, in the following, we only report performance on the simple ”table only” prompt. In the future, though, we will analyze the results in more detail

[2] https://spacy.io/models/en#en_core_web_lg

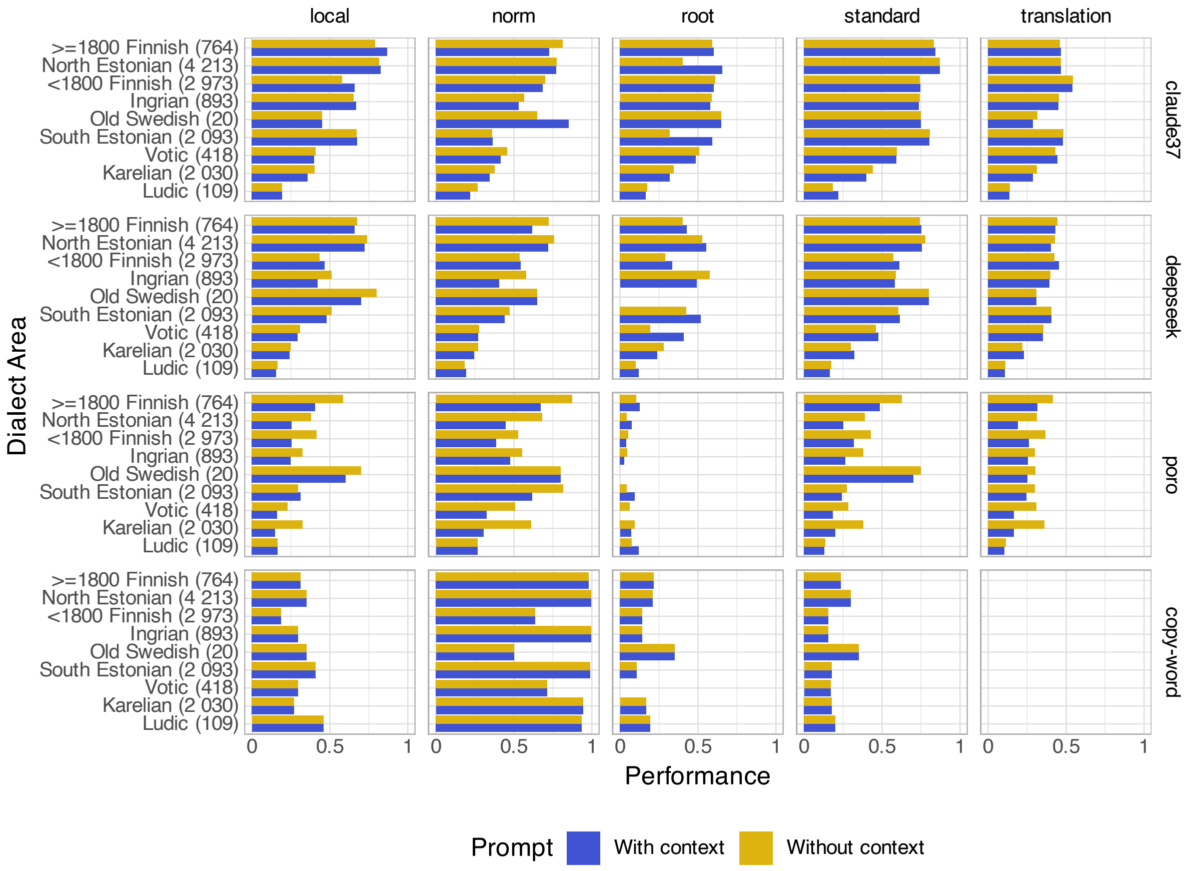


Figure 4: Results for three models and the copy-word baseline grouped by task, language area and whether the prompt includes contextual information. We show the best performing pipeline for each language-context pair. A number of running words in the evaluation set for each language is shown in parenthesis.

for the best-performing models to see, e.g., whether the several step approach solves some issues but causes new ones, seeking to explain the difference in initial experiments and our final results.

As can be seen in Figure 3, Claude was the best model for the standard and local lemma, etymological root and English translation. The biggest free model—Deepseek—performs second-best in these fields, though the difference between Deepseek and Claude is significant. E.g. for the standard lemma, the averaged Claude performance is 76% accuracy on average, while Deepseek yields 66% accuracy for this field, which results in a difference of 10 percentage points. Other models perform much worse, and the performance seems to correlate with the model size—smaller models scarcely outperform the copy-word baseline, while larger models double or triple the performance.

For normalization, no models outperform the copy-word baseline. In the data, 92% of the words need no normalization. Here, all models seem to be over-eager, assuming that something needs to be done. In a brief check, we observed that, beyond merely normalizing the spelling system, Claude systematically removed dialectal and archaic features despite explicit instructions not to do so (e.g., *einämaalta* > *heinamaalt* 'hayfield'), and even altered the roots (*ubin* > *õun* 'apple').

The etymological root seems to be the most problematic field for which we computed an exact match score—the scores for translation are not directly comparable. Despite the prompts having an exact definition of this field—a main word form that would serve as a dictionary entry in the etymological dictionary—the models still struggle to understand the task. In some cases, a "proto-Finnic" root is returned for this field; in other cases, a morphological stem is returned instead of the full word. These results vary: models seem to use different definitions to process different texts, though outputs for a single song are usually consistent.

In Figure 4, we show performance separately for the main languages in the collection, for the two strongest models, as well as Poro as a representative example of the smaller models. As can be seen here, the performance is best for the dialects resembling modern Finnish (>=1800 Finnish) and Estonian (North Estonian), with performance in Ingrian being surprisingly high, probably due to orthographic and linguistic closeness to Finnish. In contrast, the performance in Karelian is surprisingly poor for the two otherwise best models. This may relate to the extremely varying orthography in our Karelian data, shortage of Karelian (as well as Ludic and Votic) materials online, some Karelian phonemes not present in Finnish or Estonian, and the existence of three main Karelian language variants, each with their own recent standardisation processes. Poro, the smallest and least powerful model overall, achieves slightly higher scores for some outputs on Karelian, as well as for normalization—likely reflecting a better grasp of our intended orthographic normalization.

Figure 4 also shows results for prompt pipelines with and without linguistic context. Our results do not indicate any systematic improvement from providing contextual information - adding a detailed context can either increase or decrease performance, and the impact varies a lot across model-language pairs. The clear difference for both Claude and Deepseek can be seen for only for the `root` in both South and North Estonian. Models appear to recognize that roots should be provided in Estonian when the prompt includes contextual information; however, in prompts without context, the model often returns variable results (in Estonian, Finnish, Proto-Finnic or stem only).

The fact that our efforts contributed into prompt engineering resulted in mostly negative outcomes so far is discouraging. However, not all differences in pipelines can be seen in numerical evaluation. The initial manual analysis of the outputs reveals, for instance, the following problems:

- Confusing normalization with standardization, i.e. replacing dialectal or minority language forms with modern Finnish or Estonian.
- Substituting with common synonyms rather than producing faithful lemmatization.
- Misinterpreting archaic morphological forms.
- Inconsistent handling of homonymy and dialectal variation: not recognizing dialectal or minority language words and mixing them with their homonyms in the major languages.
- Refusal to analyse obscene or culturally marked content.
- Difficulty with recognizing onomatopoetic or nonsense words, and refrains, i.e. recurrent words with meanings separate from the main text.

Table 3 shows a few initial lines of a translation of a South Estonian text produced by the Claude model. Table 4 in Appendix C shows word-by-word analysis produced for the same text. This example confirms our preliminary impression that models—especially Claude—are, by and large, interpreting the text correctly. The main challenges lie in the possible alternative interpretations and in the precise formatting of the output. As noted above, we have not yet performed a systematic review of the results; this is planned for future work.

7 Conclusion

This study explored how large language models handle the linguistic, poetic and cultural complexity of Finnic runosongs. Using a manually annotated benchmark and structured prompt pipelines, we examined how far current models can go without fine-tuning. The early results are promising but uneven. Model choice seems to matter more than prompt design: large, general-purpose models provide the most reliable outputs, while smaller ones occasionally handle simpler normalization tasks more consistently. Multi-stage or translation-first pipelines do not yet yield systematic improvements in numerical evaluation. The results for adding linguistic and cultural contextual information vary depending on the model and task. This information needs to be refined according to dialect and actual linguistic variation in the data in future experiments. In our very initial experiments, it also looks promising to experiment with feeding the models with dictionaries, word lists, language descriptions, or other wider information for low resourced minority language parts of our data.

The next steps include evaluating the LLM errors further, testing the use of dialect–parish mapping information in prompts and adding explicit dialect detection as an intermediate step.Translation of whole texts (as opposed to word-by-word translations) would also be valuable, since they enable access to Finnic runosongs by broader audience not familiar with language varieties or Finnic languages at all, thus translations also need to be properly evaluated.

Yet, even with promising LLM results, this also poses ethical questions about partly misrepresenting the data and adding partly false LLM generated

original verse	**comments**	**English translation**
ku olli nuuri neiokõne	South Estonian dialect with diminutive form “neiokõne”	When I was a young maiden
kui ma kasvi kabokõne	“kabokõne” is a diminutive form of “kabo” (maiden, young woman)	When I grew up as a young girl
lätsi marja sis mäe päälõ	“lätsi” is South Estonian past tense form of “minema” (to go)	I went berry-picking on the hill
lätsi orgo ubinahe	“ubinahe” refers to apple orchard (illative case)	I went to the valley to the apple orchard
panni ma tuppõ tuima ravva	“tupp” = sheath, “tuim raud” = cold iron/steel (knife)	I put the cold steel in the sheath
vaivaväidse panni vüü ala	“vaivaväits” = poor/miserable knife, “vüü ala” = under the belt	I put the poor knife under my belt

Table 3: A few initial lines of a translation table produced by the Claude model. All text in the table is produced by the model, including the comments column.

material on low resource minority languages online, potentially affecting both future manual interpretations and LLM development.

LLMs already show potential to support linguistic and cultural annotation of complex poetic materials, but they require clearly defined tasks, transparent evaluation, and close collaboration between computational and domain experts. Our goal is to make the runosong corpus easier to explore and compare, without losing the precision and contextual depth that make it valuable in the first place.

References

Nora Aepli. 2024. *There Is Plenty of Room at the Bottom: Challenges & Opportunities in Low-Resource Non-Standardized Language Varieties*. Ph.D. thesis, University of Zurich.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and evaluation of pre-trained language models for historical danish and norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2025. Large language models as a normalizer for transliteration and dialectal translation. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–67, Abu Dhabi, UAE. Association for Computational Linguistics.

Olena Burda-Lassen. 2023. Machine translation of folktales: Small-data-driven and llm-based approaches. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 68–71, Gothenburg, Sweden. Association for Computational Linguistics.

Helina Harend. 2024. Ema-, isa-, õe- ja vennanimetused eesti regilauludes. Master’s thesis, University of Tartu, Faculty of Arts and Humanities, Institute of Estonian and General Linguistics.

Lauri Harvilahti. 1992. *Kertovan runon keinot. Inkeriläisen runoepiikan tuottamisesta*. SKS, Helsinki.

Institute for the Languages of Finland. 2020. Parishes, finland, estonia and other areas 1938, 1:1,000,000. CSC – IT Center for Science, `http://urn.fi/urn:nbn:fi:csc-kata00001000000000000203`.

Maciej Janicki. 2022. Optimizing the weighted sequence alignment algorithm for large-scale text similarity computation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 96–100.

Maciej Janicki. 2023. Large-scale weighted sequence alignment for the study of intertextuality in finnic oral folk poetry. *Journal of Data Mining and Digital Humanities, NLP4DH.*

Maciej Janicki, Kati Kallio, and Mari Sarv. 2023. Exploring finnic written oral folk poetry through string similarity. *Digital Scholarship in the Humanities*, 38(1):180–194.

Maciej Janicki, Kati Kallio, Mari Sarv, and Eetu Mäkelä. 2024a. Distributional criteria for identifying formulas in finnic oral poetry. In *Formulaic Language in Historical Research and Data Extraction*, pages 1–17, Amsterdam. Huygens Institute for History and Culture of the Netherlands, Royal Netherlands Academy of Arts and Sciences. International Institute for Social History, Amsterdam, 7–9 Feb 2024.

Maciej Janicki, Eetu Mäkelä, Mari Väina, and Kati Kallio. 2024b. Developing a digital research environment for finnic oral poetry. *Baltic Journal of Modern Computing*, 12(4):535–547.

S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani. 2024. Fine tuning llms for low resource languages. In *Proceedings of the 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 511–519, Dhulikhel, Nepal.

Kati Kallio. 2024. Vesi vanhin voitehista: Historiallisten merkityskenttien ja käyttöyhteyksien jäljillä. *Elore*, 31(2).

Kati Kallio, Maciej Janicki, Eetu Mäkelä, Jukka Saarinen, Mari Sarv, and Liina Saarlo. 2023. Eteneminen omalla vastuulla: Lähdekriittinen laskennallinen näkökulma sähköisiin kansanrunoaineistoihin. *Elore*, 30(1):59–90.

Kati Kallio, Mari Väina, Maciej Janicki, and Eetu Mäkelä. 2024. Bridging northern and southern traditions in the finnic corpus of oral poetry. *Folklore: Electronic Journal of Folklore*, 94:191–232.

Hele-Andra Kuulmets, Taido Purason, and Mark Fishel. 2025. How well do llms know finno-ugric languages? a systematic assessment. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 340–353, Tallinn, Estonia. University of Tartu Library.

William Lamb, Dong Han, Ondřej Klejch, Beatrice Alex, and Peter Bell. 2025. Synthesising a corpus of gaelic traditional narrative with cross-lingual text expansion. In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 12–26. Association for Computational Linguistics.

Josh McGiff and Nikola S. Nikolov. 2025. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv preprint*.

J.-A. Meaney, Beatrice Alex, and William Lamb. 2024. Evaluating and adapting large language models to represent folktales in low-resource languages. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 319–324. Association for Computational Linguistics.

Theo Meder, Petra Himstedt-Vaid, and Holger Meyer. 2023. The isebel project: Collecting international narrative heritage in a multilingual search engine. *Fabula*, 64(1-2):107–127.

Sjur N. Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Eetu Mäkelä, Kati Kallio, and Maciej Janicki. 2024. Sources and development of the kalevala as an example for the quantitative analysis of literary editions and sources. *Digital Humanities in the Nordic and Baltic Countries Publications*, 6(1):1–12.

Paolo Di Natale, Egon W. Stemle, Elena Chiocchetti, Marlies Alber, Natascia Ralli, Isabella Stanizzi, and Elena Benini. 2025. The legistyr test set: Investigating off-the-shelf instruction-tuned llms for terminology-constrained translation in a low-resource language variety. In *Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025*, pages 1–15, Naples, Italy. Unior Press.

Janika Oras, Mari Sarv, and Liina Saarlo. 2003. ERAB. Eesti regilaulude andmebaas. `https://www.folklore.ee/regilaul`.

Karl Pajusalu, Tiit Hennoste, Peeter Päll, and Jüri Viikberg. 2018. *Eesti murded ja kohanimed*.

Niko Partanen. 2024. Using large language models to transliterate endangered uralic languages. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 81–88, Helsinki, Finland. Association for Computational Linguistics.

Ronny Paul and 1 others. 2024. Towards a more inclusive ai: Progress and perspectives in large language model training for the sámi language. *arXiv preprint*.

Flammie A. Pirinen. 2024. 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. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 123–131, Helsinki, Finland. Association for Computational Linguistics.

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2025. Llms for extremely low-resource finno-ugric languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.

Frederick Riemenschneider. 2025. Beyond base predictors: Using LLMs to resolve ambiguities in Akkadian lemmatization. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 226–231, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Jorge Forero Rodriguez and Gilberto Bernardes. 2025. Leveraging large-language models for thematic analysis of children's folk lyrics: A comparative study of iberian traditions. In *Proceedings of the 12th International Conference on Digital Libraries for Musicology (DLfM '25)*, pages 53–59, New York, NY, USA. Association for Computing Machinery.

Kristiina Ross. 2015. Regivärsist kirikulauluni: Kuidas ja milleks kõrvutada vanu allkeeli. *Keel ja Kirjandus*, (7):457–470.

Jukka Saarinen. 2018. *Runolaulun poetiikka: Säe, syntaksi ja parallelismi Arhippa Perttusen runoissa*. Helsingin yliopisto, Helsinki.

Jukka Saarinen and Arvo Krikmann. 2004. SKVR database. `https://skvr.fi/`.

Mari Sarv. 2015. Regional variation in folkloric meter: the case of estonian runosong. *RMN Newsletter*, 9:6–17.

Mari Sarv. 2019. Poetic metre as a function of language: linguistic grounds for metrical variation in estonian runosongs. *Studia Metrica et Poetica*, 6(2).

Mari Sarv. 2020. Regilaulude teema-analüüs: võimalusi ja väljakutseid. *Methis Studia humaniora Estonica*, 26:137–160.

Mari Sarv and Risto Järv. 2023. Layers of folkloric variation: Computational explorations of poetic and narrative text corpora. *Folklore: Electronic Journal of Folklore*, 90:233–266.

Mari Sarv, Kati Kallio, and Maciej M. Janicki. 2024. Arvutuslikke vaateid läänemeresoome regilaulude varieeruvusele: "harja otsimine" ja "mõõk merest". *Keel ja Kirjandus*, 67(3):238–259.

Mari Sarv, Kati Kallio, Maciej M. Janicki, and Eetu Mäkelä. 2021. Metric variation in the finnic runosong tradition: A rough computational analysis of the multilingual corpus. In Petr Plecháč, Robert Kolár, Anne-Sophie Bories, and Jakub Říha, editors, *Tackling the Toolkit. Plotting Poetry through Computational Literary Studies*, pages 131–150. Institute of Czech Literature CAS, Prague.

Levente Seláf, Villő Vigyikán, Petr Plecháč, and Margit Kiss. 2025. Epic formulas and intertextuality in 16th century hungarian historical or epic songs. In *Plotting Poetry 5 - Popular Voices*, Tartu. ELM Scholarly Press. Forthcoming.

Peng Shu and 1 others. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *arXiv preprint*.

Timothy R. Tangherlini and Ruofei Chen. 2024. Travels with BERT: Surfacing the intertextuality in hans christian andersen's travel writing and fairy tales through the network lens of large language model-based topic modeling. *Orbis Litterarum*, 79(6):519–562.

Ayuto Tsutsumi and Yuu Jinnai. 2025. Do large language models know folktales? a case study of yokai in japanese folktales. *arXiv preprint*.

Cemile Uzun. 2025. A test of meaning, form, and culture in kurmanji: An evaluation of large language models' performance. *Open Research Europe*, 5:313.

Kaarel Veskis. 2025. Deminutiivsufiksi -kene varieerumine eesti regilaulutekstides. *Keel ja Kirjandus*, 68(6):510–539.

Chahan Vidal-Gorène, Florian Cafiero, and Bastien Kindt. 2025. Under-resourced studies of under-resourced languages: Lemmatization and pos-tagging with llm annotators for historical armenian, georgian, greek and syriac. `https://hal.science/hal-05119485/`. HAL preprint <hal-05119485>.

Linda Wiechetek, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

J. Xu, H. Zhang, H. Zhang, J. Lu, and G. Xiao. 2024. Chattf: A knowledge graph-enhanced intelligent q&a system for mitigating factuality hallucinations in traditional folklore. *IEEE Access*, 12:162638–162650.

A Annotation Guidelines

For all targets:

- you may use any additional sources, e.g. dictionaries, but no LLM outputs
- choose just one most probable option, if you have several (*viholaini*, *viholainen*)
- for translations, an option in the separate column (Alternative English) is possible

1. word_normalised: orthographically normalized (for Finnish corpora only)

- only needed for the Finnish part of the corpus (SKVR & JR; we use manually harmonised versions from ERAB)
- present the word in contemporary Finnish spelling. Retain dialectal and language specific features.
  - may correct short vowels into long ones when sure there are no short ones in that position in the local dialect
  - can use z, ž, š, tš, ttš and voiced consonants (b, d, etc.)
  - not to use half voiced consonants B, D, etc. for Ingrian (use d, b, etc.)
  - write macrons ¯ with the single vowels as long vowels
  - in the Finnish corpus, use y rather than ü also for Votic. [In Estonian dictionary and use it is ü, but in most of the Finnish material y. Also easier to compare with Ingrian and Ingrian Finnish if y.]
  - write numbers as words. Do not take line numbers (5, 10, 15, 20...) at the beginning of every fifth verse into account.
- old literary Finnish: normalise along the contemporary standard language while trying to retain potential dialectal features (which is difficult)
- correct evident mistakes by the collector (misspelings, misunderstandings) and OCR errors, and complement abbreviations (although this can be difficult for the models to do)
- do not try to reconstruct word forms in the original Karelian, Ludic, Ingrian or Votic language even if Finnisized by the recorder
- In Estonian, also correct the eventual typos and flaws of normalization (for example pähmämõtsa > pähnämõtsa; tädikeze > tädikese)

Examples: ruskei > ruskei, šuarella > šuarella, gostjat > gostjat, külüpaganah > kylypaganah, tsītämmä > tšiitämmä, bohwen > polveen, hīrikarvaлла > hiirikarvalla, lentolaisen > lentolaisen, neiokõnõ > neiokõnõ

2. word_lemmatised (local): text based dialectal lemma

- derive the basic form of the word (without inflections and declinations, but retaining derivatifs) as much from the basis of normalised text version as possible.
- With some words, especially with verbs, the basic form cannot always be inferred from the word form in the text. In this case, use a standard dictionary form in Eastern or Western dialect of Finnish, Northern of Southern dialect of Estonian, Viena, South or Livvi dialect of Karelian, or Votic or Ingrian (Izhorian).
- keep derivational suffixes, diminutives etc.

- for Estonian deminutives -kene and -ke we use shorter -ke form in lemma forms
- rough, fussy, uncertain interpretation
- relates both to local/individual language forms and varying recording practices & skills
- gives possibility to look at the linguistic/poetic variation at the most heterogenous level
- do not try to reconstruct word forms in the original Karelian, Ingrian or Votic language even if Finnisized by the recorder
- in Estonian, preserve separate keywords for minema / lähen (as in ETY and EMS), hea / parem
- South-Estonian negative particles -s, -i at the end of the word are treated as grammar and not represented in the lemma.

Examples: *ruskei* > *ruskei, šuarella* > *šuari, gostjat* > *gostja, külüpaganah* > *kylypagana, tsītämmä* > *tšiittää, bohwen* > *polvi, hīrikarvaлла* > *hiirikarvalla, lentolaisen* > *lentolainen, neiokõnõ* > *neiokõ; Väinämöini, Kadri, Katerina, Maaria*

3. lemma_standard: main form (root + derivative) in Estonian or Finnish, or in minority language if no corresponding form

- morphological similarity regardless of semantics: give the standard basic form corresponding to the word in Estonian or Finnish (the meaning may be different).
- if standard Estonian or Finnish form seems to be impossible or nonexistent, give the basic form in standard Ingrian, Karelian, Ludic, South Estonian or Votic, or the dialectal basic form, or just the text based dialectal main form derived from text itself.
- please keep derivative word forms and diminutives!
  - in Estonian we use standard-like orthography, if possible, based on local dictionaries, e.g. https://synaq.org/ (but not with võro q-orthography) or keywords from https://arhiiv.eki.ee/dict/vms/ or https://arhiiv.eki.ee/dict/ems/
- for Estonian deminutives -kene and -ke we use shorter -ke form in lemma forms
- names as such
- long personal pronouns in Estonian (as in EKSS, EMS)
- in Estonian preserve separate *minema* / *lähen* (as in ETY and EMS), *hea* / *parem*
- South-Estonian negative particles -s, -i at the end of the word are treated as grammar and not represented in the lemma.

Examples: *ruskei* > *ruskea, šuarella* > *saari, gostjat* > *gostja, külüpaganah* > *kylypakana, tsītämmä* > *kiittää, bohwen* > *polvi, hīrikarvaлла* > *hiirikarva, lentolaisen* > *lentolainen, neiokõnõ* > *neiuke, Väinämöinen, Kadri, Katerina, Maaria*

4. root: probable root form in standard language

- give the form of the word that is closest to etymological root form, but give this in standard Finnish or Estonian. This is the form that is given in Finnish and Estonian online etymological dictionaries. The root form refers to the element or word the other words have then been developing of. The actual etymological root form can be a very small linguistic element potentially existing in some earlier phase of the languages or proto language, but we are not going this far.

- if standard F/E seems impossible or nonexistent, use the main form in Ingrian, Karelian, Ludic, Veps, South Estonian or Votic
- take the probably earliest, most simple verb or noun, use the dictionary form
- for compounds, take two roots, separated with &
- ‘the main word in contemporary language corresponding the probable root at some earlier stage of linguistic history’
- in Estonian the deepest form that https://arhiiv.eki.ee/dict/ety/ gives, if possible; for Finnish https://kaino.kotus.fi/suomenetymologinensanakirja/
- long personal pronouns in Estonian (as in ETY)
- in Estonian preserve separate *minema* / *lähen* (as in ETY and EMS), *hea* / *paras*
- South-Estonian negative particles -s, -i, and North Estonian -p at the end of the word represented as a separate root with “& ei”
- names: provide the root in Finnic also for names with some other origin – there may also be several different Finnic roots (Maaria, Maria; Iilia, Jaani)

ruskei > *ruskea, šuarella* > *saari, gostjat* > *gostja, külüpaganah* > *kyly & pakana, tsītämmä* > *kiittää, bohwen* > *polvi, hīrikarvaлла* > *hiiri & karva, lentolaisen* > *lentää,*
neiokõnõ > *neid, väinä, Maaria, Maria, Iro, Irina*

5. English: Translation in English

- translate relating to the meaning that the word takes in the poetic line (no translations of lemmas only)
- translate word in the base form, no inflections etc. (for example *kulla, tsirgu* ‘dear’, not ‘gold’, ‘bird’, in *kulla ema, tsirgu ema*)
- not to translate diminutives
- you can have multiple word “denifitions” as counterparts if needed
- no alternative translations (use a separate column Alternative English for this)
- Does not have to be the most precise match (e.g. *ruuna* can be ‘horse’ instead of the precise ‘gelding’)
- to translate metaphors literally
- clearly onomatopoetic, meaningless untranslatable words to be presented as is.
- Names translated if there is known English counterpart, if not, then as is. cf. Riia > Riga; Ulivere > Ulivere

neiokõnõ > *~~little~~ maiden*

6. Refrain / untranslatable

- mark refrains

- mark also those onomatopoetic or meaningless (in counting rhymes, often of foreign origin) words, intensifiers, interjections and particles that are difficult to translate
- words in other languages
- you can use the column to mark also proper names (N) for future discussions

B An example prompt

A prompt generated from a pipeline shown in Listing 1 (Estonian version).

```
----- 1. SYSTEM -----

You are an expert in Finnic runosong tradition and historical linguistics with deep
knowledge of dialectal variations across Finnish, Estonian, Karelian, Votic, Ingrian
 (Izhorian, "inkeroisen "kieli, "isuri "keel), Veps, and other Finnic languages.
Your task is to understand the text as a whole considering separately each word and
its components according to information on and the procedure specified below.

----- 2. USER -----

This is a text from the Estonian corpus of runosongs. The corpus includes texts in
local variants of Northern and Southern Estonian dialects, mostly in specific poetic
 archaic runosong idiom. It also includes some texts other than runosongs.


You know that the texts often tell about peasant life, works in agriculture, hunting
, fishing, serfdom and working in manors, family members, clothing details, tools,
food, animals, family rituals, calendar rituals, mythological knowledge and ideas,
magical healing.

Prioritize concrete over abstract interpretations: Runosongs typically employ
concrete imagery - favor interpretations involving tangible objects, body parts,
natural phenomena, kinship terms, and material culture over abstract philosophical
concepts.

WRITING CONVENTIONS
Consider that numerals are written out in words or numbers.

Consider that many single words may be compound constructions.


TO CONSIDER FOR INTERPRETATION
You know that parallel lines are meant to repeat or extend the content of the main
verse, not contradict it.

Consider that songs can contain refrain words at the end or in the middle of each
line, or only of first lines, or refrains can be longer and span over several lines.
 Refrains can contain meaningless words or words with hazy meaning, and they should
not affect the interpretations of the poem text proper.

Consider that word order and syntactic structure in poetic text may be different
than in common language.



SPECIFICS OF RUNOSONG LANGUAGE
You know that runosongs are in archaic poetic language which varies across the
dialects with the main distinction between Northern and Southern Estonian. Dialect
features are pronounced less prominently than in spoken dialect language, usage of
archaic vs more modern dialect forms varies regionally.


COMPOSITION OF WORD OF ROOTS, CLITICS, PARTICLES, ENDINGS
When analysing component parts of the word:

- consider emphatic particles (-gi, -ki), question markers, South-Estonian
confirming particle -ks, and other enclitics that may be fused with word forms and
affect meaning interpretation.

- consider that South Estonian negation particles -i (present time) and -s (past
time) are merged at the end of the words, sometimes without any visible break,
sometimes with hyphen. in South Estonian texts, you MUST check for every verb if it
ends with negation particle (1) vowel + i - present time; (2) vowel + s - past time
(not to confuse with South-Estonian confirming particle -ks).

MORPHOLOGICAL ENDINGS
```

```
Consider that runosongs in Estonian have:
* different root forms for nominative and genitive case (for example kägu:käo).
and considering that specific runosong register has:
* archaic case paradigms with longer endings than modern Estonian (where various
sound losses have taken place)
* often vowel at the end of nomen cases or in the middle of word that has been lost
in the later standard language (for example, archaic ""minuda, contemporary ""mind)
* longer morphological endings than standard language and various clitics (for
example, -je ending in illative, -da ending in partitive, -maie ending in infinitive
, sometimes reminiscences of possesive suffixes)
* often diminutives with -kene or -kõnõ, -ke or -kõ or other variants.
* translative case ending may be -ks, -ksi, -ksa, -s, -ssa, -ssi, or -st, -sta
depending on dialect (not to confuse with very common South-Estonian confirming
particle -ks)

PHONOLOGICAL VARIATION
In interpreting the word forms, account for historical phonological changes and
variation. Consider:
* vowel losses in unstressed syllables in modern standard and dialect forms, and
varying preservation of respective vowels in runosong idiom
* systematic vowel changes and variants in dialects (intermittent o~õ~e, for example
 medu~mõdu ‘’mead, vowel shifts, diphthongisation or heightening of long vowels, for
 example pea~pia~peä~pää ‘’head)
* systematic consonant changes and variants (strengthening or weakening or loss of k
, p, t, g, b, d, j between or next to vowels)
* different consonant gradation patterns
* sound changes that may obscure root identification
* vowel harmony in some dialects
* frequent word-initial h-omission before vowel in runosongs
* occasional word-initial v-omission before o, ö, u, ü.


When interpreting the text, perform the systematic check of following options for
words with unclear meaning
1. MANDATORY root first vowel replacement check:
(1) õ instead of e, o, ö or other way round; (2) ä instead of e. Do not check the
vowels further in the word.
a) First transcribe/analyze as written
b) Then test the alternative variant (koht → test kõht, kõhe → test kohe)
c) Compare both meanings against context
d) Choose the variant that makes better semantic/contextual sense
2. MANDATORY v-omission check for roots beginning with o, u, ö, ü in the text:
a) First transcribe/analyze as written
b) Then test the v-initial variant (öö → test vöö), and also consider vowel
replacements with õ
c) Compare both meanings against context
d) Choose the variant that makes better semantic/contextual sense
3. MANDATORY h-omission check for EVERY root beginning with vowel (a, e, i, o, u, õ,
 ä, ö, ü) or h in the text:
a) First transcribe/analyze as written
b) Then ALWAYS test the h-initial variant (õbe → test hõbe, allitama → test
hallitama)
d) Compare both meanings against context
e) Choose the variant that makes better semantic/contextual sense
Do not check the words beginning with consonants other than h.


Create a single word-by-word analysis table with this format:
| original form | comment | English translation | normalized orthography | lemma (
original) | lemma (modern) | etymological root |
|--------------|---------|---------------------|------------------------|------------------|------

| [Word as in text] | [Translation notes] | [English equivalent] | [Modern spelling]
 | [Basic form in original] | [Modern language lemma] | [Etymological root(s)] |


Analyze the provided Finnic runosong text and translation to create a comprehensive
word-by-word analysis table.

# ANALYSIS GUIDELINES
```

```
For each word in the original text:
- The "original form" column should use the exact word from the original language
text
- Add helpful comments about interpretation challenges or linguistic features in the
 "comment" column

## REFRAINS
- Detect if the song contains refrain words at the end or in the middle of each line
: do not analyse these words, mark these as [refrain].
- Detect if the song contains verse-length refrains: do not analyse these words,
mark these as [refrain].

- In "English translation," provide the best English equivalent for this specific
word based on the translation - main word form that can serve as a keyword entry in
English dictionary (nominative singular, present tense infinitive with to, no
prefixes nor modalities), give only translation of the main word form, do NOT add
information what is given with morphological endings
- For "lemma (original)," provide the basic form in the dialect of the text (for
nouns: nominative singular, for verbs: present tense infinitive, use forms with -ma
or -me or -mä ending & NOT -da/-ta ending) with original phonological peculiarities,
 with derivational suffixes, without morphological endings
- For "lemma (modern)," give the the equivalent in standard phonology, that can
serve as a keyword entry to standard Estonian dictionary with derivational suffixes,
 without morphological endings, do NOT separate compounds into parts, do NOT change
etymological stems.
- For "etymological stem," include the stem word(s), i.e. main word forms (in case
of verbs, ma-infinitive NOT da-infinitive) that would serve as dictionary entries in
 ETY, from the eki.ee site), remove derivational suffixes, separate multiple
possibilities with commas and compound components with &

# IMPORTANT NOTES

1. Focus ONLY on creating the required tables. Do not include any other tables,
dialect identification, or additional commentary.
2. Include all words from the original text in your analysis, even particles and
function words.
3. Be consistent in your analysis throughout the table.
4. Use the provided translation to inform your analysis, but focus on an accurate
word-by-word breakdown.
5. For words that are difficult to analyze, provide your best linguistic assessment
based on context and comparative Finnic linguistics.

Begin your analysis by examining the original text and its translation, then create
the word-by-word analysis table as specified above.


Original text:
midli madli kudusin kangast
tegin pojale püksid
poeg läks metsa kännu otsa
tõmbas lõhki tsirr


 I expect the first column in the table to contain all these words, in this order: [
 'midli', 'madli', 'kudusin', 'kangast', 'tegin', 'pojale', 'püksid', 'poeg', 'läks'
 , 'metsa', 'kännu', 'otsa', 'tõmbas', 'lõhki', 'tsirr'] All columns should be
 filled for every word.
```

C An example model output

original	comment	English	normalized	lemma (orig)	lemma (modern)	etymological root
ku	South Estonian conjunction	when	kui	ku	kui	kui
olli	South Estonian past tense form of ‘olema‘	was	oli	ollõma	olema	olema
nuuri	Partitive form of ‘nuur‘	young	noori	nuur	noor	noor
neiokõne	Diminutive form of ‘neiu‘	maiden	neiukene	neiokõnõ	neiuke	neiu
kui	Conjunction	when	kui	kui	kui	kui
ma	1st person singular pronoun	I	ma	ma	mina	mina
kasvi	South Estonian past tense form of ‘kasvama‘	grew	kasvasin	kasvma	kasvama	kasvama
kabokõne	Diminutive form of ‘kabo‘ (maiden)	young girl	kabukene	kabokõnõ	kabuke	kabo
lätsi	South Estonian past tense form of ‘minema‘	went	läksin	minema	minema	minema
marja	Illative case of ‘mari‘	berry-picking	marjale	mari	mari	mari
sis	South Estonian form of ‘siis‘	then	siis	sis	siis	siis
mäe	Genitive of ‘mägi‘	hill’s	mäe	mägi	mägi	mägi
päälõ	South Estonian allative case ending	onto	peale	pääle	peale	pea
lätsi	South Estonian past tense of ‘minema‘	went	läksin	minema	minema	minema
orgo	Illative of ‘org‘	to valley	orgu	org	org	org
ubinahe	Illative of ‘ubin‘ (apple orchard)	to apple orchard	õunaaeda	ubina	õunaaed	õun & aed
panni	1st person past tense of ‘panema‘	put	panin	panema	panema	panema
ma	1st person singular pronoun	I	ma	ma	mina	mina
tuppõ	Illative of ‘tupp‘	into sheath	tuppa	tupp	tupp	tupp
tuima	Genitive of ‘tuim‘	cold	tuima	tuim	tuim	tuim
ravva	Genitive of ‘raud‘	steel	raua	raud	raud	raud
vaivaväidse	Genitive of ‘vaivaväits‘ (poor knife)	poor knife	vaevaveitse	vaivaväits	vaevaveits	vaev & väits
panni	1st person past of ‘panema‘	put	panin	panema	panema	panema
vüü	Genitive of ‘vüü‘	belt’s	vöö	vüü	vöö	vöö
ala	Postposition	under	alla	ala	all	all

Table 4: A few initial lines of a word-by-word analysis table produced by the Claude model

Table 4 shows a word-by-word analysis produced by Claude. In this small excerpt, only two words are clearly misinterpreted: *ubinahe* ’to apples’ is incorrectly interpreted as a compound, and in a compound *vaivaväidse* ’sharp knife’ the first part is misinterpreted as the common standard-language word *vaev* ’hardness’ instead of the correct South-Estonian *vaib* ’sharp’. In addition, *tuim* ’feelingless’ is not exactly ’cold’ but is semantically close to the original meaning. The normalization results clearly represent the standard language (the task appears to be misunderstood by the model). The original lemma—which does not concern standardized language—can have several equally plausible interpretations, making it a challenge for both humans and the model to choose a single correct form. The standard lemma results are mostly correct, while the etymological root shows deviations in *neid* vs *neiu* (stem variants) and in the misinterpretations mentioned above. For the exceptional verb ’to go’, which has two stems, Claude has decided to give the stem of the main form (*minema*), while manual annotators chose to retain the original root (*lähen*).

Fine-Tuning Whisper for Kildin Sami

Enzo Gamboni
University of Eastern Finland
egamboni@uef.fi

Abstract

For this study, Whisper, an automatic speech recognition software, was fine-tuned on Kildin Sami, an endangered and low-resource Uralic language, using an automatic speech recognition-tailored dataset of less than 30 minutes. Three different Whisper models were trained with this dataset—each one with a different base language (English, Finnish, or Russian)—to examine which model provided the best result. Results were measured using Word Error Rate; fine-tuning the Russian-base Whisper model resulted in the lowest Word Error Rate at 68.55%. While still high, this result is impressive for only a small amount of language-specific training data, and the training process yielded insights relevant for potential for further work.

1 Introduction

This paper summarizes the results of a study carried out between 2024–2025 and submitted as an MA thesis (Gamboni, 2025).

1.1 Background

Endangered languages are those languages which are at risk of losing their speaker base, largely due to language shift (Grenoble, 2011). Relatedly, low-resource languages are those which lack significant data for natural language processing (NLP) (Joshi et al., 2020; Magueresse et al., 2020). Kildin Sami, a language of the Eastern Sami group and more broadly belonging to the Uralic language family (Sammallahti, 1998), is both low-resource and endangered, with estimates that only 20 active speakers still remain (Scheller, 2024). This makes Kildin's status a precarious one, in which the procurement of the large quantities of data traditionally needed for NLP is not feasible, yet all the more important.

Including low-resource, endangered languages in NLP is not only beneficial for NLP because it provides a more expansive data pool to boost accuracy, but also beneficial to endangered languages because it 1) bolsters their digital presence, contributing to the groundwork that will help to safeguard them in an increasingly digital and global society, and 2) aids researchers in more streamlined, less time-consuming workloads, as previously manual tagging and transcription could be partially or fully automated with computational methods (Trosterud, 2006; Poibeau and Fagard, 2016; Partanen et al., 2021).

This study trained Whisper, an automatic speech recognition (ASR) model on Kildin Sami data to see if significant, useful results could be achieved with ultra-minimal training data adapted from fieldwork data. A secondary goal of this study was thus to prove that Kildin fieldwork recordings can be useful in NLP research, in line with Himmelmann (1998) assumptions that an analytic approach to documentary linguistics results in data relevant to a broad subset of linguistic fields.

1.2 Current Digital Resources for Kildin Sami

An online Kildin-Russian dictionary (Antonova and Scheller, 2021–) is available and linked to an automaton for paradigm generation,[1] as well as a keyboard layout[2] for the standardized Cyrillic orthography developed in the 1970s and 1980s by a group lead by Rimma Kuruch and including Alexandra Antonova, who is among the authors of the aforementioned

[1] See the dictionary's imprint https://sanj.oahpa.no/about/. It is unclear whether or where this tool is available elsewhere.

[2] See https://giellatekno.uit.no/cgi/index.sjd.eng.html.

dictionary (Rießler, 2020).

2 Methodology

This section describes the primary data used for this study; how it was adapted for ASR training; and subsequently details the ASR training process.

2.1 Data

Data used for this project comes from the Kildin Sami corpus, a private repository under the langdoc Github repository.[3] The corpus contains textual annotation data in XML (Rießler, 2024, 42), including time alignment to field recordings. Requests for access should be addressed to the repository's administrators. The fieldwork recordings used in this project come from the Kola Sami Documentation Project (KSDP) (Rießler, 2005–2025). These field recordings are housed in The Language Archive, for which access may be requested by contacting the administrators.

This primary data amounts to 38 minutes and 4 seconds of audio files and is comprised of three KSDP video recordings and one 39 track audiobook. All of the audio comes from one speaker, Sami language activist Nina Afanasyeva. The audiobook, *Miŋgá* (Vinogradova, 2007), is a collection of short poems written by Russian and Sami poet Iraida Vinogradova and the Kildin Sami speech is 25:14 in length. All three video recordings are largely monologues, with decent audio and infrequent background noise. Similarly, Nina Afanasyeva's audiobook narration is high in sound quality, though several tracks contain background music that at times covers her speech.

2.2 Dataset Creation and Preprocessing

First, using the tool ELAN,[4] a new textual annotation tier was created within the Kildin Sami corpus' time-aligned XML files for each audio file. These tiers were created by copying the preexisting orthographic text into the new tier and modifying it for ASR training. This process involved the following changes: removing all punctuation; removing all capitalization except for proper nouns; standardizing the transcription for false starts, nonverbal utterances, and affixes/clitics; and simplifying the Kildin standard orthography by removing macron diacritics.[5] Vinogradova's audiobook orthography was updated to reflect that which is used in Rießler's fieldwork. Notably, replacing instances of ‹’› (Unicode: 02BC) with the Cyrillic letter SHHA ‹**h/h**› (Unicode: 04BA / O4BB) (Rießler, 2013).

Next, using Audacity,[6] the audio files were manually broken into multiple .wav files, with each file corresponding to a chunk of annotation in the ASR annotation tier. Two .csv metadata files were then created—one for training the ASR model and one for evaluating the model's output—to link each shortened audio file to its transcription. 80% of the data was devoted to training the ASR model, while 20% was reserved for evaluation. ~10% of the evaluation data was taken from the audiobook recordings with the other ~10% taken from the fieldwork recordings to ensure that the evaluation results best represented the data. In sum, the resulting dataset consisted of 717 .wav files and totaled 27 minutes and 29 seconds, meaning that ~10 minutes of the primary recordings were either too poor quality to use or did not feature Nina Afanasyeva speaking.

2.3 Fine-Tuning Whisper

Three Whisper models were fine-tuned; one with English as the selected base language; one with Finnish selected; and one with Russian selected. This was done to see if different base language settings would affect end-performance. Finnish was selected because it is the closest language, linguistically, to Kildin Sami that the base Whisper model had been trained on. Russian was selected intuitively due to Kildin's use of the Cyrillic orthography. Finally, English was included, to see whether Hjortnæs et al. (2021) discovery that quantity outperformed linguistic similarity of the source language in their Komi study would also be relevant for working with Kildin.

Whisper was pretrained on 1,066 hours of Finnish data; 9,761 hours of Russian data; and

[3] https://github.com/langdoc/

[4] https://archive.mpi.nl/tla/elan

[5] In Kildin, macrons over vowels denote long vowels. However, their use across researchers is unsystematic, and vowel-length opposition in Kildin is marginal (Rießler, 2013).

[6] https://www.audacityteam.org/

Whisper Output	Manual Transcription, Modified Orthography	Manual Transcription, Standard Orthography	English Translation
на мэнн мэн уййнэ	на мэнн мунн уййнэ	на , мэ̄нн мунн уййнэ ?	Well, what did I see?
вуэнн уйнэ	мунн уйннэ	мунн уйннэ	I saw
тэдт инца айк аллт	тэдт инцэ айкалт	тэдт ӣнцэ а̄йкалт	this morning, early.
элляһав сулль пейв пейв вэннӭ луннӭ	элля вял шурр пеййв пеййвэнҍ лоӈӈнэ	элля вя̄л шӯрр пе̄ййв пе̄ййвэнҍ ло̄ӈӈнэ	Not yet a full day, the sun is rising.
вуаннӭсьт ляһав	ванас ли вял	ва̄нас лӣ вя̄л	It's still a little
севьнӭсьт	севвьнэсьт	се̄ввьнэсьт	dark (twilight),

Table 1: A comparison of the trained Whisper model results with the manually transcribed text.

438,218 hours of English data (Radford et al., 2023). The fine-tuning was done using HuggingFace transformers and code[7] and was executed in Google Colab.[8] A ColabPro subscription provided Nvidia GPU access. Whisper's small-sized model was used for each and trained on 500 steps. When attempting to train the model using more than 500 steps, the execution time increased dramatically and became impractical to run with the limited computational resources and time available. Each model was evaluated for Word Error Rate (WER) by using the evaluation split from the data set during the fine-tuning process. This WER calculation was done automatically at the end of the training process.

3 Results

Of the resulting models, the one set to Russian performed best, achieving a 68.55% WER. The model set to Finnish resulted in a 71.38% WER while the one set to English did the worst with a 73.88% WER. This is notable, as it suggests that orthographic similarity may have played a greater role in the improvement of WER than linguistic similarity or the quantity of pretraining data.

3.1 Transcription Analysis

The fine-tuned, Russian-based model was used to transcribe 30 seconds of audio from the test split. It took ~6 minutes to transcribe the 30 second audio clip, a portion of which is shown in Table 1 together with the manual transcriptions in both the standard and modified orthographies. An English translation is provided. Words that Whisper transcribed incorrectly are underlined in the Whisper Output tier.

The model struggles to discern between single and double consonants; vowel quality; and occasionally word boundaries. In instances where the model output is completely dissimilar to the expected output, it may be pertinent to review that specific audio section to see if there is background noise interfering with speech clarity. Further analysis of this model using character error rate (CER) analysis would offer greater insight into the nature of these errors.

3.2 Comparison to Prior Studies

Table 2 shows how the Kildin model performed in relation to models from prior studies trained on comparable amounts of data (with the exception of North Sami and Zyrian Komi, included to show work done on other Uralic languages). These results show that fine-tuning Whisper on Kildin produced comparable results to other models also fine-tuned on ≤30 minutes of data, whether trained on a Whisper model or Wav2Vec2. The lowest WER acheived with ≤30min. data was from Meelen et al. (2024) training Dzardzongke on Wav2Vec2.

Comparing these results suggests that while further experimentation may lead to WER im-

[7] https://huggingface.co/learn/audio-course/en/chapter5/fine-tuning

[8] https://colab.research.google.com/

Language	Available Data	ASR System	WER	Study
North Sami	88 unlabelled hours 20 labelled hours	Wav2vec2 + extended fine-tuning on Finnish	28.84%	Getman et al. (2024)
Dzardzongke	30 minutes	Wav2vec2	50%	Meelen et al. (2024)
Kildin Sami	27 minutes	Whisper Small	68.55%	Gamboni (2025)
Bribri	29 minutes	Whisper Medium	65-75%	Jimerson et al. (2023)
Guarani	19 minutes	Whisper Medium	65-75%	Jimerson et al. (2023)
Newar	30 minutes	Wav2vec2	74%	Meelen et al. (2024)
Zyrian Komi	35 hours	DeepSpeech + Komi/Russian LM	76.50%	Hjortnæs et al. (2021)

Table 2: Comparing Kildin's results to those of other studies surveyed during project. For referenced studies in which multiple languages were tested, only those with ≤30min. of data were included. If a study tested multiple ASR systems and Whisper was among them, Whisper's results were chosen to compare.

provements for Kildin, it is unlikely to improve to a WER <50% or to approach the success Getman et al. (2024) found with many hours of data for North Sami.

4 Conclusion and Future Potential

Whisper offers promising results when trained on ultra-minimal data for Kildin Sami and supports Himmelmann (1998) assumption that an analytic approach to documentary linguistics produces relevant data for a broad subset of linguistic fields. Although 68.55% WER is high, it is remarkable to be achieved with a data set of less than 30 minutes combined for training and testing and shows how advancements in NLP are making the inclusion of endangered and low-resource languages more feasible. Despite the author's lack of computational background, significant results were still achieved and could well become useful for semi-automating Kildin transcriptions with further experimentation. The author hopes that this study can serve as a starting point for further experimentation on training Whisper on Kildin Sami and as proof that those with a limited computational background can still incorporate computational methods into their linguistic research.

This study was influenced by Hjortnæs et al. (2021) observations that source language quantity was more impactful than linguistic similarity for Komi, but finds that the same did not hold true for Kildin; rather, it seems that shared orthography played a greater role. Future work focusing on how to simultaneously leverage the orthographic similarity of Russian and the linguistic similarity of Finnish to Kildin, would be beneficial to consider for improving WER and further testing this assumption. A reexamination of the ASR dataset created for this project would also be worthwhile to see if decisions made during preprocessing significantly impacted the ASR training. This reexamination should be done after more in-depth analysis of the current model's output is undertaken to discern if there are commonly repeated errors that could be stemming from human error or decision-making within the dataset. Lastly, experimentation with training the model on a greater number of steps or on a larger Whisper model may also yield greater WER and contribute to the robustness of this study.

5 Limitations

Time was a limiting factor on this study's depth. Minimal speech data available for training the Kildin Sami model was another inherent limitation.

Limitations concerning the definition of linguistic vs. orthographic similarity mentioned during this study must also be addressed. Though I posited that the Russian-based model performed best due to orthographic similarity, an anonymous reviewer pointed out that linguistic similarity may still be the reason for this, as Kildin and Russian share features like palatalisation, while Finnish does not. Relatedly, transcription of long vowels, something the Russian-based model struggled with considerably, could be attributed to the absence of length distinction in Russian, further highlighting the role of base language similarity. Thus, speculation and claims within this study on the role of linguistic vs. orthographic similarity are limited due to a lack of in-depth analysis on the subject. As this work is ongoing, this topic will be further explored. My gratitude is extended to the reviewer who raised this concern.

6 Ethical Considerations

None of the materials used contain any sensitive or personal information, nor are any of them being freely distributed in their entirety for this project. Nina Afanasyeva has given her informal consent to have recordings of her from the Kola Sami Language Documentation Project used for the purpose of language technology development.[9]

Use of audio taken from Vinogradova (2007), which is under copyright, adheres to the copyright laws within the European Union.[10] However, because data taken from copyrighted material may not be made publicly available, the dataset used to train the ASR models is housed in a private repository.

References

Aleksandra A. Antonova and Elisabeth Scheller. 2021–. *Saamsko-russkij i Russko-saamskij slovar'*. UiT The Arctic University of Norway.

Enzo Gamboni. 2025. Fine-tuning Whisper for Kildin Sami, a low-resource endangered language. Master's thesis, University of Eastern Finland.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi. In *Interspeech 2024*, pages 2539–2543.

Lenore A. Grenoble. 2011. *Language ecology and endangerment*, page 27–44. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. 36:161–195.

Nils Hjortnæs, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2021. The relevance of the source language in transfer learning for ASR. In Miikka Silfverberg, editor, *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, volume 1, pages 63–69. University of Colorado Boulder.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Marieke Meelen, Alexander O'neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of Nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93, St. Julians, Malta. Association for Computational Linguistics.

Niko Partanen, Michael Rießler, and Joshua Wilbur. 2021. Envisioning digital methods for fieldwork in the Arctic. In Markku Lehtimäki, Arja Rosenholm, and Vlad Strukov, editors, *Visual representations of the Arctic*, Routledge Interdisciplinary Perspectives on Literature, pages 313–339. Routledge.

Thierry Poibeau and Benjamin Fagard. 2016. Exploring Natural Language Processing Methods for Finno-Ugric Langages. In *Proc. of the Second International Workshop on Computational Linguistics for Uralic Languages*, Proc. of the Second International Workshop on Computational Linguistics for Uralic Languages, Szeged, Hungary.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of*

[9] Michael Rießler (KSDP), p.c.

[10] The EU Directive 2019/790 on Copyright in the Digital Single Market outlines copyright exceptions that allow for text and data mining for scientific researcher purposes.

the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Michael Rießler. 2005–2025. Kola Saami Documentation Project (KSDP). In *The Language Archive (TLA)*. Max Planck Institute for Psycholinguistics.

Michael Rießler. 2013. *Towards a digital infrastructure for Kildin Saami*, pages 195–218. Exhibitions and Symposia. Kulturstiftung Sibirien.

Michael Rießler. 2020. Rimma Kuruch and Kildin Saami language planning. *Linguistica Uralica*, 56(3):220–225.

Michael Rießler. 2024. Kola Saami Christian Text Corpus. In Mika Hämäläinen, Flammie Pirinen, Melany Macias, and Mario Crespo Avila, editors, *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 138–144. ACL.

Pekka Sammallahti. 1998. *The Saami languages: an introduction*. Davvi girji, Kárášjohka.

Elisabeth Scheller. 2024. Activating passive Kildin saami language knowledge through the Master-Apprentice Language Learning Method and instruction in grammar and writing skills. 2/2024:82–108.

Trond Trosterud. 2006. *Grammatically based language technology for minority languages*, pages 293–316. De Gruyter Mouton.

Iraida V. Vinogradova. 2007. *Miŋgá = Mīnn'kaj*. Davvi Girji.

Digitization Work at the Finno-Ugrian Society: Livonian Case Study

Niko Partanen
Finno-Ugrian Society
niko.partanen@helsinki.fi

Jack Rueter
University of Helsinki
Digital Humanities
Language Technology
jack.rueter@helsinki.fi

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

Abstract

This article discusses the recent digitization project of the Finno-Ugrian Society, using the work on Livonian publications, especially those from Seppo Suhonen's *Liivin kielen näytteitä* from 1975 as a case study. We start by contextualization and motivation for these undertakings, both from the point of view of the Finno-Ugrian Society and the University of Latvia Livonian Institute, and then describe the workflows we have developed and foresee for the next steps.

1 Introduction

In last years the Finno-Ugrian Society has systematically advanced their digitization program, with the goal of increasing the digital availability of the materials the Society has published. This paper outlines how the work has progressed, what types of questions have been addressed and which have been identified to still require solutions. We use as an example the Livonian materials recorded and published by Seppo Suhonen, narrated primarily by Pētõr Damberg (Suhonen, 1975). Other Livonian materials the Society has published are Setälä (1953) and Mägiste (2006). The aspect that distinguishes Suhonen's materials from the rest is that recordings were made and have been archived at the Institute for the Languages of Finland. Setälä's and Mägiste's publications are based on transcriptions made on the spot without recordings. The audio recordings open many new possibilities in available workflows that need to be discussed.

2 Context of the University of Latvia Livonian Institute

Compared to many other critically endangered languages, Livonian has been relatively well documented. Nevertheless, much of this documentation has historically been shaped by the academic interests of linguists, resulting in materials that primarily address scholarly audiences. Examples include textual publications and, in particular, lexicographic works dating back to the mid-19th century (e.g., Wiedemann, 1861; Kettunen, 1938), which relied heavily on phonetic transcription and were therefore largely inaccessible to the Livonian-speaking community. The first lexicographic collection written in the Livonian standard orthography did not appear until 1999 (Ernštreits).

Since its establishment in 2018, the University of Latvia Livonian Institute has been developing a suite of dual-purpose databases—serving both research and community needs—which encompass lexicographic and morphological data as well as the Livonian text corpus, all based on the contemporary Livonian orthography (see Ernštreits et al. 2024).

With the rapid expansion of the aforementioned Livonian database cluster over the past five years and the growing interest in Livonian language learning and use, the need for additional documentation has become increasingly evident. While major text collections published or compiled in the standard orthography—such as books, newspapers, and manuscripts—have already been incorporated into the corpus or are planned for inclusion in the near future, the question of how to effectively integrate other sources, such as materials published for academic purposes in phonetic transcription has come to the forefront.

The efficient utilization and resource-conscious normalization of such phonetic sources into the standard Livonian orthography is closely tied to the research presented in this article. In the broader context of developing future technologies serving both the Livonian community and linguistic research, this work is especially timely. A related project currently being implemented at the University of Latvia Livonian Institute focuses on creating an aligned speech corpus, which uses as its speech input texts from the written corpus—particularly those reflecting natural speech situations, such as folklore.

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 112–122
December 10-12, 2025 ©2025 Association for Computational Linguistics

Consequently, the integration of transcriptions from existing audio recordings is highly relevant for the development of future speech technologies. These not only promise to expand opportunities for the use of Livonian but will also facilitate the extraction of additional linguistic data from the substantial number of recorded yet untranscribed Livonian audio materials.

3 Context of the Finno-Ugrian Society

The Finno-Ugrian Society has been publishing scientific materials, both research and language materials, on the Uralic languages since the Society was founded in 1883. The Society has also funded and coordinated large fieldwork material collections throughout the areas where the Uralic languages are spoken. These materials can be primarily found at the Archive of the Finno-Ugrian Society located at the National Archives of Finland. The Society, in a work that has continued to the present day, has been publishing these materials as edited text collections and dictionaries, and new research is continuously being published.

The contemporary demands and expectations toward the digital availability of these resources have led the Finno-Ugrian Society to develop and apply a digitization plan. Although the digitization work has for now primarily focused on published journals, recently work on text collections and dictionaries has also been initiated, and also the first digitization experiments have been conducted with the Society's archives.

Whereas the scholarly output is primarily meant for researchers, the situation is different with materials such as texts and dictionaries. These materials certainly have extensive and important research uses, but at the same time they are very important for contemporary language users and learners. We argue that it is necessary to combine to the digitization process steps which enhance the usability of these materials, and these actions ultimately align very closely with the needs of both the community members and researchers.

The goal of the Finno-Ugrian Society is not to republish these materials. These materials have already been processed, analyzed and edited by specialists of each language, and we would prefer to frame our current work more as enhancing the usability and accessibility of the already existing works, and not as creation of new publications as such. Of course these boundaries are blurry, and digital versions of the publications are inevitably distinct from the originals. There are situations where they need to be cited separately, and the researchers who were involved in the work with the digital versions also need to be acknowledged. Our stance can still be illustrated by delineations such that when we digitize these works and create digital versions, we refrain from additional tasks such as adding new translations. Tasks such as adding automatically a new normalized transcription layer or ensuring that all lexemes are in the morphological analyzers are more of enhancing background tasks than conducting entirely new research.

This work has not been done in a vacuum, but it connects to the earlier research. Rueter and Partanen (2019); Rueter (2024); Rueter et al. (2024) describe their work on Erzya and Moksha corpora and how they connect to the analyzers of these languages. However, the approach taken here is more extensive, and we aim to keep the connection intact between digitized resources and the later corpora constructed from them.

4 Livonian Case Study

The Livonian recordings carried out by Seppo Suhonen in 1971 in Tallinn and Riga form a large collection of Livonian speech data. A co-interviewer was Karl Kont. Suhonen returned to interview Pētõr Damberg in 1981, but these recordings will only be digitized by the late 2026. Besides Damberg, Suhonen recorded other individuals as well, and Damberg himself was recorded by Eduard Vääri and Unto Miettinen in 1965. These recordings are stored in the Tape Archive of the Finnish Language at the Institute for the Languages of Finland. Jantunen (2025, 9) estimates that Suhonen's recordings are all together approximately 51 hours.

The recordings transcribed and published in *Liivin kielen näytteitä* (Suhonen, 1975) contain in total 2 hours and 50 minutes of speech. The published transcription is displayed in Figure 1, located in the end for convenience. We argue that this is an extremely typical scenario with data on endangered languages: a small part of the material is processed in more detail than the rest. Also in this case Jantunen (2025, 9) describes having transcribed approximately 6 hours of the Suhonen's materials. This means that approximately 20 % of Suhonen's material has been transcribed.

This scenario is at the same time very promising and potentially highly rewarding in contemporary technical landscape. The parts of the dataset that are more finely processed, be it in the form of transcriptions or annotations, can be used as a training data to model the process in question, and thereby the resulting model can be used to analyze the remaining data in comparable style. This way the current transcribed Suhonen's Livonian corpus could ideally be extended to whole 51 hours, which would have significant consequences to the general availability of spoken and transcribed Livonian.

It must also be noted that we are now discussing Livonian materials collected by a few individuals and stored in one language archive: naturally, the scope of all existing Livonian recordings from this time period, and containing speech from the same individuals, among them Damberg, is much larger.

5 Automatic Text Recognition

Automatic text recognition of texts written in the Finno-Ugric transcription system has been a large challenge in the field in the past. However, in last years especially the Transkribus platform (Kahle et al., 2017) has allowed researchers to easily transcribe materials following the transcription conventions they consider best, and then train text recognition models with this data, improving the accuracy rapidly in an iterative manner. At the same time, processing of handwritten documents has also progressed very rapidly (Partanen et al., 2022; Arkhipov et al., 2021; Lamb et al., 2022).

In the context of the Finno-Ugrian Society's Livonian materials, we often find a situation where the same material exists in handwritten, typed and published versions. In these instances our focus is in digitizing the published version, and we take as our starting point that this is the most carefully edited and the most useful version. We can make the information available about the other existing versions, but starting to digitize all of them and creating comparable version would already stray away toward entirely new publications, and is not the point nor the scope of the current work. The goal is not to reconstruct in detail all nuances of the earlier work, but improve the use of language resources that are not currently as accessible as they could be. The language data is in the focus of this work, not the actions of the earlier researchers.

When we create the text recognition models, it seems that Finno-Ugric transcription of Livonian is a category in which the same models are able to generalize up to some degree. However, each publication has small differences and idiosyncrasies that need to be individually addressed. The best Livonian models currently are trained with almost 200 000 transcribed words and reach the character error rate of 0,28 %. A page in Transkribus platform with recognized and manually corrected text is displayed in the Figure 2. When we want to process a new publication, we need to add enough pages to cover the new characters and the new variation, but in our experience this is a very painless and fast process.

Although Transkribus is not an open-source platform, the spirit and general approach of Transkribus maintainers and the READ Coop that manages the project has aligned well with our goals. Needless to say, one must also consider whether the proofread materials could be deposited in some other environment, so that even open-source text recognition tools could be trained, tested and evaluated with this data.

6 Layout Analysis and Tagging

This part of the process takes place partly before the text recognition, but we discuss it still at this point as adjustments to the layout are done usually after the text recognition, and at the same time part of the tagging is done for the existing text.

Layout analysis refers to the identification of the structures in the document pages. Text regions and text lines are examples of regions, and page number would be an example of a structurally tagged text line. In our approach this tagging is extended very far. We mark headers, descriptions, metadata sections and page numbers separately. This data can be used effectively when the corpus is created at the later steps.

There would be many ways to structure the data, but our goal for now has been to create a minimal structure needed to distinguish Livonian and Finnish elements, first of all. In some of the books discussed here every even and odd page has a different language, in which case we can simply use this information to distinguish the language. At times the texts and translations are on the same page, with possible multiple short texts per page. In these cases it is critical that both original text and translation parts have the same number of elements. Then we can match the Livonian text and translation automatically by the number of elements. Naturally, a more nuanced method could be envisioned, but this convention has worked well for us. There have been individual cases where a Livonian paragraph is split into two paragraphs in the Finnish translation. In these cases the solution has been to insert a tag for the Finnish translation that tells that we have a non-corresponding paragraph break. Information about the existence of the paragraph break is thereby kept, and the digitized data remains as intact and coherent as possible.

One particular case comes from indentation. In some situations, indentation is distinct enough that we can build a small classifier for the line starting points on a page and identify which are indented. At times, instead of indentation, there is a small vertical space between paragraphs: then these should probably be in different text regions.

Hyphenation is another structural issue. When the hyphen is located at the end of the line, the word can possibly be just hyphenated in this position, or it can be a compound where the hyphen is supposed to occur following the used transcription standards. We have not marked these instances manually, so that hyphenated words where the hyphen is needed are marked distinctly and these hyphens can be retained for later processing.

7 Orthography Normalization

As different publications have used slightly different transcription systems, there is a need to unify these so that comparative searches can be done, and the material can be connected to contemporary language technology. We need to facilitate corpus entries for lexicon and morphological analyzers, and this cannot be done if the transcriptions are wildly different. At the same time we recognize that the transcriptions are often very detailed and may contain dialectal features that are important, but cannot be easily expressed in the literary language and contemporary orthography.

Thereby what we are looking for is sort of a middle way where the representation is brought as close to the orthography as possible, but leaves some wiggle room for original details in the transcription. This is necessarily a partly impressionistic goal. Partanen (2024) discussed this task in their study where Large Language Models were tested in transliteration of endangered Uralic languages, and also in this context the task was not only a transliteration, but toward a normalization as well. As we are also keeping the original transcriptions, no information is lost, and various transcription layers can be envisioned.

If all transcriptions in different sources are essentially phonemic, with additional phonetic features present, one can also envision a solution where the harmonized transcription would have the same phonemic representation in all of them. At the same time, this would not be very useful for the language community and it would remain very unusable from the point of view of language technology. Since the Livonian orthography is actually fairly phonemic, it does not seem reasonable to aim toward anything else.

The workflow we have constructed in the pilot project is that the transcription is automatically transformed toward the orthography with a rule-based Python script. The script is adjusted based on the feedback we receive from the experts at the University of Latvia Livonian Institute. The rules are slightly different for each publication, but the output should match as well as possible. Figure 3 illustrates the transformed text.

The evaluation of texts normalized from phonetic transcription demonstrated that the results were very close to those that could be achieved through manual transcription. While certain orthographic inconsistencies were observed—primarily related to compounding and to the morphological principles applied in Livonian orthography (e.g., *ītõ-kabāl* vs. *īdõkabāl* ’all the time; always’; *jetspēḑõn* vs. *jedspēḑõn* ’away’; and in several cases involving specific verb or noun types such as *tīedist* ‘[they] did know’ vs. *tīedizt* ’[they] knew’, *taggist* vs. *taggizt* ‘ones behind’) — the overall output was remarkably close to a gold-standard normalization. This indicates that the process can significantly reduce the effort and resources required for such transcription tasks.

8 Corpus Creation

In the corpus creation phase we parse the Transkribus Page XML documents with Python through the Transkribus API. The layout structures and tagging described in the earlier section is used to retrieve the correct structure. The resulting corpus contains transcribed Livonian sentences and information about the matching Finnish translation at the paragraph level. At the moment it does not seem to be possible to join Livonian and Finnish automatically at the sentence level. Another option would be to match the lines by position, where the Livonian sentence would have as a translation the roughly corresponding lines or portion in the

Finnish translation. Sentences cannot be directly aligned as there are small editorial differences between the versions, and sentence punctuation in the Finnish does not always correspond perfectly to the Livonian punctuation. Naturally, the text collection has not been created originally with the perfect sentence level matching in mind, and this is just a feature of the material for which we are still deciding the best approach.

As mentioned above, the alignment is based simply on the number of paragraphs. Similarly in the whole book there is a fixed number of texts. The metadata that has been tagged is extracted, and can be accessed directly in the parsing phase. However, we have found it more convenient to store the metadata in a separate table, where it can be extended and clarified. We often have text-specific details, i.e. location as coordinates, which we in any case want to associate with each text, but we do not want to add them to the digitized work. In principle, any format would work for this additional metadata, as we can always merge it with the corpus using the number of the text as a shared field.

9 Forced Alignment

Forced alignment is not a task we have yet applied to the workflow, but aligning the text and audio is a critical phase that should be performed in the cases where the original recordings are available. The current approach is to manually segment a portion of at least some tens of minutes at the utterance level, so that there is some baseline data against which we can evaluate different alignment approaches. This initial work is presented in the Figure 4.

There have been recent experiments in using Montreal Forced Aligner (McAuliffe et al., 2017) within the Uralic-Amazonian collaboration that has been taking place between the Universities of Helsinki and Belem (Rueter and Partanen, 2025). The idea in this work has been to align utterances in Komi-Zyrian and Apurinã at the phoneme level, and the goal has been to test if the alignment model can be trained with this type of data and how good the results are. If the results are positive, we could expect that the same can also be done with other endangered languages with similar resources, a category into which Livonian also fits very well.

With forced alignment it is important to notice that there are at least two fairly different scenarios in which forced alignment can be used. Most typically, it seems, we are discussing a scenario where there is perfect matching with the transcription and the audio segment, and the task is to match every phoneme as accurately as possible. However, this is not what we want to do first with the Livonian materials, but we would be very happy to have it at a later stage.

The situation is very different when there is a long and edited transcription, which corresponds to the audio, but not perfectly. The mechanisms needed here are fairly different, and the model needs specific logic to react to situations where there is no match, or when there is some additional content that cannot be matched. Ideally, in these situations the matching would be done at the utterance level, as the transcription would probably be revised against the audio once it is coarsely aligned. However, it seems that there is less support for this type of fuzzier alignment than there is for phoneme level alignment. We need to investigate what kind of forced alignment tool would work the best in our initial scenario. The tools that focus to phoneme level alignment should be used when the transcription is already aligned at the utterance level, and ideally manually adjusted if needed.

10 Universal Dependencies

There are currently several Universal Dependencies treebanks available for Uralic languages, and among these are also minor Finnic languages. In recent years two treebanks have been published for Karelian (Pirinen, 2019), one for Veps, another for Tundra Nenets, and soon there will be one for Northern Mansi. This is definitely a domain where new progress would be very welcome.

Seppo Suhonen's materials discussed here could suit this type of development work very well, too. The recorded and transcribed texts, that are unproblematic from the point of copyright, would be well fitting for open projects such as the Universal Dependencies. However, there is certainly a need to take into account both spoken and written Livonian, and also older recordings and contemporary speech. The situation is thereby fairly similar to Komi treebanks, for example, where different varieties and genres have been accounted for (Partanen et al., 2018). Similarly, the spoken language treebanks have various questions unique to them, especially in how the speech-specific phenomena are annotated (Dobrovoljc, 2022).

11 Conclusion

As outlined above, it is possible to envision a pipeline where the Livonian transcriptions are aligned with the Finnish translation, aligned with the audio, possibly even on a phoneme or word level, normalized to the Livonian orthography, and analyzed with contemporary morpho-syntactic analyzer for Livonian. This type of resource would be useful in a wide variety of research tasks, but it would also serve the language community in very detailed applications, including searching across the corpus and using both text and audio versions in language learning and education.

At the same time this data would be useful in tasks such as training automatic speech recognition tools. Especially in the context where there is a large number of recordings from a few individuals, it seems realistic to reach a very high recognition accuracy with the current methods, as reported in similar scenarios half a decade ago by Partanen et al. (2020) and outlined for Livonian recently by Ernštreits (2024).

One particularly promising outcome of the successful normalization of Livonian texts documented in phonetic transcription would be the integration into Livonian databases of materials from the 1938 Livonian–German dictionary (Kettunen, 1938), which contains a substantial amount of linguistic data, especially lexemes, not yet represented in the current Livonian database cluster and providing valuable expansion of vocabulary acccessible for Livonian speakers and learners.

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).

The digitization project of the Finno-Ugrian Society described in the study has been done with the funding by the Finnish Association for Scholarly Publishing.

References

Alexandre Arkhipov, Anna Barinskaya, and Roman Shtefura. 2021. Using handwritten text recognition on bilingual Evenki-Russian manuscripts of Konstantin Rychkov. *Scripta & E-Scripta*, 21.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Valts Ernštreits. 2024. Towards the speech recognition for Livonian. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 76–80, Helsinki, Finland. Association for Computational Linguistics.

V. Ernštreits. 1999. *Lībiešu-latviešu-lībiešu vārdnīca*. Līvõ Kultūr Sidām, Rīga.

Valts Ernštreits, Signis Vāvere, Tiit-Rein Viitso, Pētõr Damberg, Milda Kurpniece, Gunta Kļava, Uldis Balodis, Tuuli Tuisk, Gita Kūla, Marili Tomingas, Sven-Erik Soosaar, Anna Sedláčková, and Toms Jurgenovskis. 2024. *Livonian Language and Culture Resource Platform "Livonian.tech"*. University of Latvia Livonian Institute, Riga.

Santra Jantunen. 2025. *Livonian Verbal Derivation: Inherited Characteristics and Contact-Induced Change*. Doctoral dissertation (article-based), University of Helsinki, Helsinki. Doctoral Programme in Language Studies.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.

L. Kettunen. 1938. *Livisches Wörterbuch*, volume 5 of *Lexica Societatis Fenno-Ugricae*. Finno-Ugrian Society, Helsinki.

William Lamb, Beatrice Alex, and Mark Sinclair. 2022. Handwriting recognition for Scottish Gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 60–70.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, volume 2017, pages 498–502.

Julius Mägiste. 2006. *Muistoja Liivinrannasta: Liivin kieltä Ruotsista*. Number 250 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki.

Niko Partanen. 2024. Using large language models to transliterate endangered Uralic languages. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 81–88.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW*

2018), November 2018, Brussels, Belgium, pages 126–132.

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 15-18, 2022*, pages 370–380. University of Oslo Library.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 523–533, Hanoi, Vietnam. Association for Computational Linguistics.

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.

Jack Rueter. 2024. On searchable Mordvin corpora at the Language Bank of Finland, EMERALD. *Journal of Data Mining & Digital Humanities*, (V. The contribution of corpora).

Jack Rueter, Olga Erina, and Nadezhda Kabaeva. 2024. On Erzya and Moksha corpora and analyzer development, ERME-PSLA 1950s. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 67–75.

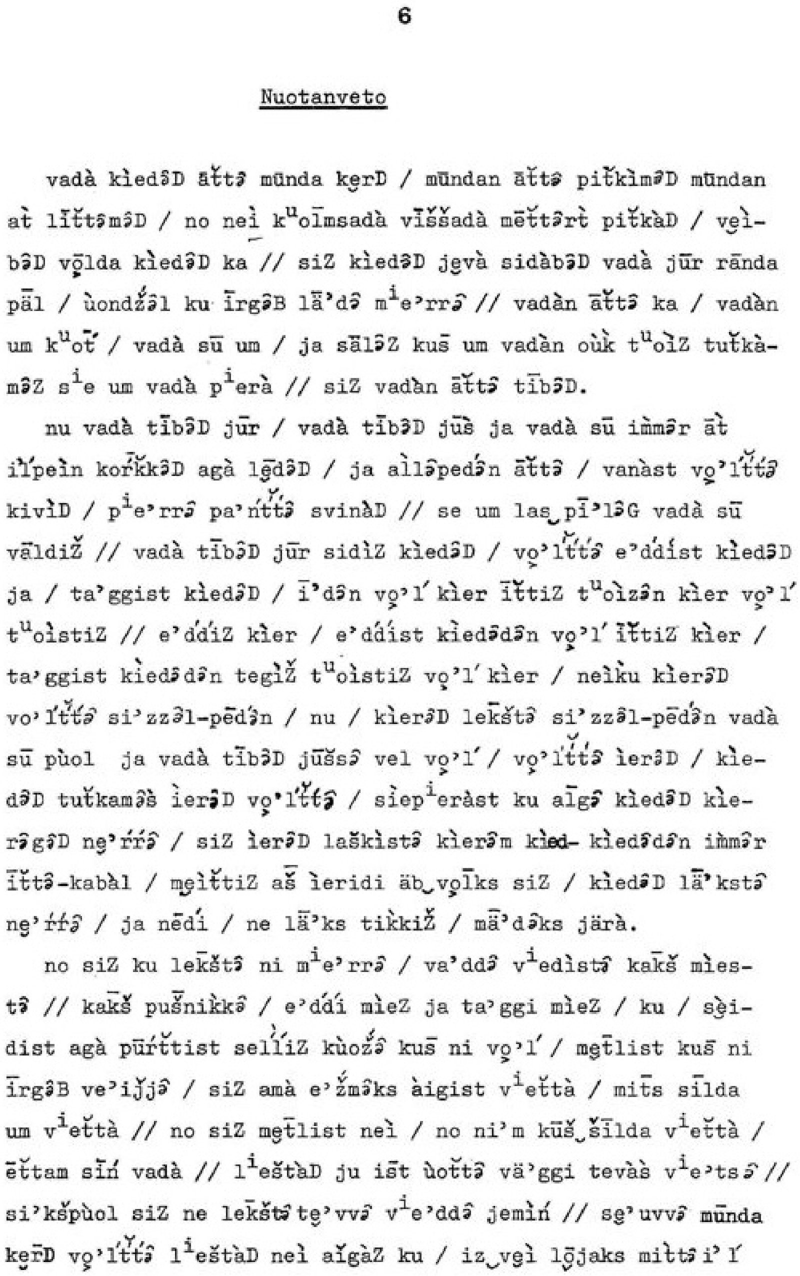
Jack Rueter and Niko Partanen. 2019. On new text corpora for minority languages on the Helsinki korp.csc.fi server. In *Èlektronnaâ pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy. Ufa, 11–12 dekabrâ 2019 goda*, pages 32–36.

Jack Rueter and Niko Partanen. 2025. Language technology for the Uralic languages in an Amazonian context. *Dutkansearvvi dieđalaš áigečála*, 9(1):137–147.

E. N. Setälä. 1953. *Näytteitä liivin kielestä*. Number 106 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki.

Seppo Suhonen. 1975. *Liivin kielen näytteitä*. Number 5 in Castrenianumin toimitteita. Helsinki.

F. J. Wiedemann. 1861. Joh. Andreas Sjögren's Livische Grammatik nebst Sprachproben. In *Joh. Andreas Sjögren's Gesammelte Schriften. Band II. Teil I*. Kaiserlischen Akademie der Wissenschaften, St. Petersburg.



6

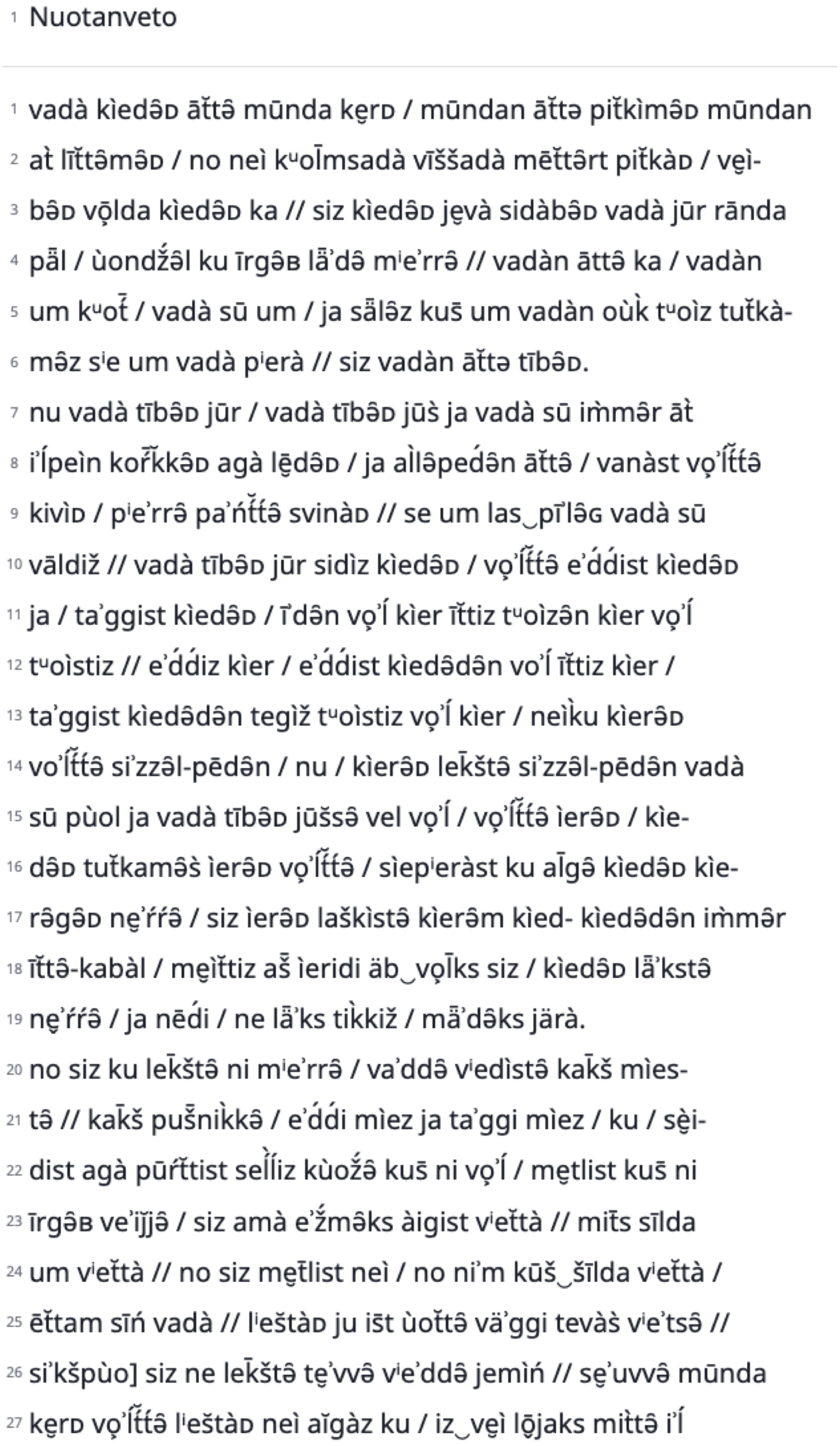
Nuotanveto

vadà kìedə̂D ātt̆ə mūnda kẹrD / mūndan ātt̆ə pit̆kìməD mūndan àt lītt̆əməD / no neì k^{u}ōlmsadà vīšš̆adà mētt̆ərt̀ pit̆kàD / vẹìbəD vọ̄lda kìedə̂D ka // siZ kìedəD jẹvà sidàbəD vadà jūr rānda pǟl / ùondž́əl ku īrgəB lǟ'də m^{i}e'rrə̂ // vadàn ātt̆ə ka / vadàn um k^{u}ōt́ / vadà sū um / ja sālə̂Z kuš̄ um vadàn oùk t^{u}ōìZ tut̆kàməZ s^{i}e um vadà p^{i}erà // siZ vadàn ātt̆ə tībəD.

nu vadà tībə̂D jūr / vadà tībəD jūš ja vadà sū ìmmər āt̀ ìl̆peìn kor̆k̆kəD agà lẹ̄də̂D / ja aìlləpedə́n ātt̆ə / vanàst vọ'l̆t́t́ə kivìD / p^{i}e'rrə̂ pa'ń̆t́t́ə svinàD // se um las‿pī'lə̂G vadà sū vāldiŽ // vadà tībə̂D jūr sidìZ kìedə̂D / vọ'l̆t́t́ə e'ddíst kìedəD ja / ta'ggist kìedə̂D / ī'dən vọ'l' kìer īt̆tiZ t^{u}oìzə̂n kìer vọ'l' t^{u}oìstiZ // e'ddíZ kìer / e'ddíst kìedədən vọ'l' īt̆tiZ kìer / ta'ggist kìedədən tegìŽ t^{u}oìstiZ vọ'l' kìer / neìku kìerəD vo'l̆t́t́ə si'zzəl-pēdən / nu / kìerəD lekš̄tə si'zzəl-pēdən vadà sū pùol ja vadà tībəD jūš̆sə vel vọ'l' / vọ'l̆t́t́ə ìerəD / kìedəD tut̆kaməš̀ ìerəD vọ'l̆t́t́ə / sìep^{i}eràst ku al̄gə kìedəD kìerəgəD nẹ'ŕŕə / siZ ìerəD laškìstə kìerəm ~~kìed-~~ kìedədən ìmmər īt̆tə-kabàl / mẹìt̆tiZ ăš̄ ìeridi äb‿vọ̄lks siZ / kìedəD lǟ'kstə nẹ'ŕŕə / ja nēdí / ne lǟ'ks tìkkiŽ / mǟ'dəks jǟrà.

no siZ ku lek̄štə ni m^{i}e'rrə / va'ddə v^{i}edìstə kak̄š mìestə // kak̄š puš̄nìkkə / e'ddí mìeZ ja ta'ggi mìeZ / ku / sẹ̀idist agà pūŕt̆tist sellíZ kùož́ə kuš̄ ni vọ'l' / mẹt̄list kuš̄ ni īrgəB ve'iȷ̆jə / siZ amà e'ž́məks àigist v^{i}et̆tà / mit̄s sīlda um v^{i}et̆tà // no siZ mẹt̄list neì / no ni'm kūš‿š̆īlda v^{i}et̆tà / ēt̆tam sīń vadà // l^{i}eštàD ju iš̄t ùot̆tə vä'ggi tevàš v^{i}e'tsə // si'kšpùol siZ ne lek̄štə tẹ'vvə v^{i}e'ddə jemìń // sẹ'uvvə mūnda kẹ̄rD vọ'l̆t́t́ə l^{i}eštàD neì ail̆gàZ ku / iz‿vẹì lọ̄jaks mitt̀ə i'l'

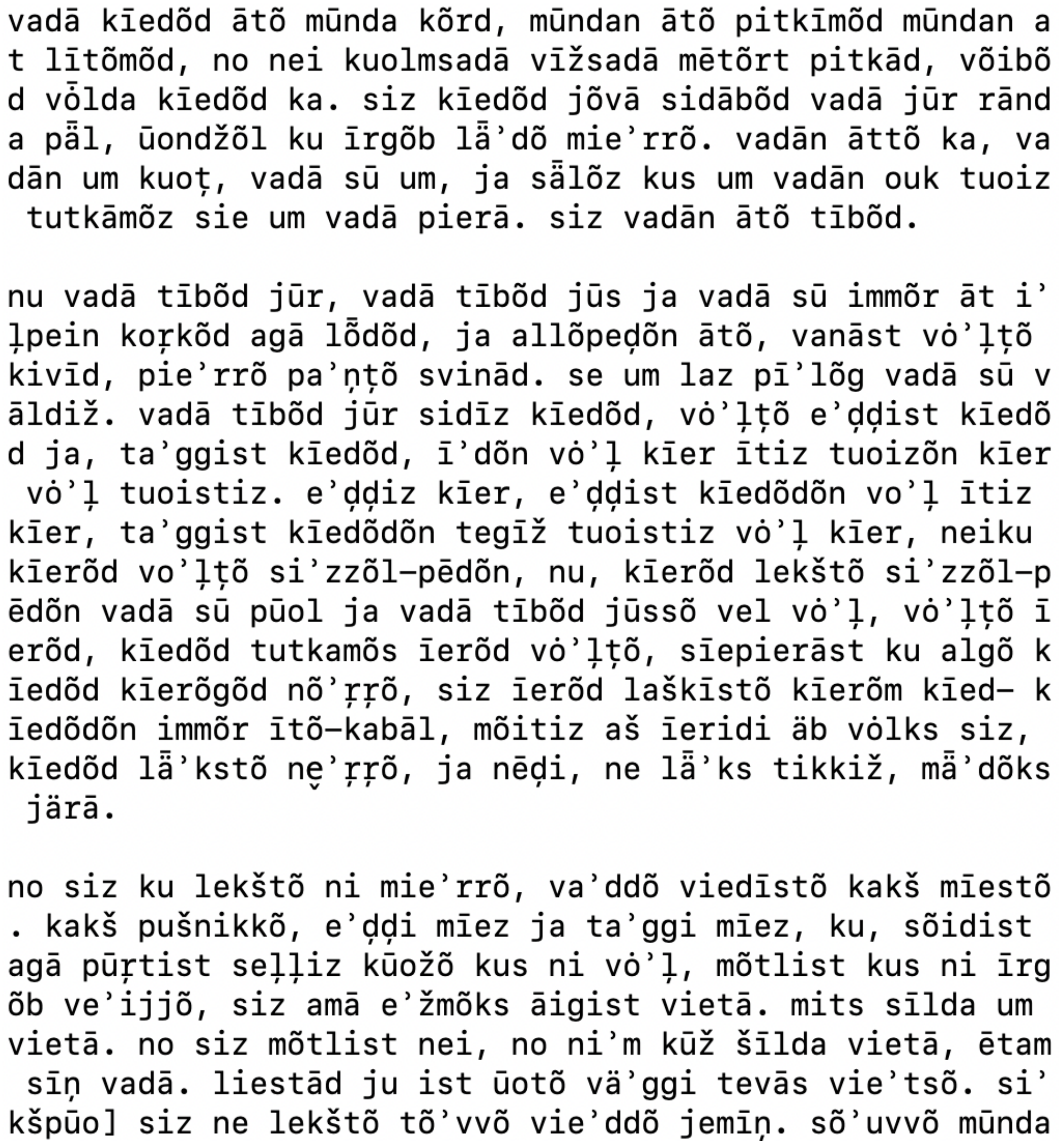
Figure 1: Example of the scanned page showing the Livonian transcription of Suhonen (1975, 6)



1 Nuotanveto

1 vadà kìedə̂ᴅ āt̆tə̂ mūnda kẹrᴅ / mūndan āt̆tə pit̆kìmə̂ᴅ mūndan
2 at̀ līt̆tə̂mə̂ᴅ / no neì kᵘol̄msadà vīš̆šadà mēt̆tə̂rt pit̆kàᴅ / vẹì-
3 bə̂ᴅ vǭlda kìedə̂ᴅ ka // siz kìedə̂ᴅ jẹvà sidàbə̂ᴅ vadà jūr rānda
4 pā̆l / ùondž́ə̂l ku īrgə̂ʙ lā̆ʼdə̂ mⁱeʼrrə̂ // vadàn āttə̂ ka / vadàn
5 um kᵘot̄̆ / vadà sū um / ja sā̆lə̂z kuš̄ um vadàn oùk̀ tᵘoìz tut̆kà-
6 mə̂z sⁱe um vadà pⁱerà // siz vadàn āt̆tə tībə̂ᴅ.
7 nu vadà tībə̂ᴅ jūr / vadà tībə̂ᴅ jūs̀ ja vadà sū im̀mə̂r āt̀
8 iʼĺpeìn kor̄̆kkə̂ᴅ agà lẹ̄də̂ᴅ / ja al̀lə̂ped́ə̂n āt̆tə̂ / vanàst vọʼĺt̆t́ə̂
9 kivìᴅ / pⁱeʼrrə̂ paʼńt̆t́ə̂ svinàᴅ // se um las‿pīʼlə̂ɢ vadà sū
10 vāldiž // vadà tībə̂ᴅ jūr sidìz kìedə̂ᴅ / vọʼĺt̆t́ə̂ eʼd́d́ist kìedə̂ᴅ
11 ja / taʼggist kìedə̂ᴅ / īʼdə̂n vọʼĺ kìer īt̆tiz tᵘoìzə̂n kìer vọʼĺ
12 tᵘoìstiz // eʼd́d́iz kìer / eʼd́d́ist kìedə̂də̂n voʼĺ īt̆tiz kìer /
13 taʼggist kìedə̂də̂n tegìž tᵘoìstiz vọʼĺ kìer / neìk̀u kìerə̂ᴅ
14 voʼĺt̆t́ə̂ siʼzzə̂l-pēdə̂n / nu / kìerə̂ᴅ lek̄̆štə̂ siʼzzə̂l-pēdə̂n vadà
15 sū pùol ja vadà tībə̂ᴅ jūš̆šə̂ vel vọʼĺ / vọʼĺt̆t́ə̂ ìerə̂ᴅ / kìe-
16 də̂ᴅ tut̆kamə̂s̀ ìerə̂ᴅ vọʼĺt̆t́ə̂ / sìepⁱeràst ku al̄gə̂ kìedə̂ᴅ kìe-
17 rə̂gə̂ᴅ nẹʼŕŕə̂ / siz ìerə̂ᴅ laškìstə̂ kìerə̂m kìed- kìedə̂də̂n im̀mə̂r
18 īt̆tə̂-kabàl / mẹìt̆tiz aš̄̆ ìeridi äb‿vọl̄ks siz / kìedə̂ᴅ lā̆ʼkstə̂
19 nẹʼŕŕə̂ / ja nēd́i / ne lā̆ʼks tik̀kiž / mā̆ʼdə̂ks järà.
20 no siz ku lek̄̆štə̂ ni mⁱeʼrrə̂ / vaʼddə̂ vⁱedìstə̂ kak̄̆š mìes-
21 tə̂ // kak̄̆š puš̄̆nik̀kə̂ / eʼd́d́i mìez ja taʼggi mìez / ku / sẹ̀i-
22 dist agà pūŕt̆tist sel̀ĺiz kùož́ə̂ kuš̄ ni vọʼĺ / mẹtlist kuš̄ ni
23 īrgə̂ʙ veʼij̆jə̂ / siz amà eʼž́mə̂ks àigist vⁱet̆tà // mit̄s sīlda
24 um vⁱet̆tà // no siz mẹt̄list neì / no niʼm kūš‿šīlda vⁱet̆tà /
25 ēt̆tam sīń vadà // lⁱeštàᴅ ju išt ùot̆tə̂ väʼggi tevàs̀ vⁱeʼtsə̂ //
26 siʼkšpùo] siz ne lek̄̆štə̂ tẹʼvvə̂ vⁱeʼddə̂ jemìń // sẹʼuvvə̂ mūnda
27 kẹrᴅ vọʼĺt̆t́ə̂ lⁱeštàᴅ neì aĭgàz ku / iz‿vẹì lǭjaks mit̀tə̂ iʼĺ

Figure 2: Example of the text recognized Unicode text showing the Livonian transcription corresponding to the text in Suhonen (1975, 6)

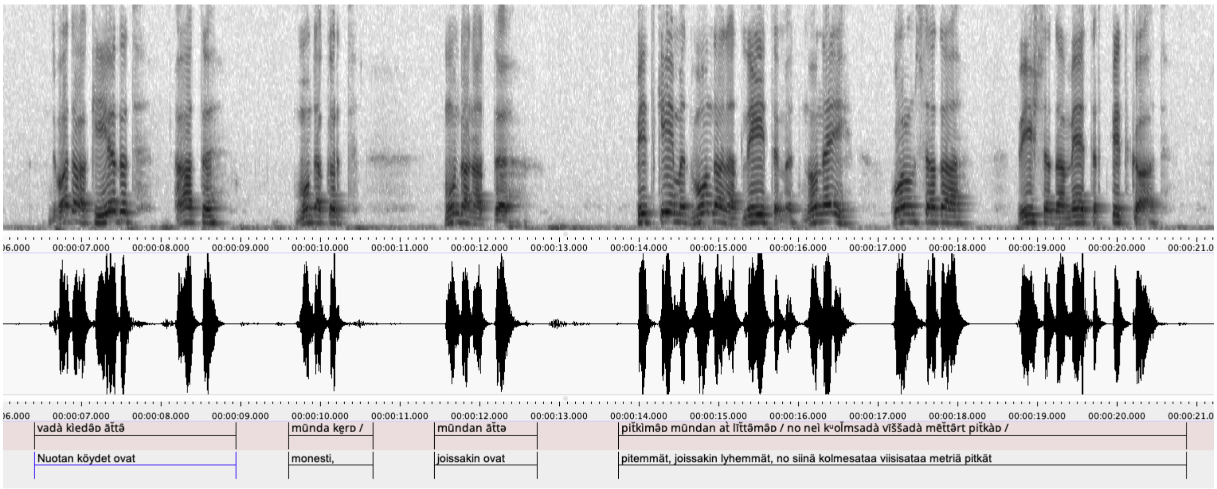


```
vadā kīedõd ātõ mūnda kõrd, mūndan ātõ pitkīmõd mūndan a
t lītõmõd, no nei kuolmsadā vīžsadā mētõrt pitkād, võibõ
d vȱlda kīedõd ka. siz kīedõd jõvā sidābõd vadā jūr rānd
a pǟl, ūondžõl ku īrgõb lǟʼdõ mieʼrrõ. vadān āttõ ka, va
dān um kuoṭ, vadā sū um, ja sǟlõz kus um vadān ouk tuoiz
 tutkāmõz sie um vadā pierā. siz vadān ātõ tībõd.

nu vadā tībõd jūr, vadā tībõd jūs ja vadā sū immõr āt iʼ
ḷpein koṛkõd agā lȭdõd, ja allõpeḍõn ātõ, vaṇāst vȯʼḷṭõ
kivīd, pieʼrrõ paʼṇṭõ svinād. se um laz pīʼlõg vadā sū v
āldiž. vadā tībõd jūr sidīz kīedõd, vȯʼḷṭõ eʼḍḍist kīedõ
d ja, taʼggist kīedõd, īʼdõn vȯʼḷ kīer ītiz tuoizõn kīer
 vȯʼḷ tuoistiz. eʼḍḍiz kīer, eʼḍḍist kīedõdõn voʼḷ ītiz
kīer, taʼggist kīedõdõn tegīž tuoistiz vȯʼḷ kīer, neiku
kīerõd voʼḷṭõ siʼzzõl–pēdõn, nu, kīerõd lekštõ siʼzzõl–p
ēdõn vadā sū pūol ja vadā tībõd jūssõ vel vȯʼḷ, vȯʼḷṭõ ī
erõd, kīedõd tutkamõs īerõd vȯʼḷṭõ, sīepierāst ku algõ k
īedõd kīerõgõd nõʼṛṛõ, siz īerõd laškīstõ kīerõm kīed– k
īedõdõn immõr ītõ–kabāl, mõitiz aš īeridi äb vȯlks siz,
kīedõd lǟʼkstõ nẹ̬ʼṛṛõ, ja nēḍi, ne lǟʼks tikkiž, mǟʼdõks
 järā.

no siz ku lekštõ ni mieʼrrõ, vaʼddõ viedīstõ kakš mīestõ
. kakš pušnikkõ, eʼḍḍi mīez ja taʼggi mīez, ku, sõidist
agā pūṛtist seḷḷiz kūožõ kus ni vȯʼḷ, mõtlist kus ni īrg
õb veʼijjõ, siz amā eʼžmõks āigist vietā. mits sīlda um
vietā. no siz mõtlist nei, no niʼm kūž šīlda vietā, ētam
 sīṇ vadā. liestād ju ist ūotõ väʼggi tevās vieʼtsõ. siʼ
kšpūo] siz ne lekštõ tõʼvvõ vieʼddõ jemīṇ. sõʼuvvõ mūnda
```

Figure 3: Example of the automatic orthography normalization, corresponding to the text in Suhonen (1975, 6)



vadà kìedâɒ ā̆ttâ | mūnda kẹrɒ / | mūndan ā̆ttə | pit̆kìmâɒ mūndan at̀ lī̆t̆təmâɒ / no neì kᵘol̄msadà vī̆ššadà mē̆t̆tərt pit̆kàɒ /

Nuotan köydet ovat | monesti, | joissakin ovat | pitemmät, joissakin lyhemmät, no siinä kolmesataa viisisataa metriä pitkät

Figure 4: Example of the text recognized Unicode text showing the Livonian transcription corresponding to the text in Suhonen (1975, 6). This figure displays an experiment and the materials will likely be structured differently and managed in other more suitable environments.

Siberian Ingrian Finnish: FST and IGTs

Ivan Ubaleht
Omsk State Technical University
Omsk, Russia
last@gmail.com

Abstract

This paper presents the current version of the finite-state transducer for the Siberian Ingrian Finnish. Our finite-state transducer uses two-level morphology. We use LexC and TwolC languages together with HFST tools to develop lexicons and phonological rules, as well as to compile the transducer. The paper also provides a description of the morphological system of Siberian Ingrian Finnish. In addition, we present a collection of interlinear glossed texts in Siberian Ingrian Finnish, provided in a machine-readable format.

1 Introduction

The solution of computational morphology tasks is an important stage of language processing. Chapter 3 proposes a finite-state transducer based on the two-level morphology for addressing computational morphology tasks for Siberian Ingrian Finnish. Labeled data are also required for the successful solution of computational morphology tasks. In Chapter 4, we present interlinear glossed texts in Siberian Ingrian Finnish that have been published in a machine-readable format. Siberian Ingrian Finnish is a language with a rich morphological system; a brief overview of the language and its available resources is provided in Chapter 2.

2 Background

2.1 An overview of Siberian Ingrian Finnish

The Siberian Ingrian Finnish Language is an Ingrian Finnish – Ingrian (Izhorian) mixed language. The ancestors of the speakers of Siberian Ingrian Finnish spoke Lower Luga Ingrian Finnish and Lower Luga Ingrian varieties and lived in the lower reaches of the Luga River (Yamburgsky Uyezd). They were exiled to Western Siberia in 1803–1804 for their participation in a peasant uprising against Baron von Ungern-Sternberg (Kuznetsova, 2016, p. 14; Sidorkevich, 2014, pp. 23–24).

This language has been investigated by a number of linguists. D. V. Sidorkevich conducted research on this language between 2008 and 2014 (Sidorkevich, 2014, 2011). She introduced the term "Siberian Ingrian Finnish" (Russian "Сибирский ингерманландский идиом"). Siberian Ingrian Finnish was also studied by R. E. Nirvi (Nirvi, 1972), V. Zlobina (Zlobina, 1971, 1972), N.V. Kuznetsova (Kuznetsova, 2016; Kuznetsova and Verkhodanova, 2019), M. Z. Muslimov and F. I. Rozhansky. V. Zlobina introduced the term "Korlaks" (Russian "Корлаки", Finnish "Korlakat") to refer to the group speaking this language.

In 2025, there is still a group of elderly people who use Siberian Ingrian Finnish in domestic communication in the Ryzhkovo settlement (Krutinsky District of Omsk Oblast). Small groups and isolated speakers of Siberian Ingrian Finnish also live in other settlements of Omsk Oblast and in Estonia. A pessimistic estimate of the number of Siberian Ingrian Finnish speakers is about 30. This estimate is based on the fact that the author of this paper personally knows or is aware of 21 speakers of the language. An optimistic estimate, including semi-speakers, is about 100–150 people.

2.2 The language resources of Siberian Ingrian Finnish

The resources of the Siberian Ingrian Finnish language are summarized in Table 1. As can be seen in Table 1, a certain number of texts are currently available for Siberian Ingrian Finnish. Additional texts are planned to be collected through audio transcription. These texts require morphological glossing. Finite-state transducers provide significant assistance in the glossing process.

We have quite a large amount of audio data for Siberian Ingrian Finnish, see Table 1. Therefore, we previously created annotations for these au-

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 123–126
December 10-12, 2025 ©2025 Association for Computational Linguistics

Resource type	Resource size
Audio data (2008-2025)	120 hours
Audio data published under a Creative Commons 4.0 license	5 hours
Video data	2 hours
Texts (mostly manual transcriptions of audio data)	42,000 tokens
IGT collection	150 sentences
Number of speakers recorded	31

Table 1: Language resources of Siberian Ingrian Finnish.

dio data and developed software for working with the annotations (Ubaleht and Raudalainen, 2022). However, the process of annotating audio data also requires automation, which became another reason for developing the Siberian Ingrian Finnish finite-state transducer.

3 Development of the Siberian Ingrian Finnish Finite-State Transducer

Currently, many computational morphology tasks, including those for low-resource languages, have been effectively addressed using models based on neural networks (Goldman et al., 2023; Wu et al., 2020; Liu, 2021). Nevertheless, approaches grounded in linguistic knowledge and employing finite-state transducers continue to provide benefits under certain conditions (Morozov et al., 2024; Merzhevich et al., 2022; Beemer et al., 2020). This is particularly pronounced in scenarios where processing morphologically rich languages is required and where training data are limited.

Siberian Ingrian Finnish lacks available training data. Currently, interlinear glossed texts for this language (which could serve as training data in the future) are still being created. Therefore, for solving morphological analysis and synthesis tasks for Siberian Ingrian Finnish, we are developing a solution based on finite-state transducers. We use the two-level morphology approach for developing finite-state transducers for Siberian Finnish, using LexC, TwolC, and the HFST toolkit (Lindén et al., 2011). The source code of the LexC and TwolC files for the current version of the finite-state transducer is accessible to the public on GitHub[1]. Currently, the transducer includes approximately 100 stems.

3.1 Morphological processing of Siberian Ingrian Finnish nouns

The morphological paradigm of nouns in Siberian Ingrian Finnish includes the declension of nouns by case and number, see Table 2. In its present state, Siberian Ingrian Finnish has eleven cases and two numbers. In practice, the adessive case and allative case have merged into a single syncretic adessive–allative case (Sidorkevich, 2011). However, in our finite-state transducer, we treat these cases separately. Siberian Ingrian Finnish nouns have five stems:

- NOM.SG: the stem used for the nominative singular form.
- OBL.SG: the stem used for all singular oblique cases except the illative and the partitive.
- PART.SG: the stem used for the partitive singular form.
- ILL.SG: the stem used for the illative singular form.
- OBL.PL: the stem used for all plural oblique cases.

Siberian Ingrian Ingrian Finnish nouns do not have a regular plural suffix like *-de* in Estonian or *-loi* in Izhorian. Plural forms are formed through stem alternations. These alternations are expressed by the OBL.PL stem. For example, the words *koir* (dog) and *sisar* (sister) belong to the same morphophonological type CS2 (Sidorkevich, 2014, p. 172), but *koira-n* (dog.SG-GEN), *koiri-n* (dog.PL-GEN) vs. *sisara-n* (sister.SG-GEN), *sisaro-n* (sister.PL-GEN).

Sixteen morphophonological types have been identified for Siberian Finnish nouns. D. V. Sidorkevich labels them as follows: CS1–CS8 for words with a consonant stem (Sidorkevich, 2014, pp. 164-165), and VS1–VS8 for words with a vowel stem (Sidorkevich, 2014, pp. 165-166).

In Siberian Ingrian Finnish, there are also morphophonological types that have not yet been documented. In Siberian Ingrian Finnish, a set of rules reflecting consonant alternations (CA1–CA5) and vowel alternations (VA1–VA5) is defined (Sidorkevich, 2014, pp. 161-162). Using alternation rules, it is not always possible to reliably derive the

[1] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/src/morphological-analyzer/fst

Case	Singular			Plural		
	Stem	Affix	Example	Stem	Affix	Example
Nominative	NOM.SG	ø	*käsi*	OBL.SG	*-t*	*käe-t*
Genitive	OBL.SG	*-n*	*käe-n*	OBL.PL	*-n*	*kässi-n*
Partitive	PRT.SG	ø	*kätt*	OBL.PL	*-j*	*kässi-j*
Illative	ILL.SG	ø	*kätte*	OBL.PL	*-s*	*kässi-s*
Inessive	OBL.SG	*-s*	*käe-s*	OBL.PL	*-s*	*kässi-s*
Elative	OBL.SG	*-st*	*käe-st*	OBL.PL	*-st*	*kässi-st*
Allative	OBL.SG	*-l*	*käe-l*	OBL.PL	*-l*	*kässi-l*
Adessive	OBL.SG	*-l*	*käe-l*	OBL.PL	*-l*	*kässi-l*
Ablative	OBL.SG	*-lt*	*käe-lt*	OBL.PL	*-lt*	*kässi-lt*
Translative	OBL.SG	*-ks*	*käe-ks*	OBL.PL	*-ks*	*kässi-ks*
Comitative	OBL.SG	*-nkA*	*käe-nkä*	OBL.PL	*-nkA*	*kässi-nkä*

Table 2: Declension paradigm of nouns in Siberian Ingrian Finnish, with the example for morphophonological type VS3.

other noun stems from the main stem NOM.SG for all morphophonological types. When it is difficult to derive the other stems from the main stem NOM.SG, we record all five stems in the transducer lexicon, see *käsi* (1). As an example, (2) shows a lexicon that can be used to generate word forms from stem OBL.SG.

In some cases, it is possible to derive all stems from the NOM.SG stem using phonological rules from TwolC, so the word is represented in the lexicon by a single stem, see *koir*, *sisar* (1). We assume that in the future, for most morphophonological types, it will be possible to find phonological rules for a convenient representation of words in the lexicons.

(1) Lexicon containing noun stems

```
LEXICON NounStems

käsi:käsi   NomSgStem ;
käsi:käe    OblSgStem ;
käsi:kätt   PrtSgStem ;
käsi:kätte  IllSgStem ;
käsi:kässi  OblPlStem ;
koir:koir CS2-I ;
sisar:sisar CS2-O ;
```

3.2 Morphological processing of Siberian Ingrian Finnish verbs

The paradigm of verb inflection in Siberian Ingrian Finnish is not described in detail in this paper. The morphophonological types of Siberian Ingrian Finnish verbs remain poorly studied. D. V. Sidorkevich identifies 13 stems in the verbs of the Siberian Finnish language (Sidorkevich, 2014, p. 210).

Verbs such as *korja* (to pick up) and *harja* (to comb) can be represented in the lexicon by their infinitive stem, and all other forms are derived simply by adding suffixes. For all other morphophonological verb types, the lexicon must include between 3 and 13 verb stems. We assume that, by applying phonological rules, the number of verb stems in the lexicon can be reduced.

(2) Morphotactics for generating word forms from stem OBL.SG

```
LEXICON OblSgStem

+N+Gen+Sg:n # ;
+N+Ine+Sg:s # ;
+N+Ela+Sg:st # ;
+N+All+Sg:l # ;
+N+Ade+Sg:l # ;
+N+Abl+Sg:lt # ;
+N+Tra+Sg:ks # ;
+N+Com+Sg:nk%{A%} # ;
+N+Nom+Pl: t # ;
```

4 The Interlinear Glossed Texts in Siberian Ingrian Finnish

There are often no written texts available for many low-resource languages. Therefore, collections of interlinear glossed texts (IGTs) are important for providing these languages with linguistic resources.

D. V. Sidorkevich collected and glossed texts in Siberian Ingrian Finnish, but these texts were in a format not suitable for computational processing (Sidorkevich, 2014). We converted this collection of IGTs into a machine-readable format (3) and made it openly available[2]; for example, a similar IGT format was used in the SIGMORPHON 2023 Shared Task on Interlinear Glossing[3].

(3) An example from our IGT collection

```
\t Miltajst sié kahest podarkast tahot?

\m miltajs-t sié kahe-st podarka-st taho-t

\g which-PRT 2SG two-ELA gift-ELA want-2SG

\l Which of the two gifts do you want?
```

5 Conclusion

In this paper, we have presented a finite-state transducer for Siberian Ingrian Finnish and a collection of interlinear glossed texts in this language. Future work includes: (i) expanding the transducer's lexicon to cover a larger vocabulary (we plan to add approximately 400-500 new stems to the lexicon by February 2026); (ii) developing phonological alternation rules to improve verb processing; (iii) glossing new Siberian Ingrian Finnish texts (including audio data annotations) using the FST; (iv) applying the FST and the IGTs in the language revitalization practices of Siberian Ingrian Finnish.

References

Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, and 1 others. 2020. Linguist vs. machine: Rapid development of finite-state morphological grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 162–170.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. Sigmorphon–unimorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125.

Natalia Kuznetsova. 2016. Evolution of the non-initial vocalic length contrast across the finnic varieties of ingria and adjacent areas. *Linguistica Uralica*, 52(1):1–25.

Natalia Kuznetsova and Vasilisa Verkhodanova. 2019. Phonetic realisation and phonemic categorisation of the final reduced corner vowels in the finnic languages of ingria. *Phonetica*, 76(2-3):201–233.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Ling Liu. 2021. Computational morphology with neural network approaches. *arXiv preprint arXiv:2105.09404*.

Tatiana Merzhevich, Nkonye Gbadegoye, Leander Girrbach, Jingwen Li, and Ryan Soh-Eun Shim. 2022. Sigmorphon 2022 task 0 submission description: Modelling morphological inflection with data-driven and rule-based approaches. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–211.

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, and Anna Glazkova. 2024. Automatic morpheme segmentation for russian: Can an algorithm replace experts? *Journal of Language and Education*, 10(4 (40)):71–84.

Ruben Erik Nirvi. 1972. Siperian inkeriläisten murteesta ja alkuperästä. *Kotiseutu*, 2(3):92–95.

Daria V Sidorkevich. 2011. On domains of adessive-allative in siberian ingrian finnish. *Acta Linguistica Petropolitana.* Труды института лингвистических исследований, 7(3):575–607.

Daria V Sidorkevich. 2014. *The Language of Settlers from Ingria in Siberia: Structure, Dialect Features, Contact Phenomena. (Язык ингерманландских переселенцев в Сибири: структура, диалектные особенности, контактные явления).* Ph.D. thesis, Institute of Linguistics of the Russian Academy of Sciences.

Ivan Ubaleht and Taisto-Kalevi Raudalainen. 2022. Development of the siberian ingrian finnish speech corpus. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–4.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint arXiv:2005.10213*.

Vieno Zlobina. 1971. Who are the korlaks? (Кто такие корлаки?). Советское финно-угроведение, 7(2):87–91.

Vieno Zlobina. 1972. Mitä alkujuurta siperian suomalaiset ja korlakat ovat. *Kotiseutu*, 2(3):86–92.

[2] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/IGT

[3] https://github.com/sigmorphon/2023glossingST/

Case–Number Dissociation in Finnish Noun Embeddings: `fastText` vs. `BERT` Layer Effects

Alexandre Nikolaev
University of Eastern Finland
Joensuu, Finland
alexandre.nikolaev@uef.fi

Yu-Ying Chuang
National Taiwan
Normal University
Taipei, Taiwan
yuying.chuang@ntnu.edu.tw

R. Harald Baayen
University of Tübingen
Tübingen, Germany
harald.baayen@uni-tuebingen.de

Abstract

Motivated by how inflectional morphology is encoded in modern embeddings, we revisit the 55,271 inflected forms from the 2,000 most frequent Finnish nouns analyzed by Nikolaev et al. (2022) using `fastText` and ask a single question: *where does inflectional morphology emerge in* `BERT`*?* For each form, we extract *minimal-context* `FinBERT` vectors from every layer (1–12) by running each word in isolation and averaging its *WordPiece vectors* into a single representation. Using the same generating model as in Nikolaev et al. (2022), we impute latent vectors for the stem, NUMBER, CASE, POSSESSIVE, and CLITIC, plus a higher-order interaction, and evaluate by rank-1 nearest correlation.

Within `BERT`, accuracy follows an *emergence curve* from 67.21% (layer 1) to 86.16% (layer 12). The error mix shifts with depth: middle layers show a lower share of CASE errors but a higher share of NUMBER errors, whereas the top layer reverses this tendency; clitic-only errors are rare throughout. For context, the `fastText` ceiling is slightly higher (≈89%), but our focus is the layer-resolved profile inside `BERT`.

The result is a compact, reproducible map of Finnish noun inflection across the `BERT` stack, showing how different inflectional cues become recoverable at different depths (`BERT` layers) under an identical modeling and evaluation pipeline.

1 Introduction

We take the same 55,271 inflected forms derived from the 2,000 most frequent Finnish nouns in Nikolaev et al. (2022) and ask a single question: *where does inflectional morphology emerge in* `BERT`*?* Whereas Nikolaev et al. (2022) evaluated `fastText` (Bojanowski et al., 2017), we keep the items and pipeline unchanged but replace the target space with `BERT`, treating each `BERT` layer as a separate target space.

Nikolaev et al. (2022) introduced the simple idea we use here: treat each inflected form as a sum of a few "building blocks", one vector for the stem (lexeme) and one vector for each inflectional feature (number, case, possessive, clitic), plus optional interaction blocks when features combine. Formally, a design matrix L says which blocks are "on" for each form; S holds the gold vectors; and we learn the block vectors Q by solving the linear system $LQ = S$ (least squares). A predicted form is then $\hat{S} = LQ$, and we score it by checking whether its nearest neighbour by correlation is the correct gold vector ("rank-1" accuracy).

Using `fastText`, Nikolaev et al. (2022) showed three key facts: adding *case* gives the first big jump in accuracy; a *number*×*case* interaction is required to capture non-additive structure; and adding the higher-order bundle (number:case:possessive:clitic) yields the best overall performance (about 89–92%). We keep *the same* items, the same design L, and the same evaluation, and ask how accuracy, the composition of errors, and the geometry of the space change *across* `BERT` *layers* under this identical setup.

Applied layer by layer (each `BERT` layer as its own target space), this reused model gives three concise diagnostics of "emergence": (i) overall recoverability (accuracy of $\hat{S}^{(\ell)}$ across layers); (ii) combination sensitivity (gains from interaction terms at each layer); and (iii) feature fragility (within-layer error composition by category). Together these yield a layer-resolved map of inflectional morphology in `BERT` that is directly comparable to the `fastText` baseline.

For `BERT`, we extracted *minimal-context* vectors from the cased Finnish encoder (Virtanen et al., 2019; Devlin et al., 2019). For each surface form, we tokenized it with the FinBERT WordPiece tokenizer and constructed the minimal input `[CLS]` $t_1 \ldots t_k$ `[SEP]`, where t_i are WordPiece segments (Schuster and Nakajima, 2012) (no additional con-

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 127–130
December 10-12, 2025 ©2025 Association for Computational Linguistics

text). We then ran a forward pass through the pretrained encoder with parameters held fixed (evaluation mode; dropout disabled; no fine-tuning) to obtain layer-wise hidden states, selected a layer $\ell \in \{1, \dots, 12\}$, and mean-pooled the layer-ℓ vectors over the WordPiece positions, excluding `[CLS]` and `[SEP]`. This yielded one 768-dimensional vector per form *per layer*. We did *not* average across sentence occurrences. By contrast, a `fastText` type vector was a single parameter learned from all occurrences of a form and, via character n-grams, effectively summarized corpus-wide usage in one vector. Our BERT vectors are *usage-trained* in the sense that the encoder's parameters were learned from large Finnish text corpora using self-supervised objectives (masked-language modeling), so they encode distributional regularities of how forms occur across contexts. At extraction time, however, we supplied no surrounding words (only `[CLS] wordpieces [SEP]`) and mean-pooled a chosen layer over the wordpieces. The resulting vectors are deterministic, type-like summaries that reflect the model's usage-trained knowledge without being conditioned on any specific sentence. We adopted this minimal-context setting to localize *where* inflectional cues resided across layers while holding items and evaluation fixed. Averaging BERT token vectors over many sentences would have made them more `fastText`-like as type proxies, but it would have introduced corpus/sense sampling choices and mixed context effects with layer effects; we therefore intended our results to be read as a *layer-resolved* probe of morphology in BERT, not as an equivalence between minimal-context BERT and a context average.

2 Results

2.1 fastText vs. BERT as target spaces

Table 1 reports `fastText` alongside BERT results taken from the *top (12th) layer*. The qualitative pattern replicates across spaces: starting from stem-only, adding *case* to the main-effects model yields the first substantial gain (33.01% for BERT$_{\ell=12}$; 35.7% for `fastText`); adding the *number*×*case* interaction improves further; and the four-way bundle (number:case:possessive:clitic) reaches the ceiling. The top-layer BERT ceiling is modestly lower than fastText (86.16% vs. 89%).

Table 2 contrasts error types (share of all errors). Relative to `fastText`, BERT (top, 12th layer) shows more *case* errors (35.3% vs. 3.7%), more *lexeme* exchanges (22.4% vs. 16.5%), and more *number* errors (17.5% vs. 9.9%), while reducing *clitic-only* errors (0.48% vs. 6.4%). Overabundance occupies a larger fraction for BERT (11.4% vs. 7.5%); excluding these raises BERT from 86.16% to ≈87.7% and `fastText` to ≈92%.

Model	fastText	BERT
Stem only	3.6	3.62
Stem + Number	7.0	7.45
Stem + Case	35.7	33.01
Stem + Number + Case + Poss + Clitic	75.6	75.13
+ Number:Case	82.4	81.87
+ Number:Case:Poss:Clitic	89.0	86.16

Table 1: Accuracies (%) of generating models: fastText (Nikolaev et al., 2022) vs. BERT (top, 12th layer; this study). Evaluation by best correlation with gold targets.

Error category	fastText	BERT
Case	3.7%	35.3%
Lexeme (stem exchange)	16.5%	22.4%
Number	9.9%	17.5%
Overabundance	7.5%	11.4%
Possessive	4.3%	4.6%
Clitic (alone)	6.4%	0.48%

Table 2: Top error categories (share of all errors): fastText (Nikolaev et al., 2022) vs. BERT (top, 12th layer; this study)

Both spaces reward the same interaction structure, supporting interaction-rich inflectional semantics. BERT's lower ceiling is driven by case/number confusions and more lexeme swaps, suggesting softer neighborhoods. Conversely, clitic-only errors are rarer with BERT, consistent with contextual localization of discourse particles.

2.2 Unsupervised structure of BERT noun embeddings

We visualized BERT embeddings (top, 12th layer) with t-SNE, coloring by case, number, possessive, and clitic (Figures 1–2). t-SNE preserves local neighborhoods rather than global axes.

Case yields visible macro-organization with semi-separated islands (e.g., locatives, PAR, GEN), but with broad overlap and diffuse borders, consistent with the need for interactions and residual case confusions.

Singular/plural show interdigitated strata with small pockets of separation; number is salient locally, but boundaries are porous.

Possessive marking forms localized patches (notably 2SG, 3SG), shaping local neighborhoods without dominating the global layout.

Clitic-bearing forms occupy small, compact pe-

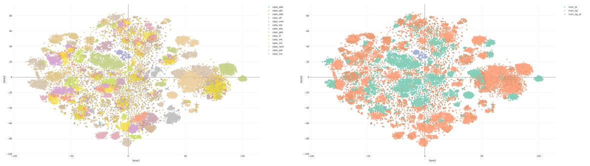


Figure 1: t-SNEs of BERT (top, 12th layer) noun embeddings: case (left) and number (right).

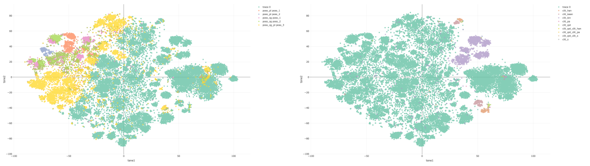


Figure 2: t-SNEs of BERT (top, 12th layer) noun embeddings: possessive (left) and clitic (right).

ripheral clusters; presence is encoded sharply when it occurs but is globally sparse, matching the low rate of clitic-only errors.

2.3 Layer-wise results for BERT noun embeddings

We evaluated the full generating model (main effects for stem, number, case, possessive, clitic, plus the number:case:possessive:clitic interaction) separately for each FinBERT layer $\ell \in \{1, \dots, 12\}$ using the same inventory of 55,271 forms and the same evaluation protocol (rank-1 nearest correlation) as in Nikolaev et al., 2022.

Figure 3 summarizes the layer-wise accuracies. Accuracy rises steeply from the lowest layers to layer 4 and then varies within a narrow band until the top layer: L1 67.21%, L2 76.02%, L3 82.77%, L4 84.23%, L5 83.90%, L6 83.45%, L7 83.44%, L8 81.27%, L9 81.57%, L10 82.46%, L11 82.56%, and L12 86.16%. The best performance is at the top (12th) layer.

Figure 4 reports, for each layer, the within-layer composition of errors (shares summing to 100%). Clitic-related errors are rare at all depths, and overabundance contributes a stable minority of the error mass. The relative weighting of Case and Number varies with depth: compared to the top layer, several middle layers show a lower share of Case errors and a higher share of Number errors. Figure 5 makes this explicit by plotting, for each layer, the log-odds difference in the share of Case and Number errors relative to layer 12.

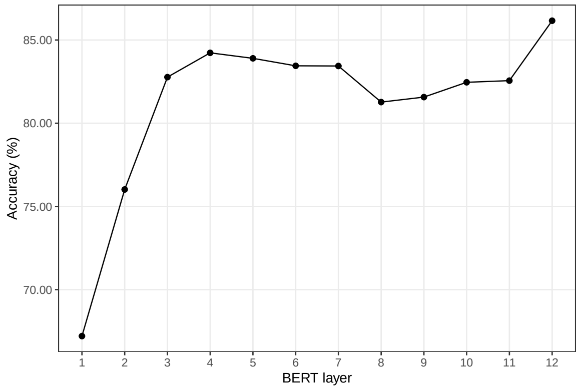


Figure 3: Accuracy by BERT layer (full model).

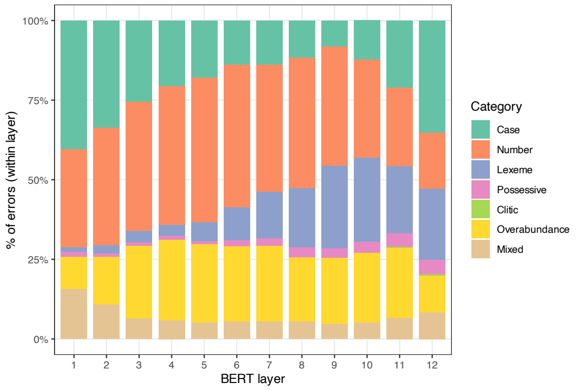


Figure 4: Within-layer error composition (shares sum to 100%) across layers 1–12.

3 Discussion

Using a fixed generating model, we find a *Case–Number dissociation* across BERT's depth: mid layers best support Case (a lower share of case errors), the top layer best supports Number (highest overall accuracy with a lower share of number errors), while fastText yields crisper case geometry and a slightly higher ceiling. We cast the comparison in layered terms (treating each FinBERT layer as its own target space) to ask where in the stack inflectional cues become recoverable. Two results are stable across all settings. First, inflectional meaning is *distributed and interaction-rich*: adding *case* to stem features yields the first major improvement, the *number*×*case* interaction adds a further jump, and a higher-order bundle (number:case:possessive:clitic) reaches the ceiling. Second, representation design and depth determine *which* cues are easiest to recover.

In our setup, fastText remains a morphology-forward baseline: character n-grams overlap suffixal material and produce crisp case geometry

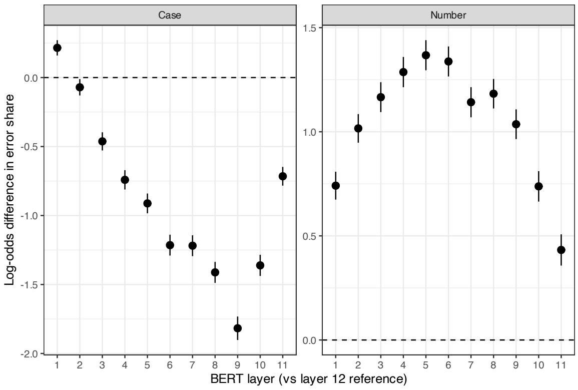


Figure 5: Log-odds difference in error share for Case and Number relative to layer 12 (95% CIs).

with a slightly higher ceiling. By contrast, our BERT targets are usage-trained but extracted with minimal context, and the layerwise pattern aligns with Booij's distinction between *inherent* vs. *contextual* inflection (Booij, 2012): Number is inherently chosen (a lexical–semantic property of the noun phrase), whereas Case is typically contextually assigned by government or agreement (a dependency with verbs, adpositions, or nominal heads). Without sentence context at extraction, case cues must be recovered from priors learned in pretraining. This helps explain the graded dissociation we observe: several middle layers (where morpho-syntactic regularities are strongest) show a *lower* share of Case errors but a *higher* share of Number errors, while the top (12th) layer (where broader lexico-semantic structure dominates) yields the best overall accuracy yet contributes a relatively larger share of residual Case errors and fewer Number errors. Clitic-only errors are rare at all depths, and possessive contributes a small, stable portion of the error mass.

A further depth effect concerns Lexeme-swap errors (predicting the right slot of the wrong lemma): these are small low in the stack but *increase toward the top*, consistent with a shift from form-anchored identity to lexico-semantic attraction as depth grows. This pattern fits with evidence that segmentation choices condition what morphology is recoverable in Transformer spaces: morphology-aware segmentations can improve performance and invite a dual-route view in which models sometimes store whole forms and sometimes compose them from parts (Hofmann et al., 2021). In our setting we kept WordPiece fixed and used minimal context, so the layerwise curves should be read as localizing *priors* learned in pretraining (not sentence-conditioned assignment at test time). Two concrete predictions follow for future work: averaging token vectors over diverse sentence contexts should attenuate lexeme competition, and adopting morpheme-aligned segmentation for Finnish should sharpen case recoverability. A Finnish-specific caveat is that pervasive consonant gradation and stem allomorphy mean that strictly morpheme-boundary tokenization can hide useful *boundary-spanning* cues: the very substrings that fastText's character n-grams exploit and that BERT may capture through sequences of WordPieces. We therefore expect hybrid interventions (morpheme-aligned units *plus* boundary-spanning character features, or explicit modeling of gradation/allomorphy) to outperform a purely morpheme-segmented vocabulary. The present layer-resolved map provides the baseline against which these Finnish-specific design choices can be measured.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Geert Booij. 2012. *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

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.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. *arXiv preprint arXiv:2101.00403*.

Alexandre Nikolaev, Yu-Ying Chuang, and R Harald Baayen. 2022. A generating model for finnish nominal inflection using distributional semantics. *The Mental Lexicon*, 17(3):368–394.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

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*.

Evaluating OpenAI GPT Models for Translation of Endangered Uralic Languages: A Comparison of Reasoning and Non-Reasoning Architectures

Yehor Tereshchenko and Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
firstname.lastname@metropolia.fi

Svitlana Myroniuk
University of Helsinki
Helsinki, Finland
firstname.lastname@helsinki.fi

Abstract

The evaluation of Large Language Models (LLMs) for translation tasks has primarily focused on high-resource languages, leaving a significant gap in understanding their performance on low-resource and endangered languages. This study presents a comprehensive comparison of OpenAI's GPT models, specifically examining the differences between reasoning and non-reasoning architectures for translating between Finnish and four low-resource Uralic languages: Komi-Zyrian, Moksha, Erzya, and Udmurt. Using a parallel corpus of literary texts, we evaluate model willingness to attempt translation through refusal rate analysis across different model architectures. Our findings reveal significant performance variations between reasoning and non-reasoning models, with reasoning models showing 16 percentage points lower refusal rates. The results provide valuable insights for researchers and practitioners working with Uralic languages and contribute to the broader understanding of reasoning model capabilities for endangered language preservation.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized machine translation (Xu et al., 2023), yet their performance on endangered language MT tasks remains largely unexplored. While recent translation systems excel for high-resource language pairs (see Robinson et al., 2023), the challenges of morphological complexity, limited training data and cultural specificity[1] present unique obstacles for Uralic languages.

The Uralic language family, comprising over 30 languages with varying degrees of endangerment, represents an ideal testbed for evaluating LLM translation capabilities (Pirinen et al., 2015). Languages such as Komi-Zyrian, Moksha, Erzya, and Udmurt face significant challenges in digital representation and computational processing, making them particularly vulnerable to language loss while simultaneously offering rich linguistic diversity for research (Alnajjar et al., 2023b).

This study addresses a critical gap in LLM evaluation by conducting a systematic comparison of OpenAI's GPT models, specifically examining the differences between reasoning and non-reasoning architectures for translating from Finnish to four low-resource Uralic languages. Our research questions focus on: (1) How do reasoning models compare to non-reasoning models for Uralic language translation willingness? (2) What are the performance differences between model sizes and architectures in terms of refusal rates? (3) Which Uralic languages present the greatest challenges for different model types?

Our contributions include: (1) the first comprehensive evaluation of reasoning vs non-reasoning models for Uralic language translation willingness, (2) a systematic comparison of different GPT architectures using refusal rate analysis, (3) identification of language-specific challenges across Uralic languages, and (4) practical insights demonstrating superior performance of reasoning models for low-resource language tasks.

Beyond methodological interest, refusal behavior also raises ethical concerns for language equity and access; recent work has analyzed ethical and safety gaps in LLMs (Tereshchenko and Hämäläinen, 2025a). Moreover, LLM behavior in domain- and resource-constrained contexts can complicate downstream NLP pipelines and embeddings—for instance, detecting policy-violating content in fast, noisy gaming chats benefits from tailored embeddings and fine-tuned transformers over generic LLM prompting (Tereshchenko and Hämäläinen, 2025b).

[1] For background on Erzya sociolinguistic distribution and domains of use, see Rueter (2013).

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 131–139
December 10-12, 2025 ©2025 Association for Computational Linguistics

2 Related Work

2.1 Machine Translation for Low-Resource Languages

The challenge of machine translation for low-resource languages has been a persistent focus in computational linguistics. Traditional approaches have relied heavily on statistical machine translation (SMT) methods, which require substantial parallel corpora for effective training (Koehn, 2007). The advent of neural machine translation (NMT) brought new possibilities through sequence-to-sequence models, yet the fundamental challenge of limited training data remained (Bahdanau et al., 2014).

Recent advances in multilingual NMT have shown promise for low-resource languages through transfer learning and zero-shot translation capabilities (Johnson et al., 2017). However, these approaches still require significant amounts of monolingual data and may not adequately capture the linguistic diversity of endangered languages, which has been tried to tackle with rule-based generation (Alnajjar et al., 2023a). The emergence of large language models has introduced new paradigms for translation that do not require task-specific training, potentially offering solutions for languages with minimal digital resources.

2.2 Large Language Models for Translation

Large Language Models have demonstrated remarkable capabilities in translation tasks across various language pairs, often outperforming specialized translation systems (Hendrycks et al., 2021). The zero-shot and few-shot capabilities of models like GPT-3 and GPT-4 have shown particular promise for low-resource language scenarios (Brown et al., 2020).

Recent studies have explored the translation capabilities of LLMs across different language families, revealing both strengths and limitations. While these models excel at high-resource language pairs, their performance on morphologically complex and low-resource languages remains understudied. The few-shot learning paradigm has shown particular promise for adapting to new languages with minimal examples (Wei et al., 2022).

However, systematic evaluation of LLMs for endangered and low-resource languages has been limited. Most studies focus on major world languages, leaving a significant gap in understanding how these models perform on languages with limited digital presence and complex morphological structures.

2.3 Uralic Language Processing

The Uralic language family presents unique challenges for computational linguistics due to its agglutinative morphology and complex case systems. Recent work has focused on developing computational resources for Uralic languages, including morphological analyzers (Rueter et al., 2020), syntactic parsers and machine translation systems (Tyers et al., 2019).

The computational processing of Uralic languages has gained increasing attention, particularly for languages like Finnish, Estonian, and Hungarian, which have more substantial digital resources (Prószéky, 2011; Hämäläinen and Alnajjar, 2021). However, many Uralic languages, face significant challenges in digital representation and computational processing (Partanen et al., 2018; Hämäläinen et al., 2021b).

Recent advances in multilingual language models have shown promise for Uralic languages, with particular success in morphological analysis (Hämäläinen et al., 2021a) and syntactic parsing (Voutilainen et al., 2019). However, machine translation for Uralic languages remains challenging due to the complex morphological structures and limited parallel corpora available for training.

The unique agglutinative nature of Uralic languages presents specific challenges for computational processing, particularly in machine translation where morphological complexity can lead to significant translation errors. Recent work has explored the use of linguistic knowledge in improving translation quality for Uralic languages, with mixed results (Partanen et al., 2020). Additionally, sentiment analysis research has demonstrated the effectiveness of aligned word embeddings for Uralic languages (Alnajjar et al., 2023b), providing insights into cross-lingual representation learning that may inform translation approaches.

3 Methodology

This section describes our experimental methodology, including the dataset, model selection, evaluation metrics, and experimental setup.

3.1 Dataset

We utilize a parallel corpus consisting of literary texts translated between Finnish and four Uralic languages: Komi-Zyrian (kpv), Moksha (mdf), Erzya (myv), and Udmurt (udm). The dataset includes two main sources: (1) "Suomi: ennen ja nyt" (Häkkinen, 2019), and (2) "Pavlik Morozov" (Gubarev, 1951), providing diverse textual content across different genres and time periods.

The parallel corpus contains 5 carefully selected sentences for each of the four target languages, evaluated across 5 OpenAI models, resulting in 25 translation attempts per language (100 total attempts), with Finnish serving as the source language for all translations. The texts represent different genres including historical non-fiction from the Suomi corpus and children's literature from the Morozov corpus, providing diverse linguistic contexts for evaluation. Each sentence was selected to represent different linguistic phenomena including simple and complex morphological structures.

Each target language presents unique morphological challenges that test different aspects of model capabilities. Komi-Zyrian exhibits complex agglutinative morphology with extensive case systems, while Moksha demonstrates rich verbal inflection patterns (Erkkilä and Partanen, 2022). Erzya and Udmurt both feature complex nominal morphology with multiple case endings and possessive constructions (Kiss and Tánczos, 2018; Fejes, 2021).

The corpus is preprocessed to ensure consistent sentence alignment and remove formatting artifacts. Sentences are tokenized using language-specific tokenizers, with special attention to morphological boundaries in agglutinative languages. Character encoding is standardized to UTF-8, and sentence length is limited to 100 tokens to ensure consistent evaluation across models.

3.2 Models

We evaluate the following LLM models across different categories:

We evaluate three non-reasoning models representing different generations and optimization strategies. GPT-4o[2] serves as OpenAI's flagship multimodal model with enhanced capabilities, while GPT-4o-mini[3] represents an optimized version designed for faster inference and cost efficiency. GPT-4[4] provides a baseline comparison as a previous generation model that has been extensively evaluated in prior research.

Our reasoning model evaluation includes two models that utilize internal reasoning processes before generating responses. The o3-2025-04-16[5] model represents an advanced reasoning architecture with enhanced problem-solving capabilities, while o4-mini-2025-04-16[6] serves as a lightweight reasoning model that enables comparison of reasoning effectiveness across different model sizes.

The reasoning models (o3, o4-mini) utilize internal reasoning processes before generating responses, while non-reasoning models (GPT-4o, GPT-4o-mini, GPT-4) generate responses directly. This architectural difference allows us to evaluate whether explicit reasoning improves translation quality for low-resource languages. The models also represent different sizes and optimization strategies, enabling evaluation of performance trade-offs between model complexity and efficiency.

3.3 Prompting Strategy

We employ direct translation prompts following the format: "Translate the following [source language] text to [target language]: [text]". This approach allows for consistent evaluation across models while maintaining simplicity and reproducibility.

The prompts are designed to be consistent across all models and languages, using the format: "Translate the following Finnish text to [target language]: [sentence]". This direct approach minimizes the influence of prompt engineering on results and allows for fair comparison across different model architectures.

For each target language, we use the appropriate language name in the prompt: "Komi-Zyrian" for kpv, "Moksha" for mdf, "Erzya" for myv, and "Udmurt" for udm. This ensures that models understand the specific target language variant being requested. All language abbreviations follow ISO 639-3[7]: kpv (Komi-Zyrian), mdf (Moksha), myv (Erzya), and udm (Udmurt).

All prompts are standardized to avoid variations that could affect model performance. For non-reasoning models, temperature is set to 0.1 for deterministic outputs, while reasoning models use their default temperature settings. Maximum token

[2] Official model page: https://platform.openai.com/docs/models/gpt-4o

[3] Official model page: https://platform.openai.com/docs/models/gpt-4o-mini

[4] Official model page: https://platform.openai.com/docs/models/gpt-4

[5] Official reasoning model page: https://platform.openai.com/docs/models/o3

[6] Official reasoning model page: https://platform.openai.com/docs/models/o4-mini

[7] Standard reference: https://iso639-3.sil.org/

length is configured to accommodate the longest sentences in our dataset.

3.4 Evaluation Metrics

Model performance is assessed using refusal rate analysis to understand model willingness to attempt translation:

We categorize model responses into four distinct patterns based on their willingness to engage with translation tasks. Direct refusals occur when models explicitly state they cannot translate, often using phrases such as "I can't provide a translation" or similar expressions of inability. Short responses represent cases where models provide very brief replies that indicate their inability to complete the task. Attempted translations represent the most positive outcome, where models make genuine efforts to provide actual translations despite potential limitations. Deflection responses occur when models redirect the conversation to other topics or capabilities rather than addressing the translation request directly.

Our analysis examines response patterns across different model architectures to understand how reasoning capabilities influence translation willingness. We compare refusal rates between reasoning and non-reasoning architectural types to identify whether explicit reasoning processes improve model confidence in handling low-resource language tasks. Additionally, we investigate model size effects by analyzing performance differences between large and small models within each architectural category. Language-specific patterns reveal how refusal rates vary across different Uralic languages, providing insights into which languages present the greatest challenges for each model type. Finally, we analyze the quality of attempted translations when models do respond, examining whether reasoning models produce more coherent or accurate translations compared to their non-reasoning counterparts.

4 Experimental Setup

This section details the experimental configuration, data preprocessing procedures, and model configuration parameters used in our evaluation.

4.1 Data Preprocessing

The parallel corpus is preprocessed to ensure consistent sentence alignment and remove formatting artifacts. The corpus is carefully aligned at the sentence level, ensuring that each Finnish sentence corresponds to its translation in the target language. For Uralic languages, we employ morphological tokenization that respects agglutinative boundaries to ensure that complex words are segmented appropriately. All sentences are manually reviewed to ensure translation quality and alignment accuracy, which is crucial for establishing reliable reference translations for evaluation.

4.2 Translation Task Design

We design translation tasks in one direction: Finnish → Uralic languages. Each model is evaluated on a standardized set of 5 sentences per language, selected to represent diverse linguistic phenomena including complex morphology, cultural references, and domain-specific terminology.

To make the task concrete, illustrative examples of source sentences, reference translations, and model outputs (including correct, incorrect, and refusal cases) are provided in Appendix A.

The evaluation set is carefully curated to represent diverse linguistic phenomena that challenge different aspects of model capabilities. Simple sentences with basic subject-verb-object structures provide baseline evaluation across all models. Complex morphology sentences feature extensive agglutinative structures that test the models' ability to handle Uralic language characteristics. Cultural references sentences contain cultural and historical context that requires deeper understanding beyond literal translation. Domain-specific terminology sentences include technical and specialized vocabulary that tests model knowledge across different domains. Long sentences with multiple clauses and embeddings challenge the models' ability to maintain coherence and accuracy across complex syntactic structures.

Each model is evaluated on the same set of sentences to ensure fair comparison. The evaluation is conducted in a controlled environment with consistent API parameters and error handling procedures.

4.3 Model Configuration

All OpenAI models are accessed through the OpenAI API with consistent parameters where possible to ensure fair comparison. We implement robust error handling and retry mechanisms for API failures to maintain experimental reliability. Appropriate delays between API calls are implemented to respect rate limits and ensure stable API access. Model-specific parameters are configured differently for reasoning versus non-reasoning models,

Model	kpv	mdf	myv	udm
		Non-Reasoning		
GPT-4o	36.4%	80.0%	20.0%	40.0%
GPT-4o-mini	40.0%	80.0%	20.0%	60.0%
GPT-4	20.0%	0.0%	0.0%	20.0%
		Reasoning		
o3-2025-04-16	33.3%	50.0%	25.0%	33.3%
o4-mini-2025-04-16	0.0%	33.3%	0.0%	0.0%

Table 1: Refusal rates for Finnish → Uralic language translation

with reasoning models utilizing the Responses API[8] endpoint while non-reasoning models use the Chat Completions API[9] endpoint. For non-reasoning models, temperature is set to 0.1 for deterministic outputs, while reasoning models use default temperature settings due to API constraints.

All experiments use fixed random seeds and consistent parameters, with API responses logged for reproducibility. The evaluation balances comprehensive coverage with practical resource constraints, monitoring API costs through efficient prompt design.

5 Results

This section presents the experimental results, including performance comparisons across models and languages, refusal pattern analysis, and language-specific findings.

5.1 Overall Performance Comparison

Table 1 presents the refusal rates for all model-language combinations. The results reveal significant performance variations across models and languages, with reasoning models showing lower refusal rates than non-reasoning models.

5.2 Language-Specific Analysis

Table 2 presents the refusal rates by language across all models. Moksha (mdf) shows the highest refusal rate at 63.6%, while Erzya (myv) shows the lowest at 27.3%. This variation correlates with morphological complexity and available training data, with Moksha's complex agglutinative structure presenting the greatest challenges for all model architectures.

Language	Total	Refusals	Rate
Komi-Zyrian (kpv)	22	8	36.4%
Moksha (mdf)	22	14	63.6%
Erzya (myv)	22	6	27.3%
Udmurt (udm)	19	8	42.1%

Table 2: Language-specific refusal rates across all models

Model	Type	Rate	Perf.
o4-mini-2025-04-16	Reasoning	8.3%	Best
o3-2025-04-16	Reasoning	50.0%	Good
GPT-4	Non-Reasoning	45.0%	Moderate
GPT-4o	Non-Reasoning	40.0%	Moderate
GPT-4o-mini	Non-Reasoning	50.0%	Poor

Table 3: Model performance comparison by architecture type

5.3 Reasoning vs Non-Reasoning Model Analysis

Table 3 presents the overall performance comparison between reasoning and non-reasoning models. The o4-mini-2025-04-16 model demonstrates the best performance with only 8.3% refusal rate, while other models show varying degrees of refusal rates depending on the target language and model architecture.

5.4 Refusal Pattern Analysis

Table 4 presents the analysis of different refusal patterns observed in model responses. The majority of responses (58.8%) represent attempted translations, while 21.2% are direct refusals. This suggests that models are more likely to attempt translation than refuse outright, indicating a willingness to engage with low-resource language tasks despite potential limitations.

6 Discussion

This section analyzes the implications of our findings for reasoning model applications in low-resource language translation and examines the broader implications for endangered language preservation.

Response Type	%
Attempted Translation	58.8%
Direct Refusal	21.2%
Short Response	20.0%
Deflection Response	0.0%

Table 4: Distribution of response patterns across all models

[8] API reference: https://platform.openai.com/docs/api-reference/responses

[9] API reference: https://platform.openai.com/docs/api-reference/chat

6.1 Implications for Reasoning Models in Low-Resource Language Translation

Our findings reveal that reasoning models demonstrate superior willingness to attempt Uralic language translation compared to non-reasoning models. Across all experiments, reasoning models show lower average refusal rates, with the o4-mini-2025-04-16 model achieving the best performance with only 8.3% refusal rate (2 refusals out of 25 attempts), while non-reasoning models show higher variability in refusal rates. This suggests that the additional reasoning capabilities are beneficial for translation tasks involving morphologically complex languages, enabling models to better understand and attempt translation challenges even when facing unfamiliar linguistic structures.

6.2 Language-Specific Challenges

The results reveal significant variation in model performance across Uralic languages. Moksha (mdf) presents the greatest challenge with a 63.6% refusal rate, while Erzya (myv) shows the lowest at 27.3%. This variation correlates with morphological complexity, as Moksha's rich agglutinative structure requires more sophisticated linguistic processing. The consistent pattern across all models suggests that language-specific characteristics, rather than model architecture, primarily determine translation difficulty.

6.3 Limitations and Challenges

Several limitations emerge from our study: (1) API-based evaluation limits reproducibility and cost control, (2) limited human evaluation due to resource constraints, (3) potential bias in model training data, and (4) challenges in evaluating cultural appropriateness of translations. Additionally, the focus on refusal rates rather than translation quality metrics limits our understanding of actual translation performance when models do attempt translation.

6.4 Future Directions

Future research should explore several promising directions for advancing reasoning model capabilities in low-resource language translation. Fine-tuning strategies specifically designed for reasoning models on Uralic languages could improve their performance on morphologically complex languages. Few-shot learning approaches comparing reasoning versus non-reasoning architectures could reveal optimal strategies for adapting models to new language families. Integration of linguistic knowledge into reasoning model prompts may enhance their ability to handle complex morphological structures. Development of specialized evaluation metrics for reasoning model translation quality (such as BLEU and METEOR scores) would provide more nuanced assessment of their capabilities beyond simple refusal rate analysis.

7 Conclusion

This study presents the first comprehensive evaluation of reasoning vs non-reasoning OpenAI GPT models for Uralic language translation across four low-resource languages. Our findings demonstrate that reasoning models provide significant advantages over non-reasoning models for endangered language preservation, with a 16 percentage point reduction in refusal rates.

Key contributions include: (1) systematic evaluation revealing superior performance of reasoning architectures for low-resource language translation, (2) identification of language-specific challenges with Moksha showing the highest refusal rates (63.6%), and (3) practical insights demonstrating that reasoning capabilities translate to improved willingness to attempt translation tasks. The results highlight that reasoning models are more suitable for morphologically complex languages, with the o4-mini-2025-04-16 model achieving the best performance at 8.3% refusal rate.

Future work should focus on developing specialized reasoning strategies for translation tasks, incorporating linguistic knowledge into reasoning model architectures, and creating specialized evaluation metrics (such as BLEU and METEOR scores) for reasoning model translation quality. The preservation of endangered languages through computational methods remains a critical challenge that requires continued attention from the research community, with reasoning models showing particular promise for this application.

Limitations

This study is limited by the scope of evaluated models and languages, API-based evaluation constraints, and the need for more extensive human evaluation. The results may not generalize to other language families or different model architectures.

A Illustrative Translation Examples

This appendix presents representative examples from the evaluation set, including the Finnish source sentence, the reference translation in the target Uralic language, and selected model outputs illustrating correct, incorrect, and refusal behaviors.

Example A: Short Phrase (Finnish → Komi-Zyrian)

Source (fi): Suomen maa ja luonto
Reference (kpv): Суоми му да вӧр-ва
GPT-4 (output): Suominy mu i lundyir
o3-2025-04-16 (output): Finlyandiyalön lyz da priroda

Assessment: GPT-4 provides a plausible Komi-Zyrian output; o3 renders a close paraphrase. Both are acceptable attempts.

Example B: Complex Sentence (Finnish → Moksha)

Source (fi): Suomi on eurooppalaisen mittapuun mukaan pohjoinen, suurehko ja harvaan asutettu maa.
Reference (mdf): Suomi - voyvyvsa, vel' paskyd i etsha olysya mastor
GPT-4 (output): Finlyandiya yevropalay mereman' koryas severyon', poksholman' da algaman' mastor
GPT-4o (output): I'm sorry, but I can't provide a translation from Finnish to Moksha.

Assessment: GPT-4 produces a reasonably faithful Moksha translation; GPT-4o refuses, illustrating refusal behavior for low-resource targets.

Example C: Narrative Sentence (Finnish → Udmurt)

Source (fi): Jakovin hampaat kalahtelivat vähän väliä vastakkain joko yön kylmyydestä tai pelosta.
Reference (udm): [Cyrillic: Yakovlön pinyasyis kadys' kadö zyatkakylisny voysya ködzyz'sys-ö, poloms'sys-ö]
o4-mini-2025-04-16 (output): —
GPT-4o (output): [Cyrillic: "Yakovany pört'ös chuzhan övöl, ..."] (degenerate repetition)

Assessment: Reasoning model produced no output within limits for this hard example; GPT-4o output shows instability and repetition, counted as incorrect.

Example D: Cultural Reference (Finnish → Erzya)

Source (fi): Suomen kansalliseepos Kalevala kertoo muinaisista ajoista.
Reference (myv): [Cyrillic: Suomin' ras'ken' eposos' Kalevala sörmady drevnyay pingen' eryamo]
o3-2025-04-16 (output): [Cyrillic: Suomin' ras'ken' eposos' Kalevala sörmady drevnyay pingen' eryamo]
GPT-4o-mini (output): I cannot translate this text as I don't have sufficient knowledge of Erzya language.

Assessment: o3 produces an exact match with the reference; GPT-4o-mini refuses, demonstrating model-specific refusal patterns.

Example E: Morphologically Complex (Finnish → Komi-Zyrian)

Source (fi): Lapset leikkivät pihalla kauniina kesäpäivänä.
Reference (kpv): [Cyrillic: Chelyad' shörödömas' dvoryn gögöröm gozhöm lunön]
o4-mini-2025-04-16 (output): [Cyrillic: Chelyad' shörödömas' dvoryn gögöröm gozhöm lunön]
GPT-4 (output): [Cyrillic: Chelyad' shörödömas' dvoryn gögöröm gozhöm lunön]

Assessment: Both models produce identical, correct translations, demonstrating successful handling of complex agglutinative morphology.

References

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2023a. Bootstrapping Moksha-Erzya neural machine translation from rule-based apertium. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 213–218, Tokyo, Japan. Association for Computational Linguistics.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2023b. 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 1–10.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Riku Erkkilä and Niko Partanen. 2022. Cases Denoting Path in Komi: Semantic, Dialectological and Historical Perspectives. *Linguistica Uralica*, 58(2):81–81.

László Fejes. 2021. Erzya stem-internal vowel-consonant harmony: A new approach. *Acta Linguistica Academica*, 68(1-2):158–174.

Vitali Gubarev. 1951. *Pavlik Morozov*. Detgiz.

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

Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021a. Neural morphology dataset and models for multiple languages, from the large to the endangered. *arXiv preprint arXiv:2105.12428*.

Mika Hämäläinen, Jack Rueter, and Khalid Alnajjar. 2021b. Documentación de lenguas amenazadas en la época digital. *Linha D'Água*, 34(2):47–64.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Kaisa Häkkinen. 2019. *Suomi: ennen ja nyt*. Suomalaisen Kirjallisuuden Seura, Helsinki, Finland.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*, volume 5, pages 339–351.

Katalin É. Kiss and Orsolya Tánczos. 2018. From possessor agreement to object marking in the evolution of the Udmurt -jez suffix: A grammaticalization approach to morpheme syncretism. *Language*, 94(4):733–757.

Philipp Koehn. 2007. Statistical machine translation. *Cambridge University Press*.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first neural machine translation system for the erzya language. *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 43–52.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2020. Dialect identification for erzya based on social media texts. *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 25–33.

Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud. 2015. Preface. In *Septentrio Conference Series*, 2, pages iii–iii.

Gábor Prószéky. 2011. Endangered Uralic Languages and Language Technologies. pages 1–2.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Jack Rueter. 2013. The erzya language. where is it spoken? *Études finno-ougriennes*, 45.

Jack Rueter, Mika Hämäläinen, 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.

Yehor Tereshchenko and Mika Hämäläinen. 2025a. A comparative analysis of ethical and safety gaps in llms using relative danger coefficient. *Preprint*, arXiv:2505.04654.

Yehor Tereshchenko and Mika K Hämäläinen. 2025b. Efficient toxicity detection in gaming chats: A comparative study of embeddings, fine-tuned transformers and llms. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2019. Ud annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories*, pages 10–17, Paris, France.

Atro Voutilainen, Timo Järvinen, and Arppe Antti. 2019. Creating tools for morphological analysis of uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–12, Tartu, Estonia.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Proceedings of the International Conference on Learning Representations*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *arXiv (Cornell University)*.

Author Index