Building an Aligned Speech Corpus for Northern Russian

Kseniia Protonina
Higher School of Economics
kiprotonina@edu.hse.ru

Daniil Ignatev
Utrecht University
d.ignatev@uu.nl

Abstract

Russian dialects remain underrepresented in speech recognition resources despite phonological and lexical divergence from the standard language. The paper reports the results of a pilot study on constructing a corpus for dialect-aware speech processing. We first make an argument in favor of dialect-aware ASR – particularly, for Russian. We then describe a reproducible procedure for converting long archival recordings, coarse timestamp indices and manual transcripts into utterance-level aligned data, and introduce a pilot corpus of the Vologda dialect of Russian that is highly scored by expert evaluators. Testing zero-shot ASR models against the corpus reveals that current models remain sensitive to dialectal speech, motivating further expansion of our data collection.

1 Introduction and Background

In recent years, there have been growing efforts to extend speech recognition technologies to underrepresented languages and dialects (Faisal et al., 2024; Joshi et al., 2025), particularly in the corpus domain (Bentum et al., 2024; Oberkircher et al., 2026; Thillainathan et al., 2026). This development broadens the ASR coverage of natural speech and voice-controlled applications, while highlighting performance disparities between language varieties, even when the standard language is well supported (Faisal et al., 2024; Dipto et al., 2025).

Russian is well-represented in the existing speech corpora such as the Whisper dataset (Radford et al., 2022), Russian LibriSpeech (Pratap et al., 2020), Mozilla Common Voice (Ardila et al., 2020), and Golos (Karpov et al., 2021). However, dialectal Russian remains limitedly represented: regional variation is rarely distinguished in large ASR corpora, while dialect-focused resources are often designed for linguistic documentation rather than speech recognition and appear scant in size or accessibility. The dialectal part of the Russian National Corpus (Kachinskaya and Sitchinava, 2015; Savchuk et al., 2024) is an instance of this trend, although it includes a large amount of dialect recordings. Similarly, smaller corpora also provide valuable dialectological material (e.g., Daniel et al. (2013); Knyazev (2023)), but they are often only accessible via their search interfaces and thus are not readily available as ASR data collections. This data shortage limits the robust handling of Russian dialects by speech-to-text tools and may, thereby, reinforce their perceived marginal societal position.

In our pilot study, we motivate dialect-aware ASR and introduce a novel aligned speech corpus of the Vologda dialect, a subtype of Northern Russian. Specifically, our research questions are: (1) how accurately does the pipeline recover word-level temporal boundaries, evaluated against manual annotation? (2) how well do current ASR models perform on the resulting dialectal data in a zero-shot setting?

We describe the dialect and source materials, our audio-transcription alignment method and evaluation, and the resulting corpus. Finally, we benchmark ASR models on the corpus in a zero-shot setting. [1]

2 Source Data

2.1 Dialectal Properties

Vologda dialect is considered a subtype of Northern Russian, a group of dialects spoken primarily in the northern regions of European Russia, including the Arkhangelsk, Vologda, and Kostroma regions (Daniel et al., 2019). Northern Russian dialects are characterized by a set of features that distinguish them from the Moscow-based standard. The most

[1] Our data is available at https://huggingface.co/datasets/hse-prs-folklore/pez. The code is available at https://github.com/ruthenian8/alignment.

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 257–263
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_026

prominent phonological feature is *okanye*: the absence of vowel reduction in unstressed syllables, where unstressed /o/ is preserved as a distinct vowel rather than merging with /a/. Additional features include *tsokanye* (the merger of the affricates *ch* and *ts* sounds in some subdialects) and the retention of archaic morphological and lexical forms.

Our study does not cover Vologda Russian in its entirety, as we focused on a set of recordings collected in Pezhma, a village in the South of Arkhangelsk oblast that falls within the Vologda dialectal area. Still, although a limited number of younger speakers exhibit none or few dialectal features, most speakers in the set do represent Vologda Russian.

2.2 Provenance and Formatting

The Folklore Archive of the Russian State University for the Humanities (Moroz, 2015) incorporates the field data collected in expeditions to the Arkhangelsk oblast, taking place from 1993 on a yearly basis. These expeditions secure novel linguistic materials and focus thematically on folklore studies and dialectology. The collection procedure involves conducting in-depth recorded interviews with local participants, followed by partial transcription of the recordings.

Interviews are manually transcribed during expeditions by linguistically trained collectors who receive prior ear training in the local dialect. Transcriptions cover only thematically relevant fragments, leaving parts of the audio untranscribed. Manually curated index files for each interview record the transcribed and skipped segments with timestamps, along with metadata such as recording time, location, and speaker and collector names.

The archive currently comprises interviews from 12 localities in the Arkhangelsk region totaling to 808 audio recordings and approximately 1046 hours of recorded speech. For our pilot study, we focus on the Pezhma subcorpus, including 48 recordings (around 57 hours). Discarding 3 files with corrupt transcripts leaves us with a final set of 45 interviews. Each recording is accompanied by a manual transcription and a coarse timestamp index serving as the entry point for our pipeline.

The manual transcriptions pose a challenge for straightforward processing due to their specific transcription format. First, they follow a non-standard orthography reflecting dialectal pronunciation, including stress marks on vowels. This orthographic divergence can affect audio-text alignment, as the dialectal spellings differ from the forms represented in standard Russian acoustic models, and can consequently impact word-level ASR metrics. At the same time, they retain standard syntax as opposed to pause marks, standard in phonetic transcription. Secondly, utterances from two or more speakers, often overlapping, are transcribed in a linearized manner as opposed to multilevel transcription. Finally, collector's notes (e.g., [laughter] or [unclear]) are interleaved with the transcription and only marked by brackets. Cf. the excerpt below (backward slash denotes stress by convention):

[S002:] Ну бы\ли э\ти зна\хари... [нрзб.]
nu byli èti znaxari
"Well, there were these healers... [unclear]"

3 Alignment

3.1 Protocol

In our approach, we draw on the increasingly popular method of leveraging ASR outputs as a proxy for their respective audio sources, shifting the alignment task to text-to-text matching of gold captions and these intermediary targets. Some examples of this approach are the procedure adopted by Bentum et al. (2024) in the Spoken Dutch Corpus and the system for Russian speech presented in Borodin et al. (2025). These works leverage intermediary ASR transcriptions from Wav2vec 2.0 (Baevski et al., 2020) and a family of models including GigaAM (Kutsakov et al., 2025) respectively.

Inspired by the said studies, we designed a five-stage pipeline. The first step is audio preprocessing, which involves converting all audio to 16kHz WAV as well as additional loudness normalization and denoising, as described in Appendix A. It is followed by the joining of timestamped indices and fragment-level manual transcriptions, initially loosely matched, thereby ensuring a strict one-to-one mapping. These are joined by their positions; in case of mismatches, word overlap is used as a heuristic. Third, we transcribe each timestamped fragment with WhisperX (Bain et al., 2023), resulting in utterance-level segments. Our choice of WhisperX is dictated by its out-of-the-box capabilities in segmentation and diarization. Thus, even though the direct ASR performance of WhisperX appears inferior to newer, Russian-focused models (see Section 5), this does not impact our pipeline strongly.

The fourth step is transcript alignment and corpus export. The alignment stage finds the best match between each WhisperX-produced segment and a fragment of the manual transcript via dynamic programming, using the token-set F1 score ($\frac{2|A \cap B|}{|A|+|B|}$) as a similarity measure. The use of intermediary WhisperX output is restricted to the temporal scaffold. Manual transcripts remain the source of truth for the exported text. Speaker labels are inferred from the manual transcript when possible, including speaker initials at the end of transcript blocks.

At the fifth and final, export stage, we filter interviewer turns from the output, so as to keep them separate from dialectal speech. Our reason is that these utterances are intentionally left unattributed in some of the corpus files, unlike the speaker responses. Nevertheless, we retain this data in the collection for future use.

3.2 Evaluation

Our evaluation protocol focused on the utterance level. In **question 1**, we adopted the schema from Bentum et al. (2024) that covers 5 categories: "good" for full match, "bad" for full mismatch, as well as "start match", "end match", and "middle match". Meanwhile, we also noted that this taxonomy does not reflect the quality of partial matches. To estimate the latter, we introduced **question 2**: a clip and a caption are "well-matched" if 75% of the caption words match the audio and not otherwise. For example, for a 6-word clip, a mismatch of 2 words would trigger the answer "no", while a shorter mismatch still triggers "yes". Despite its coarseness, this metric appeared intuitive for the annotators, resulting in high inter-worker consistency.

We recruited 5 paid workers among trained students with linguistic background to score a sample of 1000 clips, stratified across the 45 source interviews, based on the above schema. Each annotator received from one to three batches of 100 items. We used a self-hosted Potato (Jurgens et al., 2026) instance for data collection.

500 of the clips were doubly annotated to estimate the inter-worker agreement; we chose Krippendorf's alpha as our IAA metric. The obtained scores equal 0.872 for question 1 and 0.934 for question 2, pointing at acceptable annotation quality. Likewise, we scored all workers using the MACE algorithm (Hovy et al., 2013), resulting in scores above 0.7, i.e., moderate to high reliability. Doubly-annotated labels were aggregated based on the MACE scores.

Label	Ratio
Q1: Good	57.90%
Q1: Start match	16.30%
Q1: End match	16.20%
Q1: Middle match	5.30%
Q1: Bad	4.30%
Q2: Well-matched	78.03%
Q2: No	21.97%

Table 1: Evaluation label distribution.

The results are reported in Table 1. We observe that completely mismatched files constitute a relatively small fraction of 4.3%, indicating that our solution is generally viable. At the same time, the proportion of partial matches is high, similarly to results by Bentum et al. (2024). In turn, Q2 shows that the ratio of files with none- or minor mismatches is high overall.

4 Corpus Design

4.1 Metadata Schema

The metadata is stored in a single "manifest" table, where each row corresponds to one audio segment and its associated transcription. Speaker information is included separately for every utterance.

The folder structure is multilevel. At the global level, files are organized per interview. Within each interview folder, utterances are stored in subfolders per thematic fragment, reflecting the original index structure of the archive. Each segment is identified by a filename including the interview ID, the fragment number, the pseudonymized speaker tag, and the timestamp. Each audio file is paired with a plain-text transcription file. The corpus manifest includes, for each segment: the audio path, the transcript path, the speaker label and quality flags indicating whether the segment was manually reviewed or flagged as possibly misaligned.

4.2 Corpus Statistics

Table 2 provides an overview of the pilot subcollection of the Pezhma corpus. To compute the overall duration and utterance number, we automatically diarized multiple speakers per interview using WhisperX and filtered out the interviewer turns. Considering the speaker groups, male and female speakers are comparably represented. Meanwhile,

the age groups are highly imbalanced, as more than half of the speakers fall into the 40+ category. However, this also entails a better representation of the dialectal features, as they are more prominent in the 40+ stratum.

Parameter	Value
Selected interviews	45
Total duration	$\approx$ 57 h.
Aligned utt.	9 641
Mean utt. length (words)	8.5
Median utt. length (words)	7
Max utt. length (words)	53
Total speakers	63
Female / male speakers	37 / 26
Age: 0-20 / 20-40 / 40+	13 / 2 / 48

Table 2: Final corpus overview.

4.3 Consent and Anonymization

All participants provided informed consent before recording, and their identities were concealed for privacy reasons; they are pseudonymized using hash-based speaker tags in both the transcripts and the corpus metadata.

5 Model Benchmarking

5.1 Models

In order to verify how well existing speech recognition solutions handle the dialect under scrutiny, we evaluate a number of pretrained ASR models, both multilingual and Russian-specific, on our data. This evaluation is performed zero-shot, without further fine-tuning or dialect-specific adaptation. The list of models is as follows. **Whisper Large-v3** (Radford et al., 2022), a state-of-the-art solution for multilingual ASR. Second, **MMS**, Meta's massively multilingual model. Finally, **GigaAM** (Kutsakov et al., 2025), a Russian-specific Conformer-based model yielding SOTA results on Russian; GigaAM offers **RNNT**- and **CTC**-decoding versions, both of which are covered in our evaluation.

5.2 Evaluation

As the evaluation is fully zero-shot, we report the metrics obtained on the entirety of corpus data. We use WER as our primary evaluation criterion and also report CER, which helps account for whether the phonetic spelling makes a difference. Punctuation marks and capitalization are ignored in the evaluation. Diacritics used in transcripts to indicate phone duration during evaluation are stripped away and ignored in the WER and CER computation.

5.3 Results

Table 3 reports zero-shot ASR performance on the corpus data.

Model	WER	CER
GigaAM-RNNT	36.02	24.57
GigaAM-CTC	38.00	24.10
Whisper Large-v3	42.33	29.27
MMS	67.94	34.94

Table 3: Zero-shot ASR results on the corpus data. WER and CER are percentages computed on normalized references and predictions.

5.4 Discussion

The strongest result is obtained by GigaAM RNNT, but even this model yields a WER above 36%. The two GigaAM variants show lower CER than Whisper, suggesting that many errors are short character-level deviations rather than complete word failures, but the word-level mismatch rate remains high for all systems. MMS performs substantially worse, as expected from a primarily multilingual model, not tailored to Russian dialectal speech.

At the same time, pretrained models appear to struggle to process dialectal data as opposed to regular, normative speech. This trend is further confirmed by the error analysis below.

Qualitative inspection of low-overlap segments revealed three systematic error types. (1) *Dialectal phonology*: northern Russian features *okanye*—the preservation of the /o/–/a/ distinction in unstressed syllables—that has been lost in standard Russian; standard-trained models often fail to recognize these words (Daniel et al., 2019). (2) *Non-standard lexicon*: dialect-specific words absent from standard training data are substituted or omitted. In addition, the systems may struggle with the general domain shift due to the focus of the interviews on folklore and traditional culture. (3) *Spontaneous speech phenomena*: filled pauses, false starts and ellipses create mismatches between manual transcripts and model output. Filtering interviewer turns and manually reviewing boundary segments mitigated these issues.

6 Conclusion

We presented a compact pipeline for turning partially transcribed archival field recordings into an aligned utterance-level corpus for dialect-aware ASR. The pilot on Pezhma recordings shows that coarse archival indices and manual transcripts can be converted into ASR-ready data while preserving the original dialectal orthography. The boundary evaluation indicates that the resulting segmentation is usable but still imperfect, and zero-shot ASR benchmarks reveal a clear performance gap on dialectal speech. Future work will expand the pipeline to the remaining localities in the archive, improve manual validation of boundary cases and use the resulting data for dialect-aware model adaptation.

7 Limitations

The authors admit that the proposed alignment procedure is not error-free, and that the corpus can occasionally contain erroneous alignments. However, thanks to the corpus metadata, files with low aligner confidence, which largely coincide with errors, can be filtered out in downstream applications.

As indicated earlier, our study focused on recordings from one village for pilot purposes; hence, our current release does not represent Northern Russian in full.

We did not test the produced alignment system on different languages and dialects, even though it should be compatible with more WhisperX-compatible languages.

7.1 Disclosure of AI assistance

In the current project, we used AI-based coding assistant to refactor and consolidate an early version of our code-base. The authors manually verified the code and the results.

8 Acknowledgments

We are grateful to Daria Savina and Eleonora Izmailova for their contributions to the project. We thank the anonymous reviewers for helping us improve the paper. Finally, we are deeply thankful to the makers and maintainers of the RSUH Folklore archive as well as the recorded speakers.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. ArXiv:1912.06670 [cs.CL].

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Martijn Bentum, Eric Sanders, Antal PJ van den Bosch, Douwe Zeldenrust, and Henk van den Heuvel. 2024. Corpus creation and automatic alignment of historical dutch dialect speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4021–4029.

Kirill Borodin, Nikita Vasiliev, Vasiliy Kudryavtsev, Maxim Maslov, Mikhail Gorodnichev, Oleg Rogov, and Grach Mkrtchian. 2025. A data-centric framework for addressing phonetic and prosodic challenges in russian speech generative models. *arXiv preprint arXiv:2507.13563*.

Michael Daniel, Nina Dobrushina, and Ruprecht von Waldenfels. 2013. The language of the ustja river basin. a corpus of north russian dialectal speech.

Michael Daniel, Ruprecht von Waldenfels, Aleksandra Ter-Avanesova, Polina Kazakova, Ilya Schurov, Ekaterina Gerasimenko, Daria Ignatenko, Ekaterina Makhlina, Maria Tsfasman, Samira Verhees, and et al. 2019. Dialect loss in the russian north: Modeling change across variables. *Language Variation and Change*, 31(3):353–376.

Tawsif Tashwar Dipto, Azmol Hossain, Rubayet Sabbir Faruque, Md Rezuwan Hassan, Kanij Fatema, Tanmoy Shome, Ruwad Naswan, Md Foriduzzaman Zihad, Mohaymen Ul Anam, Nazia Tasnim, et al. 2025. Are asr foundation models generalized enough to capture features of regional dialects for low-resource languages? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 178–188.

European Broadcasting Union. 2011. Recommendation r 128: Loudness normalization and permitted maximum level of audio signals. EBU Technical Recommendation. First issued February 2010; revised 2011, 2014, 2020 and 2023.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

David Jurgens, Michael Chen, and Lina Iyer. 2026. Potato 2.0: A comprehensive annotation platform with support for ai-in-the-loop and agentic systems. In *Proceedings of the The 64th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Irina B. Kachinskaya and Dmitri V. Sitchinava. 2015. Today's state of the dialectal subcorpus. In *Proceedings of the V.V. Vinogradov Russian Language Institute. № 6.*, pages 142–163.

Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. 2021. Golos: Russian Dataset for Speech Research. ArXiv:2106.10161 [eess.AS].

S. V. Knyazev. 2023. Corpus of the russian dialect spoken in the village of veegora. [Online resource]. Available at: https://lingconlab.ru/veegora (accessed: 2026-05-19).

Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, and Fyodor Minkin. 2025. GigaAM: Efficient Self-Supervised Learner for Speech Recognition.

Andrei Moroz. 2015. The folklore archive of the russian state university for the humanities in moscow. *Studia mythologica Slavica*, 7:175–179.

Lena S Oberkircher, Jesujoba O Alabi, Dietrich Klakow, and Jürgen Trouvain. 2026. Saar-voice: A multi-speaker saarbr\" ucken dialect speech corpus. *arXiv preprint arXiv:2604.11803*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Resemble AI. 2023. Introducing resemble enhance: Ai powered speech denoising and enhancement. Blog post. Accessed 20 May 2026.

S. O. Savchuk, T. Arkhangelskiy, A. A. Bonch-Osmolovskaya, O. V. Donina, Yu. N. Kuznetsova, O. N. Lyashevskaya, B. V. Orekhov, and M. V. Podryadchikova. 2024. Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, 2:7–34.

Hendrik Schroter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier. 2022. Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7407–7411. IEEE.

Adsajan Thillainathan, Nishanthini Kanthakumar, Nivethiga Rasan, and Kengatharaiyer Sarveswaran. 2026. Talk-corpus: A regionally diverse evaluation set for sri lankan tamil speech. In *Proceedings of Speech Language Models in Low-Resource Settings: Performance, Evaluation, and Bias Analysis (SPEAKABLE) @ LREC 2026*, pages 194–200, Palma de Mallorca, Spain. ELRA and ICCL.

A Normalization

All source audio was converted to 16kHz mono WAV prior to alignment for compatibility with ASR. Denoising was applied selectively based on signal-to-noise ratio (SNR) estimates. Table 4 summarizes our three-level denoising strategy. For severely noisy segments (SNR $<$ 9 dB) we employed the Resemble Enhance denoiser–enhancer pipeline (Resemble AI, 2023). For moderate noise (9 dB $\leq$ SNR $<$ 11 dB) we used DeepFilterNet, a low-complexity deep filtering framework for full-band speech enhancement (Schroter et al., 2022). Segments with SNR $\geq$ 11 dB were left unaltered. A high-pass filter at 80Hz and a low-pass filter at 7,500 Hz removed low-frequency rumble and high-frequency interference. We then performed loudness normalization using the `ffmpeg` double-pass `loudnorm` filter implementing the EBU R128 standard for loudness normalization and permitted maximum level (European Broadcasting Union, 2011), targeting an integrated loudness of -16 LUFS. Table 5 summarizes the effect of preprocessing on audio quality across the corpus.

SNR (dB)	Method
<9	Resemble AI (2023)
9–11	Schroter et al. (2022)
>11	-

Table 4: Denoising strategy by SNR level

After normalization the loudness standard deviation dropped from 3.97 to 2.39 LUFS, reflect-

Metric	Before	After
Mean LUFS	-12.4	-17.4
Std LUFS	3.97	2.39
Mean SNR (dB)	14.1	15.1
SNR range (dB)	6.0–27.9	6.7–32.2
Files with SNR < 10dB	7.6% (731)	< 1% (91)

Table 5: Audio preprocessing statistics (N = 9,617 segments)

ing greater consistency across segments. The final mean LUFS of -17.4 remains slightly below the -16 LUFS target, likely due to wind noise and fluctuating speech activity in field recordings, which affect the gating mechanism of the EBU R128 algorithm.