Using Large Language Models to Transliterate Endangered Uralic Languages

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

Abstract

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

1 Introduction

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

The use of orthographies in contemporary language documentation has been seen preferable to phonemic or detailed phonetic transcription (Gerstenberger et al., 2016, 32), and also in this study the primary question is how to transform the transcriptions into currently used orthographies. The use of some other systematic transcription system could be likewise suggested, and it seems reasonable to expect that our approach could be extended to other comparable transformations. The target chosen here, contemporary orthography, comes with its own problems that some other choices would have avoided. There is inevitably more subjectivity in the details of the result than what i.e. phonemic representation would entail.

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

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

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

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 81–88
November 28-29, 2024 ©2024 Association for Computational Linguistics

2 Related work

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

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

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

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

3 Data and Experiment Design

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

The linguistic sources include: Udmurt texts published in Munkácsi (1952), edited by D. R. Fuchs. The Northern Mansi example is published in Kannisto (1956), which is part of a large collection of Mansi texts in several volumes. The Komi-Zyrian example is published in Uotila (1986). These are published by the Finno-Ugrian Society[1]. As stated, the Kildin Saami text differs from the other examples. It was published in Genetz (1879) by a Hungarian publisher, it represents a religious genre and, given its age, it is clearly in the public domain. The latter appears to be true for all of the earliest contributors to these materials.

The materials used in this study have been published in GitHub[2], which allows replicating the results and makes it easy to test the outcome with different preferences and choices. The Table 1 shows the size of the current dataset. See the appendices A and B for example of how the prompts were formatted.

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

[2]https://github.com/nikopartanen/finno-ugric-transliteration-examples

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

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

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

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

Four different models were selected and tested further. These are Claude 3 and 3.5 Sonnet[3], Gemini 1.5 Pro[4] and ChatGPT 4o[5]. We conduct two different experiments, first is a zero-shot scenario where the model is asked to transliterate the text with no additional information. This is not an ideal setting, since as explained, it is not obvious what kind of representation is actually wanted. However, it gives information about the model's capabilities. It can be questioned why a zero-shot experiment is needed, as it is clear that example data should improve the result, but at the same time zero-shot scenario does give valuable information about what the models are capable to do without any further guidance. In the second experiment the same prompt is used, but there is an additional example provided, which illustrates the style and type of transformation that is desired. This example is taken from the same text as the transliterated sample, so that the dialect is certainly the same. The test text is one paragraph and the added example little bit longer.

[3] https://www.anthropic.com/news/claude-3-5-sonnet

[4] https://deepmind.google/technologies/gemini/pro/

[5] https://openai.com/index/hello-gpt-4o/

4 Results

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

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

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

- **Transcription**
  - Komi-Zyrian: me t̮šuži ti̮śa̧ ̮t́śa ok-mi̮ş́-ş́o vited vo̮i̮n e̮t́iked juĺɛ d̮́erevńa p u st u· ń ai̮n ì ź̧ - v a rai̮jonin̮.
  - Mansi: kit i̯ōrn ōlèɣ. i̯ànii̯ᵏχumìtɛ nēŋ, māń ᵏχumìtɛ nētāl nētāl ᵏχȯs ōls, βāt́ ōls, nē βis. nē βis, ūsᵊn mìnȧs, βētrȧ βìnȧ βis.
  - Udmurt: ud-murt ki̮šnojos ni̮l-pi vajondi̮rjazi̮, ni̮lzi̮-pizi̮ t́ś́irkkäm-kä vordkono, tuž kapt́ś́iän vajo.
  - Kildin Saami: A mañña Vavilon' vealhtɵtmužest Ieχonia sḁγi̊j Salafiil;
- **Ground Truth**
  - Komi-Zyrian: Ме чужи тысяча öкмыссё витöд воын öтикöд юльö деревня Пустыняын Изьва районын.

Table 2: Zero-shot results

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

Table 3: Extended prompt results

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

– Mansi: Кит ёрн олэг. Яныг хумитэ нэӈ, мань хумитэ нэтӑл нэтӑл хос олыс, вӑть олыс, нэ вис.

– Udmurt: Удмурт кышноёс нылпи ваён дыръязы, нылзы-пизы чырккем ке вордконо, туж капчиен ваё.

– Kildin Saami: А маӈӈа Вавилон вя̄лл̄тэдтмужэсьт Иехония соа̄гэ Салафӣл;

• **Claude 3.5 Sonnet**

– Komi-Zyrian: Ме чужи тысяча ӧкмысьсё витӧд воын ӧтикӧд июлӧ деревня Пустӧйын Изьва районын.

– Mansi: кит ёрн олэг. яныг хумитэ нэӈ, мань хумитэ нэтӑл нэтӑл хос олс, вӑть олс, нэ вис.

– Udmurt: удмурт кышноос ныл-пи ваён дыръязы, нылзы-пизы чирккем ке вордсконо, туж капчиен ваё.

- Kildin Saami: А маӈӈа Вавилонь веалхтэтмужэст Иехониа сāгый Салафиил;

- **Gemini 1.5 Pro**
  - Komi-Zyrian: Ме тшужы тіся окмисӧ витӧд воин ӧтикед юльӧ деревня пустуняын Изьва районын.
  - Mansi: Кити йӧӈн ӧлёх. Йāнӈх хумитэ нэнь, мань хумитэ нэтāл-нэтāл хӧс ӧлс, вāт ӧлс, нэ вис.
  - Udmurt: Удмурт киӟноёс нилпи ваён-дыръязи, нилзи-пизи чирккам-ка вордсконо, туж капчиян ваё.
  - Kildin Saami: А мāнньа Вавилон веалӽтӧтмужест Иехония сāйий Салафиил;

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

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

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

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

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

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

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

[6] Коми кыв корпус by FU-Lab Team contains over 85 million tokens in October 2024: https://komicorpora.ru

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

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

5 Conclusion

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

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

As far as we have been testing these models in last years, this is the first time an LLM has been able to process this proficiently smaller Uralic languages. This is in itself a major development, and their capabilities should be extensively tested against different tasks. These could include machine translation, interlinear glossing, disambiguation or dependency parsing, just to mention a few usually manual work phases that the researchers of Uralic languages have been engaging with very regularly, and where automatization could have a major impact. Transliteration and normalization in themselves are tasks that the researchers may not have performed that often before. We have only recently started to receive high quality Unicode versions of the older transcriptions, and thereby the need may have not been acute yet either.

Ethics statement

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

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

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

Acknowledgments

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

References

Alidar Asvarov and Andrey Grabovoy. 2024. Neural machine translation system for Lezgian, Russian and Azerbaijani languages. *arXiv preprint arXiv:2410.05472*.

Jeremy Bradley. 2017. Transcribe.mari-language.com Automatic transcriptions and transliterations for ten languages of Russia. *Acta Linguistica Academica*, 64(3):369–382.

Jeremy Bradley and Rogier Blokland. 2023. Mansi et al. in Print before and under Unicode. *Linguistica Uralica*, 59(4):243–257.

Jeremy Bradley and Elena Skribnik. 2021. The many writing systems of Mansi: challenges in transcription and transliteration. *Multilingual Facilitation*, page 12.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R Rabiee. 2024. LLM-powered grapheme-to-phoneme conversion: Benchmark and case study. *arXiv e-prints*, pages arXiv–2409.

Arvid Genetz. 1879. Orosz-lapp nyelvmutatványok (máté evangélioma és eredeti textusok). *Nyelvtudományi közlemények*, 15(1):74–152.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2020. Normalization of different Swedish dialects spoken in Finland. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 24–27.

Csilla Horváth. 2019. The 'extraordinary thing': the only Mansi newspaper on online presence and social media practices. In *Digitalne medijske tehnologije i društveno-obrazovne promene 8*, pages 165–176, Serbia. Univerzitet u Novom Sadu. International scientific conference The Bridges of Media Education ; Conference date: 14-09-2018 Through 15-09-2018.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Artturi Kannisto. 1956. *Wogulische Volksdichtung. III*. Number 111 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki. Gesammelt und übersetzt von Artturi Kannisto, bearbeitet und herausgegeben von Matti Liimola.

Bernhard Munkácsi. 1952. *Volksbräuche und Volksdichtung der Wotjaken*. Number 102 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki. Aus dem Nachlasse von Bernhard Munkácsi herausgegeben von D. R. Fuchs.

Niko Partanen, Rogier Blokland, Michael Rießler, and Jack Rueter. 2022. Transforming archived resources with language technology: From manuscripts to language documentation. In *The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference, Uppsala, Sweden, March 15-1, 2022.*, volume 3232, pages 370–380. CEUR-WS.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Workshop on Noisy User-generated Text*, pages 141–146. The Association for Computational Linguistics.

Bhargav Shandilya and Alexis Palmer. 2024. Boosting the Capabilities of Compact Models in Low-Data Contexts with Large Language Models and Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.00387*.

Florian Siegl and Michael Rießler. 2015. Uneven steps to literacy: The history of the Dolgan, Forest Enets and Kola Sámi literary languages. *Cultural and linguistic minorities in the Russian Federation and the European Union: Comparative studies on equality and diversity*, pages 189–230.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of Large Language Models. *arXiv preprint arXiv:2404.02456*.

T. E. Uotila. 1986. *Syrjänische Texte. Band II. Komi-Syrjänisch: Ižma-, Pečora- und Vym-Dialekte*, volume 193 of *Mémoires de la Société Finno-Ougrienne*. Finno-Ugrian Society. Übersetzt und herausgegeben von Paula Kokkonen.

A Prompt example 1

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

i̯ɛɣpɣaɣɛ ȭńɣaɣɛ ĺūńš́èɣ tistèɣ. tȧβ ta· ᵏ⩽χui̯i̯. ᵏ⩽χȯsȧ ᵏ⩽χùi̯ȧs, βāťi ᵏ⩽χùi̯ȧs, nȭŋᵏ⩽χaĺ su̯nsi: i̯ɛɣpɣaɣɛ ȭǹɣaɣɛ ĺūńš́èɣ tistèɣ. nȯmsi: »a·maňri̯ɣ̆ ĺūńš́èɣ?» nȯᵏ⩽χsāikȧtaχ₍ts. kiti̯ɣ̆lii̯aɣɛ: »maňri̯ɣ̆ ĺūńši̯ⁱn?» lāβèɣ: »nȧŋ nɛ̄·n àlas₍ʟɴ.» tȧβ lāβi: »am ma·n u̯₍rl àlas₍ʟᴍ?» »nȧŋ nɛ̄n ȯsmal pinu̯ŋk lāβslamēn.» »tɛ̄!» lāβi.

B Prompt example 2

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

i̯ɛɣpɣaɣɛ ȭńɣaɣɛ ĺūńš́èɣ tistèɣ. tȧβ ta· ᵏ⩽χui̯i̯. ᵏ⩽χȯsȧ ᵏ⩽χùi̯ȧs, βāťi ᵏ⩽χùi̯ȧs, nȭŋᵏ⩽χaĺ su̯nsi: i̯ɛɣpɣaɣɛ ȭǹɣaɣɛ ĺūńš́èɣ tistèɣ. nȯmsi: »a·maňri̯ɣ̆ ĺūńš́èɣ?» nȯᵏ⩽χsāikȧtaχ₍ts. kiti̯ɣ̆lii̯aɣɛ: »maňri̯ɣ̆ ĺūńši̯ⁱn?» lāβèɣ: »nȧŋ nɛ̄·n àlas₍ʟɴ.» tȧβ lāβi: »am ma·n u̯₍rl àlas₍ʟᴍ?» »nȧŋ nɛ̄n ȯsmal pinu̯ŋk lāβslamēn.» »tɛ̄!» lāβi.

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

kit i̯ȭrn ōlèɣ. i̯ȧni̯ɣ̆ ᵏ⩽χumìtɛ nɛ̄ŋ, māń ᵏ⩽χumìtɛ nētāl nētāl ᵏ⩽χȯs ōls, βāť ōls, nɛ̄ βis. nɛ̄ βis, ūsᵊn mìnȧs, βētrȧ βìnȧ βis. ȧi̯i̯u̯ŋk tūltᵏ⩽χàtas. ȧi̯nɛ̄tɛ palìtᵊl i̯ȯl aʽt pàti. βìnàtɛ ᵏ⩽χȍlas i paʽ₍ᴅs. rȧɣ̆ɣȧʽ₍ts ta· ᵏ⩽χui̯i̯. i̯ɛɣpɣaɣɛ ȭńɣaɣɛ lāβèɣ mȧńēn nūpᵊl: »ȯsmal pinɛ̄₍ln, ĺūĺĺššȧŋ paʽ₍ᴅs.» ɛ̄₍kβȧ ȯsmȧ pinu̯ŋk tu̯β kβāls. tȧβ kāt nȭŋᵏ⩽χaĺ tȍti̯ɣ̆₍ls, kāʿtnȧ ᵏ⩽χȭiβᵊs, ɛ̄₍kβȧ š́ȧmrȧɣ̆ɣȧʽᴅs.

Кит ё̄рн олэ̄г. Яныг хумитэ нэ̄ӈ, мāнь хумитэ нэ̄тāл. Нэ̄тāл хос ōлыс, вāть ōлыс, нэ̄ вис. Нэ̄ вис, ӯсын минас, вē̄тра вина вис. Аюӈкве тӯлтхатас. Айнэ̄тэ палытыл ёл ат паты. Винатэ холас и патыс. Рагатас та хуи. Ягпыгагеоньгаге лāвēг маньнэ̄ нупыл: «Осмал пинэлн, лю̄льщаӈ патыс.» Э̄ква осма пинуӈкве тув квāлыс. Тав кāт нōӈхаль тотыглыс, кāтна хōйвес, э̄ква щам-рагатас.