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;

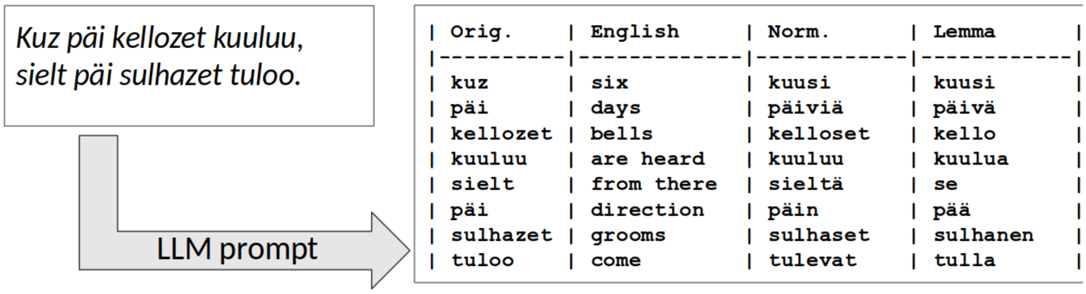

Figure 1: An illustration of our pipeline: on the left is a runosong in Livvi Karelian, on the right the same text processed with an LLM (DeepSeek-R1-BF16 in this case): for each input word, the model returns its English translation, the same word in standard orthography, and lemma. First two words are 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.

```json
{    "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 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

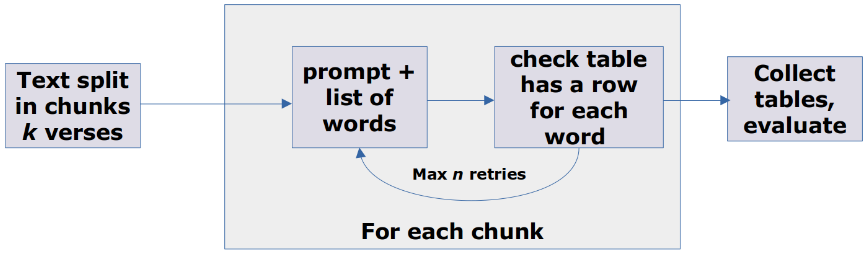

Figure 2: Data processing pipeline.

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-

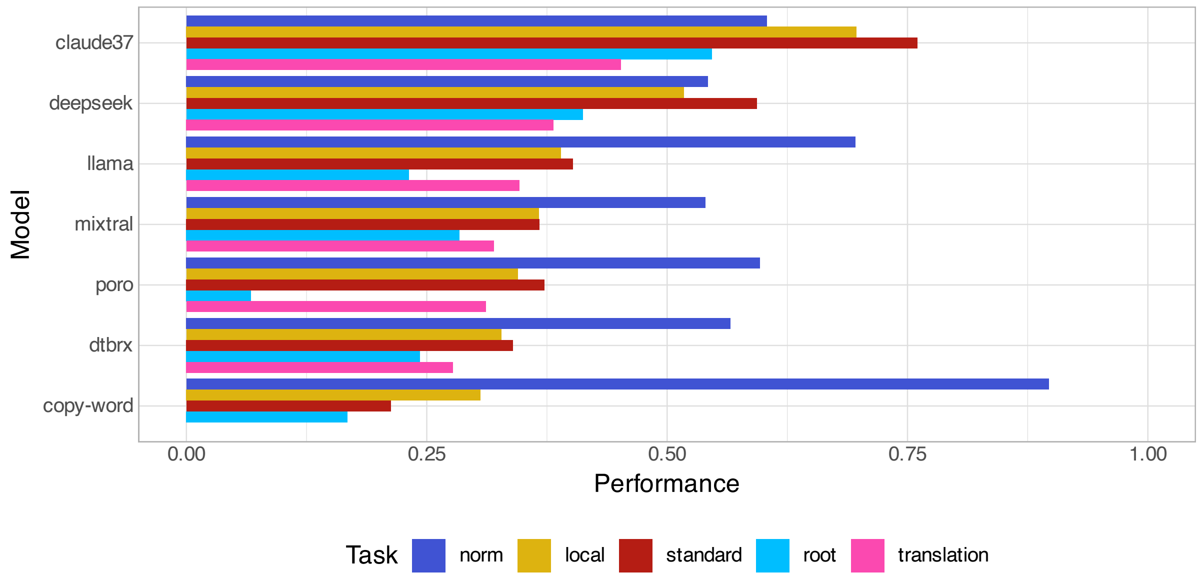

Figure 3: Overall results, averaged across all texts in the collection. We show exact string match for all fields except for English translation, where we show cosine distance between ground truth and model output embeddings. We show the best performing pipeline for each model-task pair.

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

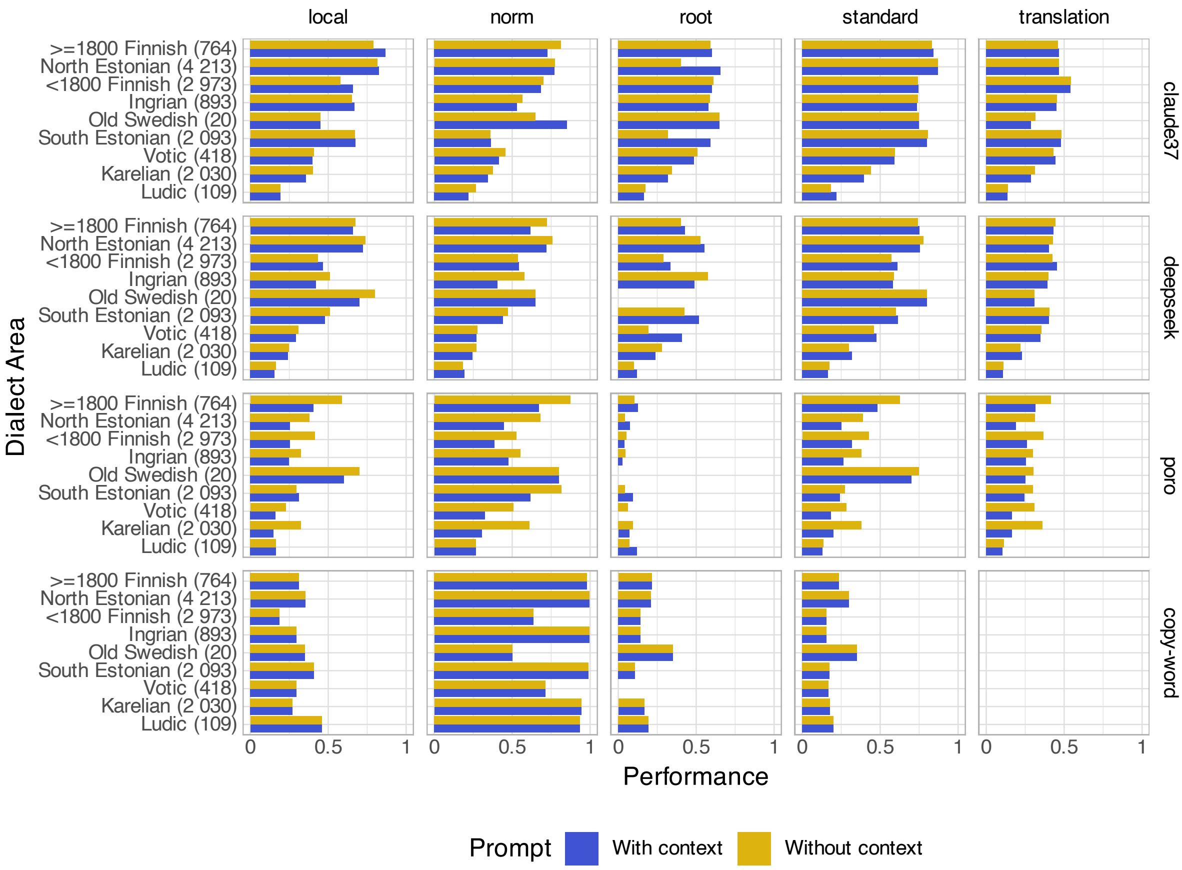

Figure 4: Results for three models and the copy-word baseline grouped by task, language area and whether the prompt includes contextual information. We show the best performing pipeline for each language-context pair. A number of running words in the evaluation set for each language is shown in parenthesis.

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.

Acknowledgments

This research was supported by the IKAKE project funded by the KONE foundation; REFOP project funded by the Research Council of Finland (No 346342); research project PRG1288 funded by the Estonian Research Council; and the Estonian Roots Centre of Excellence (TK215, Estonian Ministry of Education and Research).

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*).