A data-centric approach to performance improvement in under-resourced ASR: The case of Dënë Sųłné

Olga Kriukova[1], **Olga Lovick**[1], **Antti Arppe**[2]

[1]University of Saskatchewan, [2]University of Alberta

Correspondence: olga.kriukova@usask.ca

Abstract

This paper presents a study focused on advancing Automatic Speech Recognition (ASR) for the under-resourced language Dënë Sųłné through data-centric approaches. We explore multiple strategies to enhance the quality of training data—both audio recordings and transcriptions—to address the challenges posed by mixed-quality datasets. Our experiments investigate which data preparation techniques most effectively improve ASR performance in this context. Our findings show that reducing spelling variants of the same lexeme in the corpus significantly improves model generalization, resulting in a substantial increase in recognition accuracy. Additionally, we demonstrate that increasing manually reviewed transcriptions consistently improves word and character error rates, while audio enhancement slightly reduces performance, highlighting the complex trade-offs in low-resource ASR development.

1 Introduction

Neural models have significantly expanded the tools and resources available for processing major languages. However, recent advances in machine learning have enabled languages with scarce linguistic data to also benefit from these models. Performance of neural models heavily depends on the quality of the training data (Sambasivan et al., 2021), yet under-resourced languages face a compounding problem: data is not only scarce but also particularly hard to obtain in high-quality form (Kjartansson et al., 2018; Liang and Levow, 2025).

Researchers working with major languages often take data quality for granted, relying on well-curated, validated benchmark datasets (cf. Joshi et al., 2020) such as—in the case of speech recognition technologies—Mozilla Common Voice (Ardila et al., 2020) or VoxPopuli (Wang et al., 2021). These datasets, while not of uniform quality, benefit from systematic curation processes and validation that under-resourced languages typically lack. This abundance of relatively high-quality data has naturally led the field toward model-centric approaches in machine learning, where researchers focus primarily on architectural innovations and hyperparameter optimization (Bhatt et al., 2024; Jakubik et al., 2024) or adding more data. However, this model-centric paradigm is ineffective for under-resourced languages, where factors such as non-standardized orthography, inconsistent transcriptions, and variable audio quality compound data scarcity (Lin et al., 2024; Liu et al., 2022; Wisniewski et al., 2020; Zhong et al., 2024). When training data contains numerous incorrect labels, even optimal model configurations will yield poor results. This is why, data-centric approaches—which prioritize improving the quality and accuracy of training data over model tuning—are essential for effective machine learning (Whang et al., 2023).

While the field increasingly recognizes this challenge (Luger et al., 2025), data-centric approaches for under-resourced languages remain largely underexplored. Through our work on fine-tuning an Automatic Speech Recognition (ASR) model for Dënë Sųłné, an under-resourced, endangered language spoken in Canada, we demonstrate how systematic dataset improvement can yield significant gains. We present experiments across data refinement stages, identify key preprocessing steps, and highlight the need for the computational linguistics community to engage more deeply with data-centric methodologies for advancing under-resourced language technologies.

2 Backround

2.1 Dënë Sųłné

Dënë Sųłné (ISO 639-3: chp), belongs to the Dene (Athabaskan) language family and spoken across Alberta, Saskatchewan, Manitoba, and the Northwest Territories by approximately 10,000 people

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 95–106

(Statistics Canada, 2021). The data for this study were collected from speakers of the neighbouring communities of Clearwater River and La Loche in Saskatchewan.

Like other Dene languages, Dënë Sųłné is polysynthetic and predominantly prefixing, with a large consonant inventory and phonemic contrasts of tone and nasality (Cook, 2004). These typological properties, combined with the absence of a fully standardized orthography, create specific obstacles for corpus-based work that are detailed in Sections 2.2 and 2.3.

2.2 Language data

The dataset for the present study was compiled from three sources. The first and most significant portion (85%) comes from the *Talking Dene* project (2020-2024; PI Olga Lovick). The second portion (9%) was collected by Kriukova for this study. The third portion (4.4%) consists of verb paradigm recordings collected by Nial Willems (2025) from one speaker. In all three cases, participating speakers explicitly consented to the use of their language data for training the ASR model. In total, the whole corpus for this study contained language data from 22 speakers. The total length of the audio data for the final iteration of the model was 12 hours 35 minutes, and the number of utterances was 18,779; however, these numbers changed over the course of the study, as explained in detail in Section 4.

The corpus contains predominantly Dënë Sųłné utterances, along with some English and Dënë Sųłné-English mixed utterances. To preserve the model's ability to recognize both languages, we included all utterance types in the training and testing datasets. This approach helps to prevent catastrophic forgetting of English that can occur when fine-tuning on a single language (Simmons, 2025).

The audio recordings in the dataset also vary in quality. The majority of them have good or satisfactory quality, which is sufficient for successful ASR training. Nevertheless, some recordings feature interviews with multiple speakers and thus contain occasional overlapping speech by two or more speakers. Additionally, some audio clips have significant background noise that, in rare cases, makes a speech signal inaudible.

2.3 Orthographic challenges in Dënë Sųłné

The orthographic instability of Dënë Sųłné has direct, measurable consequences for ASR training data. Our training data reflects the natural variation that arises when a language lacks a common written standard. As a result, the same word frequently appears in multiple written forms, with transcribers independently applying their own perceptual spelling strategies, especially regarding nasality and tone (Kriukova et al., 2026). Generational variation compounds this: younger speakers often produce innovative verb forms—deleting prefixes, altering morpheme order—which transcribers may render phonetically or "correct" to conservative forms, generating additional surface variants for the same pronunciation (Lovick et al., 2023, 2024).

The practical effect is a training vocabulary inflated by spelling variants without semantic distinction. Prior to any standardization, our corpus contained 17,355 unique token types; targeted standardization of frequent forms reduced this to 14,892 (see Section 4.1.1). This reduction directly decreased label noise and improved model generalization, motivating the data-centric approach described in the remainder of this paper.

3 Literature Review

There is a growing body of research on ASR for under-resourced and endangered languages. Several studies have demonstrated that pre-trained multilingual models generally achieve higher accuracy in low-resource settings than monolingual models (Jimerson et al., 2023; Sadeque, 2022; Yadav and Sitaram, 2022). However, Jimerson et al. (2023) found that no single ASR architecture consistently outperforms others across under-resourced languages. In their comparison of Whisper, Wav2Vec2, Kaldi DNN, and ESPnet2 across 11 under-resourced languages with datasets ranging from 26 minutes to 19 hours, they found no correlations between Word Error Rates (WER) and morphological properties, dataset size, or recording quality. For Whisper, the resulting WER was ranging from 5% to almost 75%. Notably, Whisper demonstrated superior performance for languages with large phone sets (>37 phones).

Another critical consideration in low-resource ASR is data partitioning, which presents unique challenges when datasets contain few speakers or recordings from a single speaker. Liu et al. (2023) demonstrated that random splits provide better training and test sets than the common "hold speaker out" method when data is scarce. Nonetheless, they recommend testing multiple partitioning methods given the individual characteristics of

small datasets.

Furthermore, Wisniewski et al. (2020) emphasize that training data preprocessing for under-resourced language corpora presents different challenges than preprocessing for well-resourced languages. Unlike standard NLP preprocessing pipelines, which can rely on established tools and conventions, preprocessing endangered language data is highly non-trivial and often requires developing language-specific solutions (as Section 4.1.1 illustrates). These tasks are not only time-consuming but also demand substantial familiarity with both the target language and the specific corpus characteristics. As the authors note, the severity of these challenges can sometimes even discourage computational linguists from working with the data.

Beyond these methodological considerations, certain linguistic factors complicate ASR development for under-resourced languages. For instance, code-switching, common in minority-language communities (cf. Moore, 2018), poses significant challenges for ASR systems (Coto-Solano et al., 2022; Guillaume et al., 2022; Simmons, 2025). Similarly, languages with complex orthographies or large phoneme inventories pose particularly demanding challenges for ASR (Adams et al., 2019; Gauthier et al., 2016; Prud'hommeaux et al., 2021). As Liang and Levow (2025) demonstrate, even state-of-the-art multilingual models struggle to accurately capture nuanced phonological contrasts such as tone distinctions, nasality, consonant length, and vowel length—features that are phonemic in many under-resourced languages (cf. Jimerson et al., 2023). Furthermore, Liang and Levow (2025) and Ćavar et al. (2016) also emphasize that fieldwork speech data—characterized by spontaneous speech and varied recording conditions—poses challenges for ASR models trained on standardized corpora.

The corpus we used in this study exemplifies many of the abovementioned challenges. Dënë Sųłiné lacks a fully standardized orthography and employs written representation of nasality and tone. Additionally, the corpus contains code-switching, and most audio recordings were captured in fieldwork settings with variable quality. The following section details our approach to navigating these obstacles.

4 Methodology

4.1 Dataset improvement

The results of the baseline model iteration were occasionally plausible but mainly unsatisfactory. It became evident that without altering the training data, we would be unlikely to improve the transcription results. Therefore, we carefully examined our training corpus and identified the main issues to be addressed to achieve higher-quality transcriptions. First, spellings of individual words varied widely across transcribers, making the training data very noisy. Second, our initial dataset contained too many sentences that were not reviewed by a linguist, resulting in inconsistent word boundary distributions and missing sentence-final enclitics in some transcripts. Most critically, innovative verb forms were spelled inconsistently and often required a manual review. Third, some audio clips had excessive background noise or significant speech overlap, which could also interfere with a model's pattern generalization.

In the following subsections, we describe how we addressed these data inconsistencies, while Section 5 demonstrates the effect of these actions on our ASR model performance.

4.1.1 Standardization of orthography

As discussed in Section 2.3, orthographic variation in Dënë Sųłiné transcriptions stems from both phonetic transcription practices and inconsistent spelling conventions. High variation makes training data noisy, hindering the model's ability to generalize speech-to-text patterns. This issue is not exclusive to word-level representations: subword tokenizers, which segment words into recurring character sequences, are equally sensitive to orthographic inconsistency. When the same word form appears under multiple spellings, it may be segmented into different subword units across instances, further compounding the noise in the training data. To address this, we reduced non-phonemic variation across multiple standardization stages and model iterations, focusing on the most frequent word types (for full details on the decision-making process and variant identification, see Kriukova et al. (2026)).

In the first stage, we targeted verbal enclitics and enclitic combinations, writing them separately from verbs and each other to reduce verbal vocabulary size (e.g., *łí nį á*—habitual+past+assertive—instead of *łínįá*),

along with frequent postpositions and adverbs. In the second and third stages, we orthographically standardized pronouns, possessive prefixes, numerals, remaining adverbs and postpositions, and frequent nouns and verbs—for example, replacing reduced numeral spellings with full forms (e.g., *tae* → *taghë*) and standardizing frequent verb forms such as *nëzǫ* 'it is fun', *bënasnį* 'I remember', *nësthën* 'I think'. In the fourth stage, we standardized word tokens belonging to frequent verb paradigms such as 'to be', 'to say', and 'to talk'. In total, these changes reduced the ASR training vocabulary (word-level) from 17,355 to 14,892 unique types.

It is worth noting that orthographic inconsistency affects not only the number of word-level representations but also the number of subword units. The model used in this study (see Section 4.4) employs a subword tokenizer, which segments words into recurring character sequences rather than treating each word as an atomic unit. When the same word form appears under multiple spellings, it may be segmented into different subword units across instances, producing inconsistent token representations that hinder model learning. Orthographic standardization is therefore important for both word-level and subword-level tokenization. For more information, see Kriukova et al. (2026).

4.1.2 Review of data

To address other inconsistencies in our corpus, such as reduced verb forms, differences in word boundaries, and missing sentence-final enclitics, we used two strategies. The main goal here was to increase the number of fully reviewed transcriptions in our corpus to improve the overall data quality. One strategy, performed by Lovick, involved manually reviewing and standardizing all utterances. This strategy was the most time-intensive, requiring 30–60 minutes per minute of multi-speaker recording and up to 20 minutes per single-speaker recording (see also Amith et al. (2021) for another example of manual data correction). The second strategy, performed by Kriukova, involved partial review of the utterances: fixing typing mistakes in both Dënë Sųłiné and English, correcting word boundaries, and adding missing enclitics. This approach took significantly less time to complete.

As a result, during this study, we completed a partial review, targeting particular tokens and types, of all utterances in the corpus before the last model iteration was trained. While time constraints prevented a complete manual review, over 400 utterances from different speakers were reviewed. Additionally, some previously reviewed transcriptions were added to the training corpus after the first model iteration was trained.

4.1.3 Review of audio clips

Our last step in improving the dataset was the review of audio clips. Our first dataset contained 14,974 audio clips with their corresponding transcriptions. The majority were of satisfactory quality; however, some clips exhibited excessive background noise or speech overlap, rendering individual speakers nearly unintelligible. Since speech overlap is a well-known cause of deterioration in transcription quality (Alharbi et al., 2021; Meng et al., 2024), we excluded or shortened audio clips that contained cross-talk. Furthermore, we removed clips from recordings with poor audio quality to assess the impact on model performance. Since full manual review was not feasible given the corpus size, these recordings were identified through listening to samples, inspecting spectrograms, and number of speakers on each recording, targeting those characterized by excessive background noise and speech overlap.

Nevertheless, during audio cleaning, we removed only speech overlaps and the noisiest segments, adjusting transcriptions when necessary, and limited removal to clips with substantial quality issues (n=129), since Whisper models benefit from varied audio quality (Radford et al., 2023). These removed clips predominantly featured impulsive noise, such as doors opening or closing, objects falling, phones ringing, or children shouting during play.

4.2 Training dataset

As a result of the dataset improvement efforts described in Section 4.1, our training dataset was continuously enhanced and updated. Therefore, each model iteration was trained on a slightly modified dataset. All training utterances could be divided into four categories: 1) reviewed utterances – utterances that were manually reviewed and standardized by Lovick; 2) partly reviewed utterances – utterances that underwent only a partial review by Kriukova; 3) partly standardized utterances–utterances that contain tokens that were standardized through targeted, corpus-wide standardization process, but were not comprehensively reviewed; and 4) unreviewed utterances – utterances

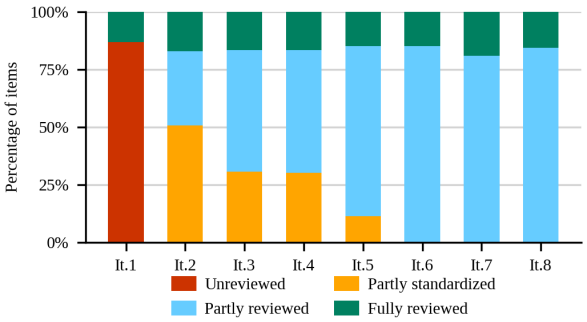

Figure 1: Status of utterances across iterations.

that were neither reviewed nor standardized. The changes in the distribution of utterances' statuses from iteration to iteration are illustrated in Figure 1. The log of the changes made to the dataset is outlined in Table 1.

Iter.	Utts.	Improvements
1	14,974	None
2	16,442	The number of fully reviewed utterances was increased; the first round of standardization was completed.
3	16,936	The number of reviewed and partly reviewed utterances was increased; the second round of standardization was completed.
4	16,807	Bad-quality audio files were removed from the corpus. Some audio files were edited to remove speech overlap.
5	18,615	Newly partly reviewed utterances were added to the corpus.
6	18,615	The partial review was completed for all the utterances. The third round of standardization was completed.
7	19,583	Reviewed recordings of paradigms were added to the corpus.
8	18,779	The fourth round of standardization was completed. Recordings of paradigms were concatenated into longer files.

Table 1: A summary of the ASR model's iterations.

4.3 Testing dataset

To track how changes to the dataset affect model performance, we prepared a dedicated test dataset of 100 utterances (3min 44sec). We decided to keep the testing dataset small, so we could thoroughly review all the test transcriptions and audio files within a reasonable time. Following Liu et al.'s (2023) recommendation to avoid the "hold-speaker out" partitioning method (see Section 3) in low-resource settings, we selected test utterances randomly from different speakers. The number of utterances required for each gender and age group was precalculated to ensure a representative test set. This decision is rooted in the fact that Dënë Sųłınné exhibits significant age-based variation (Jung et al., 2025), and it is important to assess how well the model generalizes across age groups. The division by gender was done to account for the acoustical differences between male and female voices.

Using a Python script, utterances were randomly sampled to preserve both gender and age balance. Each utterance filename encodes a unique three-letter speaker identifier, which the script used to group utterances by speaker and retrieve their corresponding gender and age decade. From each gender–decade group, the maximally even number of utterances was drawn across the 22 speakers, ensuring every speaker is represented in the test set. The resulting test set comprises 50 utterances from female speakers and 50 from male speakers. Speaker age decades were calculated from each speaker's age as of 2025.

After the testing utterances were selected, each recording–utterance pair was reviewed to ensure that the sound quality was good enough for the testing set and that there was no significant voice overlap. The recordings discarded for failing to meet the aforementioned criteria were replaced with randomly selected utterances from the same speaker. All transcriptions were then reviewed by Lovick.

The utterances in the final testing dataset were then tagged for the presence of codeswitching (n=35) and proper nouns (or named entities; n=16) and marked if they were English sentences (n=4). This coding system was applied to determine whether the presence of codeswitching or proper nouns affects Word and Character Error Rates (WER and CER). Additionally, to measure the effect of the audio quality on the transcription results, each recording was tagged as "Good" or "Satisfactory". The tag "Satisfactory" was given in the cases when: 1) the audio from a speaker is faint (usually due to distance from the microphone); 2) low to moderate volume static noise is present in the background (e.g., running water, TV in another room, refrigerator, etc.), 3) low to moderate volume non-speech vocalizations by another participant are present. All other recordings were tagged as "Good" (n=64). To estimate the effect of each tag, we used linear mixed-effects models with the lme4 package (Bates et al., 2015) in R, using lmerTest (Kuznetsova et al., 2017) to

obtain p-values via Satterthwaite's method. WER and CER were modeled as dependent variables. Fixed effects included speaker gender (M/F), audio quality (Quality: Good/Satisfactory), presence of code-switching (CDSW: TRUE/FALSE), presence of proper nouns (NE: TRUE/FALSE), fully English sentences (Full_ENG: TRUE/FALSE), speaker age (calculated from year of birth), and iteration (1–9) (added in R). Speaker (n = 20) and audio clip (n = 100) were included as random effects to account for the repeated-measures structure of the data.

4.4 ASR architecture

Since this study's focus was data-centric, the model architecture for all the iterations described below was identical. All model versions were fine-tuned from the Whisper-medium multilingual ASR model (Radford et al., 2023) with a learning rate of 5e-5, using the model's default subword tokenizer. The fine-tuning was performed on a high-performance computing cluster at the University of Saskatchewan. The GPU used for the training was Nvidia A100 (80G). Training time averaged 24 hours.

5 Results

5.1 Iteration 1

The initial model iteration served as a baseline for measuring improvements in transcription quality in subsequent iterations. This model was trained on 14,974 utterances, of which only 1,974 were fully reviewed, while the rest were normalized but not reviewed or standardized.

The baseline model achieved a WER of 81.2% and a CER of 46.3%. These error rates, notably higher than for major languages (cf. Prud'hommeaux et al., 2021), stem primarily from substantial spelling variation in the training dataset and inconsistent word boundaries between verbs and enclitics. Statistical analysis revealed no significant correlation between the presence of code-switching or proper nouns, or audio quality and the observed error rates. However, fully English utterances demonstrated significantly lower WER (p = .002).

5.2 Iteration 2

The second iteration was trained after our dataset underwent the first stage of orthographic standardization (see Section 4.1.1), which included standardizing the most frequent tokens in the corpus: enclitics, pronouns, some postpositions, and some adverbs. In total, the dataset for this iteration contained 16,442 utterances. For this iteration, we increased the number of reviewed sentences in the dataset from 1,974 to 2,805 by reviewing additional raw utterances and adding one already-reviewed transcript, for which we obtained re-consent after our experiments began. Moreover, 5,321 sentences were partly reviewed. As a result of our corpus manipulations, WER and CER for this iteration improved to 68.6% and 38.5%, respectively. Fully English sentences continued to show significantly lower WER in this iteration (p = .010).

5.3 Iteration 3

Due to time constraints with full reviewing, we prioritized increasing the number of partly reviewed utterances in the third iteration. Additionally, we expanded the dataset from 16,442 to 16,936 utterances by adding new, already partly reviewed utterances and initiated a second round of standardization (see Section 4.1.1). Model retraining revealed that WER improved to 65% in this iteration, while CER remained almost unchanged at 38.6%. Compared to Iteration 2, WER dropped by almost 4 percentage points, indicating that more words were now transcribed correctly. These improvements can be partly attributed to the slightly increased dataset but mainly to the higher number of partly reviewed utterances, which provided a better ratio of standardized tokens in the corpus. Fully English sentences again showed significantly better WER (p = .011).

5.4 Iteration 4

For the fourth iteration, we assessed how removing poor-quality recordings and cleaning of recordings with episodic noise would affect our model performance. We removed 129 audio files and their corresponding transcriptions from the training dataset due to poor quality (see Section 4.1.3). Additionally, 2,177 audio clips were shortened to remove loud noises overlapping with the speakers. As a result, the WER and CER scores for this iteration slightly worsened to 68% and 40.4%, respectively. This decrease in transcription quality can be attributed to a reduced training dataset by 129 (0.7%) deleted utterances. Furthermore, many recordings were shortened, reducing the overall length of the training dataset. A mixed-effects model for this iteration also showed that good audio quality became a significant factor for transcription accuracy,

having lower WER and CER (p = .006, p = .008). The effect of fully English sentences on WER was no longer significant in this iteration.

5.5 Iteration 5

By Iteration 5, an additional set of transcribed files became available for training, following a participant's consent to include their recordings. To compensate for data removed during the audio cleaning process in Iteration 4, we partly reviewed and added 1,810 new transcriptions, thereby increasing the overall proportion of partly reviewed data in the dataset. The testing results for this iteration showed 65.6% WER and 38.7% CER, improving in comparison to both Iterations 3 and 4. These improvements can be attributed entirely to the expanded training dataset and the increased proportion of the partly reviewed data. The significant effect of good audio quality on lower WER and CER remained in this iteration (p = .024 and .045, respectively).

5.6 Iteration 6

For the sixth iteration, we continued experimenting with transcription quality. During this stage, we fully completed the partial review of the utterances, so that the training dataset for this iteration consisted entirely of reviewed and partly reviewed utterances. Additionally, we completed the third round of standardization (see Section 4.1.1), focusing on the remaining frequent adverbs, nouns, and verb forms. Both WER and CER for this iteration decreased slightly to 65.3% and 37.9%, further demonstrating that standardization and review of transcription consistently reduce the model's error rates.

5.7 Iteration 7

For the seventh iteration of the model, we focused on improving the transcription of verbs. To achieve this, we added paradigm recordings (made by Willems (2025)) because their transcriptions were fully reviewed and covered all inflectional forms for multiple verbs. The WER and CER for this iteration increased to 67.25% and 40.1%, respectively. This deterioration in quality was somewhat unexpected. In investigating this increase, we discovered that short, single-word recordings can negatively affect transcription outcomes for longer multi-word recordings (Trabelsi et al., 2024). Given that the Dënë Sųłıné recordings we worked with were predominantly in this format (one verb or verb phrase per recording), the observed deterioration may be attributable to the short duration of the newly added files (see Section 6 for discussion).

5.8 Iteration 8

Although the addition of paradigms in Iteration 7 led to a deterioration in model performance, we did not remove them from the dataset for the fourth and final round of standardization (see Section 4.1.1). Instead, following standardization, we conducted an experiment using two datasets: one with paradigm recordings in their original format (one recording per verb form) and one with verb paradigms concatenated into longer files (4–10 seconds), with silences between individual words reduced to a minimum. The model trained on the dataset with concatenated paradigms yielded lower error rates (WER 63.7%, CER 36.8%) than the model trained on individual paradigm recordings (WER 64.6%, CER 39.8%). Therefore, we decided to use concatenated paradigms, and chose the better-performing model as our result for Iteration 8.

5.9 Observations across iterations

Across all iterations, speaker gender, code-switching, and the presence of proper nouns did not show significant effects on transcription error rates. Fully English utterances in the testing dataset had significantly lower WER in the first three iterations (p = .002, p = .010, p = .011); however, this effect disappeared by Iteration 4. Audio quality significantly influenced transcription error rates only in Iterations 4 and 5 (see Section 5.4 and 5.5).

While individual speakers showed some variation in WER (ranging from -6.9% to +6.0% relative to average), these differences were not statistically significant. Notably, WER and CER were increasing with speaker age in Iterations 5 (p = .045) and 8 (p = .022). However, significant variation existed among individual audio files (SD = 0.25, 95% CI [0.21, 0.29]), indicating that some recordings were inherently more difficult to transcribe regardless of speaker characteristics, audio quality, or other factors.

6 Discussion

In this study, we applied multiple strategies to improve the training dataset, which mostly led to reductions in model error rates (see Figure 1). Below, we first discuss strategies that improved performance, then those that proved ineffective, and

finally examine additional factors that may influence transcription accuracy.

The method that yielded the best results for our model is orthographic standardization (Iter. 2, 3, 6, 8). It reduced inconsistencies in written representations of words, decreased the vocabulary size and allowed the model to generalize more efficiently. From Iterations 1 to 3 only, spelling standardization contributed to decreases of 16.2 and 7.7 percentage points in overall WER and CER, respectively.

Increasing the number of reviewed and partly reviewed utterances (Iter. 2, 3, 5, 6, 7)—either by reviewed existing utterances or adding new verified data—also boosted ASR performance. These dataset reviews helped reduce typing mistakes (in both Dënë Sųłiné and English), made word boundaries more consistent, and slightly decreased spelling variation. For Iteration 5, data reviewing allowed for 2.4 and 1.7 percentage points decreases in WER and CER, respectively. Overall, our systematic data reviewing approach directly enabled performance gains across model iterations.

Our concatenation experiment (Iter. 8) also proved effective, demonstrating that short-format recordings can be retained in the dataset without increasing error rates. Combining paradigm recordings into longer audio files (4–10 seconds) mitigated the negative effects of single-word recordings. This finding may be particularly relevant for under-resourced languages where large portions of existing audio data were recorded in short formats (e.g., for talking dictionaries), though it should be noted that such transcripts are not suitable for language model training.

In contrast, manual audio cleaning and the removal of poor-quality recordings (Iter. 4) proved inefficient, leading to higher WER and CER metrics. Moreover, this process caused our model to perform significantly better on high-quality test recordings for two iterations, thereby reducing its robustness to the sound quality of the fieldwork recordings. These results can be attributed to the Whisper-medium model's documented robustness to varying audio quality (Gong et al., 2023), which appears to enable our fine-tuned model to handle background noise effectively even in low-resource settings. However, Trabelsi et al. (2024) note that for Whisper-base and Whisper-small, transcription quality may benefit from enhanced audio quality in the dataset, at least for English and French. Therefore, we suggest that the decision to enhance audio quality should be guided by the characteristics of the dataset, the size of the Whisper model, and the intended use cases for the fine-tuned model. Given that our model will be used to transcribe fieldwork recordings, which are typically noisier (Liang and Levow, 2025), and may also be deployed in classroom settings, robustness to background noise is essential.

Similarly, retraining the model on the dataset containing paradigm recordings in their original format (Iter. 7) increased error rates, likely due to the single-word-per-recording structure. Although ASR models can be fine-tuned on various utterance types, (Trabelsi et al., 2024) demonstrated that including single-word recordings can negatively affect the recognition of multi-word utterances. The inability to use such recordings—particularly those with high audio quality and verified transcriptions—is undesirable in low-resource settings, which motivated our concatenation approach described above.

Beyond these dataset optimization strategies, we examined whether specific linguistic and speaker characteristics affected transcription accuracy. Based on our earlier experiments with Whisper-small conducted before this study, we anticipated that proper nouns and code-switching would pose challenges for the model. Specifically, Whisper-small had difficulties recognizing Saskatchewan- and Canada-related proper nouns (e.g., Turnor Lake, Ile-a-la-Crosse, Manitoba), and speech with code-switching to English often resulted in fully English transcriptions. However, contrary to our expectations, neither factor significantly affected transcription accuracy in any Whisper-medium iteration. For instance, out of 76 English words in the test set 66 were transcribed as English words (though not always correctly), 8 were transcribed as Dënë Sųłiné words (e.g., *nowadays* → *now dë́* '(lit.) now if/when'), and one was omitted entirely. In the reverse direction, only two Dënë Sųłiné words were transcribed as English (e.g., *bazë́* 'regarding' → *pause*. This improvement in code-switching detection, along with better transcription of proper nouns, may be attributable to the more optimal model size and the larger training set used in this study.

Fully English sentences had significantly better transcriptions during the first three iterations, but later this effect had disappeared. It is most likely related to the fact that, after some standardization and data reviewing, the model improved at transcribing Dënë Sųłiné to the point where fully English

sentences were no longer advantaged. Speaker gender did not affect transcription quality in any iteration, despite the training dataset's gender imbalance (68% female speakers). However, speaker age showed a sporadic effect on WER, reaching significance only in Iterations 5 and 8. The changes made to the dataset before these iterations were unlikely to have triggered this effect. Hence, given the inconsistent pattern across iterations and the absence of a theoretical explanation for such a result, this may reflect a Type I error due to multiple comparisons rather than an actual effect. Further testing on a larger dataset is required to determine whether speaker age really affects WER.

7 Conclusions

Overall, this study contributes to the growing discourse on data-centric approaches for low-resource datasets and aims to inspire further exploration of computational methodologies for enhancing noisy, resource-constrained data. While the importance of data quality is widely acknowledged, our work provides quantitative evidence of exactly how much data-centric approaches can improve model training outcomes. Our findings demonstrate that effective ASR dataset preparation for an under-resourced language should prioritize quality of transcriptions and spelling standardization (if applicable) over audio enhancement. Crucially, out results suggest that the researchers working with imperfect audio recordings may use them for fine-tuning the Whisper-medium model without dramatically compromising performance. These insights require validation across other multilingual pre-trained ASR architectures, including Wav2Vec2 (Baevski et al., 2020), to establish their broader applicability.

As Jimerson et al. (2018) note, under-resourced communities need to make informed decisions about how to invest their resources when developing ASR tools for their languages. This also applies to time investment during dataset preparation: most language communities have unlimited time and financial resources for data curation, making it essential to identify which preprocessing steps lead to the best results. Through this study, we hope to have shown which aspects of dataset preparation should be prioritized, particularly for languages with mixed-quality datasets. Moreover, our work has revealed specific features of Whisper's behaviour—such as the minimal optimal recording length in training datasets—that are relevant to decisions about dataset structure. We hope that these insights will help language communities, linguists, and developers prioritize their dataset development efforts more efficiently.

In our work on the *Talking Dene* corpus, we plan to continue standardizing orthography to further improve transcription quality. We also plan to investigate the model's feasibility for retranscribing yet unstandardized transcriptions to reduce the corpus-reviewing workload and accelerate processing of the *Talking Dene* corpus. Finally, we plan to test the model with Dënë Sųłıné speakers and assess its effectiveness for real-life transcription scenarios.

Limitations

This study has several limitations that should be acknowledged. First, our orthographic standardization efforts require continuation to yield more statistically significant and comprehensive results. The current findings, while promising, represent only an initial exploration. Second, our analysis was conducted using only one ASR architecture: the Whisper-medium acoustic model. To establish the generalizability of our findings, validation across other state-of-the-art ASR architectures, such as Wav2Vec2, is necessary. The performance patterns observed here may be model-specific and require broader empirical validation. Finally, this study did not provide solutions to the out-of-vocabulary (OOV) problem arising from a combination of the morphological richness of Dënë Sųłıné verbs and orthographic inconsistencies in their spelling, as addressing this challenge was beyond the scope of the current work. However, we recognize this as a critical limitation and plan to investigate potential solutions in future research.

Ethical considerations

This study is approved by the University of Saskatchewan Board of Ethics (Beh-REB-4918). All speakers participating in this research gave their explicit consent for the use of their audio recordings for the Dënë Sųłıné Automatic Speech Recognition model training. The model cannot be made publicly available until the Clearwater River Dene Nation and La Loche communities decide how they want to distribute it.

Acknowledgments

We are grateful to the Clearwater River Dene Nation and La Loche (SK, Canada) Dene communities for the opportunity to work with their language. We especially want to thank the research assistants from the Clearwater River for their help in the data collection and transcription for this study: Trina Lemaigre and Chastity Sylvestre. Moreover, we want to thank all participants, whose recordings were used for the training of the Automatic Speech Recognition model (some referred to by pseudonym): Rebecca Dene, Teresa Dene, Mitchell Guetre, Gerald E. Haineault, Brenda Herman, Rhonda Herman, Sharon Kennedy, Allison Lemaigre, Andrea Lemaigre, Antoinelle Lemaigre, Edainya Lemaigre, Jeannie Lemaigre, Jennifer Lemaigre, Johny Lemaigre, Mikki Lemaigre, Miranda Lemaigre, Randall Lemaigre, Taitlyn Lemaigre, Taylon Lemaigre, Tina Lemaigre, Trina Lemaigre, Tyanne Lemaigre, Doreen Moise, Ernie Piche, Heather Piche, Ursula Piche, and Jeff Toulejour. We also want to thank Nial Willems for his help with verb-paradigm reviewing and for providing his reviewed transcriptions and recordings for this study. This study was funded by the SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages". The data collection for the *Talking Dene* study was funded by the SSHRC Insight Grant 435-2020-1197.

References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively Multilingual Adversarial Speech Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Altruki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9.

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-end automatic speech recognition: Its impact on the workflow in documenting Yoloxóchitl Mixtec. In *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222. European Language Resources Association.

Alexei Baevski, Henri Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on Neural Information Processing Systems*, pages 12449–12460, Vancouver, Canada.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Nikita Bhatt, Nirav Bhatt, Vishal Sorathiya, Samah Alshathri, and Walid El-Shafai. 2024. A data-centric approach to improve performance of deep learning models. *Scientific Reports*, 14.

Eung-Do Cook. 2004. *A grammar of Dëne Sųłiné (Chipewyan)*. Number 17 in Algonquian and Iroquoian Linguistics. University of Manitoba, Winnipeg.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3863–3867, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of Interspeech 2023*, pages 2798–2802.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering*, 66:507–515.

Robbie Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1008–1016.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for Documenting Acutely Under-Resourced Indigenous Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Dagmar Jung, Olga Lovick, Alison Lemaigre, Jakaterina Mazara, and Olga Kriukova. 2025. Mixed constructions across ages: Comparing two Dene corpora [Conference presentation]. Presented at the Conference of the Society for the Study of the Indigenous Languages of the Americas (SSILA), January 2025.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proceedings of the 6th Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55.

Olga Kriukova, Gabrielle Fontaine, Alison Lemaigre, Dagmar Jung, Antti Arppe, and Olga Lovick. 2026. Using automatic speech recognition to assist with standardization of Dënë Sųłıné transcripts. *(Submitted)*.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1–26.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. *arXiv preprint*. ArXiv:2506.17459 [cs].

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves NLP performance for Nigerian pidgin. In *LREC-COLING 2024*, pages 11510–11522.

Zoey Liu, Crystal Richardson (Karuk), Richard Hatcher Jr, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:3933–3944.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131.

Olga Lovick, Dagmar Jung, Olga Kriukova, Allison Lemaigre, and Barb Hannah. 2023. Variation and change across generations in current Dene: Reduction in verbs [Conference presentation]. Presented at the Annual Conference of the Canadian Linguistic Association, Toronto, Canada, May 2023.

Olga Lovick, Dagmar Jung, Olga Kriukova, Allison Lemaigre, and Barb Hannah. 2024. Variation and change in Dene verbs [Conference presentation]. Presented at the Conference of the Society for the Study of the Indigenous Languages of the Americas (SSILA), New York, USA, January 2024.

Sarah Luger, Rafael Mosquera-Gómez, Alex Miłowski, Thom Vaughan, Sara Hincapie-Monslave, Pedro Ortiz Suarez, and Kurt Bollacker. 2025. Building data infrastructure for low-resource languages. In *Proceedings of the 8th Workshop on Technologies for Machine Translation of Low-Resource Languages*, pages 154–160.

Lingwei Meng, Jiawen Kang, Yuejiao Wang, Zengrui Jin, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Empowering Whisper as a joint multi-talker and target-talker speech recognition system. In *Proc. Interspeech 2024*, pages 4653–4657.

Patrick James Moore. 2018. Re-valuing code-switching: Lessons from Kaska narrative performances. In Julia Christensen, Christopher Cox, and Lisa Szabo-Jones, editors, *Activating the heart: storytelling, knowledge sharing, and relationship*, pages 53–88. Wilfrid Laurier University Press.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

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 the 40th International Conference on Machine Learning*, pages 28492–28518.

Zarif al Sadeque. 2022. Automatic speech recognition for documenting endangered First Nations languages. Master's thesis, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not

the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Mark Simmons. 2025. Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper. In *Proceedings of the 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 155–161, Honolulu, HI, USA.

Statistics Canada. 2021. Mother tongue by geography, 2021 Census.

Asma Trabelsi, Laurent Werey, Sebastien Warichet, and Emanuel Helbert. 2024. Is noise reduction improving open-source ASR transcription engines quality? In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*, volume 3, pages 1221–1228.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 993–1003. Association for Computational Linguistics.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, 32:791–813.

Nial Austen Willems. 2025. The ts'ë- passive in Dëne Sųtłıné. Master's thesis, University of Saskatchewan, Saskatoon, Canada.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 306–315, Marseille, France. European Language Resources association.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 5071–5079, Marseille, France.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research. *arXiv preprint*. ArXiv:2412.04497 [cs] version: 2.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced Aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 4004–4011, Slovenia.