Fine-Tuning Whisper for Kildin Sami

Enzo Gamboni
University of Eastern Finland
egamboni@uef.fi

Abstract

For this study, Whisper, an automatic speech recognition software, was fine-tuned on Kildin Sami, an endangered and low-resource Uralic language, using an automatic speech recognition-tailored dataset of less than 30 minutes. Three different Whisper models were trained with this dataset—each one with a different base language (English, Finnish, or Russian)—to examine which model provided the best result. Results were measured using Word Error Rate; fine-tuning the Russian-base Whisper model resulted in the lowest Word Error Rate at 68.55%. While still high, this result is impressive for only a small amount of language-specific training data, and the training process yielded insights relevant for potential for further work.

1 Introduction

This paper summarizes the results of a study carried out between 2024–2025 and submitted as an MA thesis (Gamboni, 2025).

1.1 Background

Endangered languages are those languages which are at risk of losing their speaker base, largely due to language shift (Grenoble, 2011). Relatedly, low-resource languages are those which lack significant data for natural language processing (NLP) (Joshi et al., 2020; Maguer-esse et al., 2020). Kildin Sami, a language of the Eastern Sami group and more broadly belonging to the Uralic language family (Sammallahti, 1998), is both low-resource and endangered, with estimates that only 20 active speakers still remain (Scheller, 2024). This makes Kildin's status a precarious one, in which the procurement of the large quantities of data traditionally needed for NLP is not feasible, yet all the more important.

Including low-resource, endangered languages in NLP is not only beneficial for NLP because it provides a more expansive data pool to boost accuracy, but also beneficial to endangered languages because it 1) bolsters their digital presence, contributing to the groundwork that will help to safeguard them in an increasingly digital and global society, and 2) aids researchers in more streamlined, less time-consuming workloads, as previously manual tagging and transcription could be partially or fully automated with computational methods (Trosterud, 2006; Poibeau and Fagard, 2016; Partanen et al., 2021).

This study trained Whisper, an automatic speech recognition (ASR) model on Kildin Sami data to see if significant, useful results could be achieved with ultra-minimal training data adapted from fieldwork data. A secondary goal of this study was thus to prove that Kildin fieldwork recordings can be useful in NLP research, in line with Himmelmann (1998) assumptions that an analytic approach to documentary linguistics results in data relevant to a broad subset of linguistic fields.

1.2 Current Digital Resources for Kildin Sami

An online Kildin-Russian dictionary (Antonova and Scheller, 2021–) is available and linked to an automaton for paradigm generation,[1] as well as a keyboard layout[2] for the standardized Cyrillic orthography developed in the 1970s and 1980s by a group lead by Rimma Kuruch and including Alexandra Antonova, who is among the authors of the aforementioned

[1] See the dictionary's imprint https://sanj.oahpa.no/about/. It is unclear whether or where this tool is available elsewhere.

[2] See https://giellatekno.uit.no/cgi/index.sjd.eng.html.

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 106–111
December 10-12, 2025 ©2025 Association for Computational Linguistics

dictionary (Rießler, 2020).

2 Methodology

This section describes the primary data used for this study; how it was adapted for ASR training; and subsequently details the ASR training process.

2.1 Data

Data used for this project comes from the Kildin Sami corpus, a private repository under the langdoc Github repository.[3] The corpus contains textual annotation data in XML (Rießler, 2024, 42), including time alignment to field recordings. Requests for access should be addressed to the repository's administrators. The fieldwork recordings used in this project come from the Kola Sami Documentation Project (KSDP) (Rießler, 2005–2025). These field recordings are housed in The Language Archive, for which access may be requested by contacting the administrators.

This primary data amounts to 38 minutes and 4 seconds of audio files and is comprised of three KSDP video recordings and one 39 track audiobook. All of the audio comes from one speaker, Sami language activist Nina Afanasyeva. The audiobook, *Miŋgá* (Vinogradova, 2007), is a collection of short poems written by Russian and Sami poet Iraida Vinogradova and the Kildin Sami speech is 25:14 in length. All three video recordings are largely monologues, with decent audio and infrequent background noise. Similarly, Nina Afanasyeva's audiobook narration is high in sound quality, though several tracks contain background music that at times covers her speech.

2.2 Dataset Creation and Preprocessing

First, using the tool ELAN,[4] a new textual annotation tier was created within the Kildin Sami corpus' time-aligned XML files for each audio file. These tiers were created by copying the preexisting orthographic text into the new tier and modifying it for ASR training. This process involved the following changes: removing all punctuation; removing all capitalization except for proper nouns; standardizing the transcription for false starts, nonverbal utterances, and affixes/clitics; and simplifying the Kildin standard orthography by removing macron diacritics.[5] Vinogradova's audiobook orthography was updated to reflect that which is used in Rießler's fieldwork. Notably, replacing instances of ‹’› (Unicode: 02BC) with the Cyrillic letter SHHA ‹**h/h**› (Unicode: 04BA / O4BB) (Rießler, 2013).

Next, using Audacity,[6] the audio files were manually broken into multiple .wav files, with each file corresponding to a chunk of annotation in the ASR annotation tier. Two .csv metadata files were then created—one for training the ASR model and one for evaluating the model's output—to link each shortened audio file to its transcription. 80% of the data was devoted to training the ASR model, while 20% was reserved for evaluation. ~10% of the evaluation data was taken from the audiobook recordings with the other ~10% taken from the fieldwork recordings to ensure that the evaluation results best represented the data. In sum, the resulting dataset consisted of 717 .wav files and totaled 27 minutes and 29 seconds, meaning that ~10 minutes of the primary recordings were either too poor quality to use or did not feature Nina Afanasyeva speaking.

2.3 Fine-Tuning Whisper

Three Whisper models were fine-tuned; one with English as the selected base language; one with Finnish selected; and one with Russian selected. This was done to see if different base language settings would affect end-performance. Finnish was selected because it is the closest language, linguistically, to Kildin Sami that the base Whisper model had been trained on. Russian was selected intuitively due to Kildin's use of the Cyrillic orthography. Finally, English was included, to see whether Hjortnæs et al. (2021) discovery that quantity outperformed linguistic similarity of the source language in their Komi study would also be relevant for working with Kildin.

Whisper was pretrained on 1,066 hours of Finnish data; 9,761 hours of Russian data; and

[3] https://github.com/langdoc/

[4] https://archive.mpi.nl/tla/elan

[5] In Kildin, macrons over vowels denote long vowels. However, their use across researchers is unsystematic, and vowel-length opposition in Kildin is marginal (Rießler, 2013).

[6] https://www.audacityteam.org/

Whisper Output	**Manual Transcription, Modified Orthography**	**Manual Transcription, Standard Orthography**	**English Translation**
на мэнн мэн уййнэ	на мэнн мунн уййнэ	на , мэ̄нн мунн уййнэ ?	Well, what did I see?
вуэнн уйнэ	мунн уйннэ	мунн уйннэ	I saw
тэдт инца айк аллт	тэдт инцэ айкалт	тэдт ӣнцэ а̄йкалт	this morning, early.
элляhав сулль пейв пейв вэннӭ луннӭ	элля вял шурр пеййв пеййвэнҍ лоӈӈнэ	элля вя̄л шӯрр пе̄ййв пе̄ййвэнҍ ло̄ӈӈнэ	Not yet a full day, the sun is rising.
вуаннӭсьт ляhав	ванас ли вял	ва̄нас лӣ вя̄л	It's still a little
севьнӭсьт	севвьнэсьт	се̄ввьнэсьт	dark (twilight),

Table 1: A comparison of the trained Whisper model results with the manually transcribed text.

438,218 hours of English data (Radford et al., 2023). The fine-tuning was done using HuggingFace transformers and code[7] and was executed in Google Colab.[8] A ColabPro subscription provided Nvidia GPU access. Whisper's small-sized model was used for each and trained on 500 steps. When attempting to train the model using more than 500 steps, the execution time increased dramatically and became impractical to run with the limited computational resources and time available. Each model was evaluated for Word Error Rate (WER) by using the evaluation split from the data set during the fine-tuning process. This WER calculation was done automatically at the end of the training process.

3 Results

Of the resulting models, the one set to Russian performed best, achieving a 68.55% WER. The model set to Finnish resulted in a 71.38% WER while the one set to English did the worst with a 73.88% WER. This is notable, as it suggests that orthographic similarity may have played a greater role in the improvement of WER than linguistic similarity or the quantity of pretraining data.

3.1 Transcription Analysis

The fine-tuned, Russian-based model was used to transcribe 30 seconds of audio from the test split. It took ~6 minutes to transcribe the 30 second audio clip, a portion of which is shown in Table 1 together with the manual transcriptions in both the standard and modified orthographies. An English translation is provided. Words that Whisper transcribed incorrectly are underlined in the Whisper Output tier.

The model struggles to discern between single and double consonants; vowel quality; and occasionally word boundaries. In instances where the model output is completely dissimilar to the expected output, it may be pertinent to review that specific audio section to see if there is background noise interfering with speech clarity. Further analysis of this model using character error rate (CER) analysis would offer greater insight into the nature of these errors.

3.2 Comparison to Prior Studies

Table 2 shows how the Kildin model performed in relation to models from prior studies trained on comparable amounts of data (with the exception of North Sami and Zyrian Komi, included to show work done on other Uralic languages). These results show that fine-tuning Whisper on Kildin produced comparable results to other models also fine-tuned on ≤30 minutes of data, whether trained on a Whisper model or Wav2Vec2. The lowest WER acheived with ≤30min. data was from Meelen et al. (2024) training Dzardzongke on Wav2Vec2.

Comparing these results suggests that while further experimentation may lead to WER im-

[7] https://huggingface.co/learn/audio-course/en/chapter5/fine-tuning

[8] https://colab.research.google.com/

Language	Available Data	ASR System	WER	Study
North Sami	88 unlabelled hours 20 labelled hours	Wav2vec2 + extended fine-tuning on Finnish	28.84%	Getman et al. (2024)
Dzardzongke	30 minutes	Wav2vec2	50%	Meelen et al. (2024)
Kildin Sami	27 minutes	Whisper Small	68.55%	Gamboni (2025)
Bribri	29 minutes	Whisper Medium	65-75%	Jimerson et al. (2023)
Guarani	19 minutes	Whisper Medium	65-75%	Jimerson et al. (2023)
Newar	30 minutes	Wav2vec2	74%	Meelen et al. (2024)
Zyrian Komi	35 hours	DeepSpeech + Komi/Russian LM	76.50%	Hjortnæs et al. (2021)

Table 2: Comparing Kildin's results to those of other studies surveyed during project. For referenced studies in which multiple languages were tested, only those with ≤30min. of data were included. If a study tested multiple ASR systems and Whisper was among them, Whisper's results were chosen to compare.

provements for Kildin, it is unlikely to improve to a WER <50% or to approach the success Getman et al. (2024) found with many hours of data for North Sami.

4 Conclusion and Future Potential

Whisper offers promising results when trained on ultra-minimal data for Kildin Sami and supports Himmelmann (1998) assumption that an analytic approach to documentary linguistics produces relevant data for a broad subset of linguistic fields. Although 68.55% WER is high, it is remarkable to be achieved with a data set of less than 30 minutes combined for training and testing and shows how advancements in NLP are making the inclusion of endangered and low-resource languages more feasible. Despite the author's lack of computational background, significant results were still achieved and could well become useful for semi-automating Kildin transcriptions with further experimentation. The author hopes that this study can serve as a starting point for further experimentation on training Whisper on Kildin Sami and as proof that those with a limited computational background can still incorporate computational methods into their linguistic research.

This study was influenced by Hjortnæs et al. (2021) observations that source language quantity was more impactful than linguistic similarity for Komi, but finds that the same did not hold true for Kildin; rather, it seems that shared orthography played a greater role. Future work focusing on how to simultaneously leverage the orthographic similarity of Russian and the linguistic similarity of Finnish to Kildin, would be beneficial to consider for improving WER and further testing this assumption. A reexamination of the ASR dataset created for this project would also be worthwhile to see if decisions made during preprocessing significantly impacted the ASR training. This reexamination should be done after more in-depth analysis of the current model's output is undertaken to discern if there are commonly repeated errors that could be stemming from human error or decision-making within the dataset. Lastly, experimentation with training the model on a greater number of steps or on a larger Whisper model may also yield greater WER and contribute to the robustness of this study.

5 Limitations

Time was a limiting factor on this study's depth. Minimal speech data available for training the Kildin Sami model was another inherent limitation.

Limitations concerning the definition of linguistic vs. orthographic similarity mentioned during this study must also be addressed. Though I posited that the Russian-based model performed best due to orthographic similarity, an anonymous reviewer pointed out that linguistic similarity may still be the reason for this, as Kildin and Russian share features like palatalisation, while Finnish does not. Relatedly, transcription of long vowels, something the Russian-based model struggled with considerably, could be attributed to the absence of length distinction in Russian, further highlighting the role of base language similarity. Thus, speculation and claims within this study on the role of linguistic vs. orthographic similarity are limited due to a lack of in-depth analysis on the subject. As this work is ongoing, this topic will be further explored. My gratitude is extended to the reviewer who raised this concern.

6 Ethical Considerations

None of the materials used contain any sensitive or personal information, nor are any of them being freely distributed in their entirety for this project. Nina Afanasyeva has given her informal consent to have recordings of her from the Kola Sami Language Documentation Project used for the purpose of language technology development.[9]

Use of audio taken from Vinogradova (2007), which is under copyright, adheres to the copyright laws within the European Union.[10] However, because data taken from copyrighted material may not be made publicly available, the dataset used to train the ASR models is housed in a private repository.

References

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

Enzo Gamboni. 2025. Fine-tuning Whisper for Kildin Sami, a low-resource endangered language. Master's thesis, University of Eastern Finland.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi. In *Interspeech 2024*, pages 2539–2543.

Lenore A. Grenoble. 2011. *Language ecology and endangerment*, page 27–44. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. 36:161–195.

Nils Hjortnæs, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2021. The relevance of the source language in transfer learning for ASR. In Miikka Silfverberg, editor, *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, volume 1, pages 63–69. University of Colorado Boulder.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Marieke Meelen, Alexander O'neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of Nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93, St. Julians, Malta. Association for Computational Linguistics.

Niko Partanen, Michael Rießler, and Joshua Wilbur. 2021. Envisioning digital methods for fieldwork in the Arctic. In Markku Lehtimäki, Arja Rosenholm, and Vlad Strukov, editors, *Visual representations of the Arctic*, Routledge Interdisciplinary Perspectives on Literature, pages 313–339. Routledge.

Thierry Poibeau and Benjamin Fagard. 2016. Exploring Natural Language Processing Methods for Finno-Ugric Langages. In *Proc. of the Second International Workshop on Computational Linguistics for Uralic Languages*, Proc. of the Second International Workshop on Computational Linguistics for Uralic Languages, Szeged, Hungary.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of*

[9] Michael Rießler (KSDP), p.c.

[10] The EU Directive 2019/790 on Copyright in the Digital Single Market outlines copyright exceptions that allow for text and data mining for scientific researcher purposes.

the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Michael Rießler. 2005–2025. Kola Saami Documentation Project (KSDP). In *The Language Archive (TLA)*. Max Planck Institute for Psycholinguistics.

Michael Rießler. 2013. *Towards a digital infrastructure for Kildin Saami*, pages 195–218. Exhibitions and Symposia. Kulturstiftung Sibirien.

Michael Rießler. 2020. Rimma Kuruch and Kildin Saami language planning. *Linguistica Uralica*, 56(3):220–225.

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

Pekka Sammallahti. 1998. *The Saami languages: an introduction*. Davvi girji, Kárášjohka.

Elisabeth Scheller. 2024. Activating passive Kildin saami language knowledge through the Master-Apprentice Language Learning Method and instruction in grammar and writing skills. 2/2024:82–108.

Trond Trosterud. 2006. *Grammatically based language technology for minority languages*, pages 293–316. De Gruyter Mouton.

Iraida V. Vinogradova. 2007. *Miŋgá = Mīnn'kaj*. Davvi Girji.