Development of Burushaski Speech – English Text Translation Dataset

Tauqeer Saleem, Dr Abdul Samad, Azkaa Nisar, Fatima Faisal, Adina Adnan Mansoor, Mahrukh Yousuf

Habib University
Karachi, Pakistan
{tauqeer.saleem, abdul.samad}@sse.habib.edu.pk, {an08017, ff08435, am08349, my08055}@st.habib.edu.pk

Abstract

Burushaski is a language isolate spoken in northern Pakistan with a predominantly oral tradition, limited standardized orthography, and virtually no existing speech technology infrastructure. These characteristics make conventional text-centric NLP pipelines unsuitable and position speech data collection as the primary scientific challenge. This paper introduces an audio-first, linguistically informed methodology for developing a Burushaski–English speech translation resource. Rather than prioritizing model architecture, we focus on principled corpus design tailored to the language's morphological complexity, ergative-absolutive alignment, and four-gender agreement system. The dataset combines structured elicitation targeting high-frequency and morphologically diverse constructions, functional and formulaic speech, and oral narratives that capture discourse-level phenomena.

We describe the design of a custom data collection application, community-embedded crowdsourcing strategy, and translation-aligned workflow for generating parallel speech–English data. The resulting pilot corpus comprises approximately 10 hours of curated audio from 42 speakers across controlled and naturalistic settings.

While we present the results of preliminary translation experiments using Whisper, the primary contribution of this work is methodological: a scalable framework for speech-first corpus development in morphologically rich, under-resourced, and predominantly oral languages. We argue that for languages lacking stable orthography and large textual corpora, data design, not model selection constitutes the central research problem

Keywords: Burushaski, Translation, Textless NLP

1. Introduction

Recent advances in self-supervised speech modeling have significantly reduced the dependence of automatic speech recognition (ASR) and speech translation systems on large manually transcribed corpora. Models such as wav2vec 2.0 and XLS-R demonstrate that robust acoustic representations can be learned from raw audio at scale. However, these advances do not eliminate a more fundamental bottleneck: the absence of well-designed, representative speech datasets for languages that lack standardized orthography, large textual corpora, and existing speech technology infrastructure. For predominantly oral and morphologically complex languages, corpus design remains the central challenge.

Burushaski, spoken in the Hunza, Nagar, and Yasin valleys of northern Pakistan, exemplifies this setting. As a language isolate with a primarily oral tradition and limited orthographic standardization, Burushaski has minimal digitized textual resources and virtually no publicly available speech corpora. With an estimated 90,000 -120,000 speakers across three mutually intelligible but structurally distinct dialects - Hunza, Nagar, and Yasin. While multilingual self-supervised models may offer cross-lingual acoustic transfer, effective downstream adaptation still requires carefully structured language-specific data. In addition, Burushaski exhibits typological properties, such as including ergative-absolutive alignment, a four-gender noun class system, and rich verbal morphology that increase data sparsity under naive collection strategies. Opportunistic or read-speech corpora risk underrepresenting inflectional paradigms and discourse patterns essential for generalizable modeling.

In such contexts, the primary scientific question shifts from model architecture to dataset construction. How should speech data be elicited, structured, and validated to ensure morphological coverage, phonetic diversity, and functional relevance for downstream translation tasks? How can parallel speech–translation resources be created when textual norms are unstable and literacy practices vary across speakers? Addressing these questions requires integrating linguistic field methodology, speech technology constraints, and community-embedded collection practices.

This paper presents an audio-first, linguistically informed framework for developing a Burushaski–English speech translation resource. Rather than prioritizing architectural novelty, we focus on principled corpus design tailored to the structural properties of the language and the practical realities of its speaker community. Our approach combines (i) structured elicitation targeting high-frequency vocabulary and systematic coverage of

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 9–17

tense–aspect–mood and agreement paradigms, (ii) collection of functional and formulaic language relevant to real-world applications, and (iii) oral narratives and folktales capturing discourse-level phenomena and natural variation. Data collection is facilitated through a custom mobile application designed to standardize prompts while enabling scalable community participation.

The resulting pilot corpus comprises approximately ten hours of curated Burushaski speech paired with English translations, collected from 42 speakers across controlled and semi-naturalistic settings.

We report corpus design principles, speaker distribution, and linguistic coverage characteristics, and we outline baseline experiments using Whisper for translation. These modeling experiments are exploratory; the principal contribution of this work lies in establishing a replicable methodology for speech-first corpus development in morphologically rich, low-resource languages with limited textual infrastructure. Concretely, this paper makes the following contributions: (1) the first parallel Burushaski–English speech corpus, comprising approximately 10 hours of audio from 42 speakers across controlled and semi-naturalistic settings; (2) a custom mobile data collection application designed for low-infrastructure, low-literacy field settings; and (3) baseline translation experiments with Whisper establishing a reference point for future work.

2. Related Work

2.1 Speech Technology in Low-Resource and Predominantly Oral Languages

Recent advances in self-supervised learning have reshaped low-resource speech technology. Models such as Whisper, wav2vec 2.0 (Baevski et al,. 2020) and XLS-R (Conneau et al., 2021) demonstrate that acoustic representations can be learned from large volumes of unlabeled speech and subsequently fine-tuned for downstream tasks, including automatic speech recognition (ASR) and speech translation. Multilingual pretraining has further enabled cross-lingual transfer, allowing models trained on high-resource languages to adapt to lower-resource settings with comparatively limited labeled data.

However, the effectiveness of such transfer depends on the availability of at least modest quantities of representative speech data for fine-tuning. For languages with unstable orthography, limited digitized text, and little existing speech infrastructure, the primary bottleneck is not model architecture but the absence of systematically designed corpora. Predominantly oral languages introduce additional challenges: read-speech datasets may not reflect natural prosody or discourse patterns, and inconsistent orthographic practices complicate alignment and evaluation. These constraints shift the focus from model selection to data design and collection methodology.

2.2 Corpus Design and Linguistic Coverage in Morphologically Rich Languages

Low-resource speech performance is strongly influenced by the linguistic coverage of the underlying dataset. In morphologically rich languages (MRLs), surface form variation increases data sparsity: a single lexical root may appear across numerous inflectional forms encoding case, agreement, tense, aspect, or mood. Opportunistic corpus collection—such as harvesting spontaneous or read text without structural planning—often underrepresents critical morphological paradigms, limiting generalization.

Prior work highlights different strategies to address this issue. The Mboshi corpus described by Godard et al. (2017) integrated language documentation practices with computational alignment, collecting speech paired with French translations using mobile elicitation tools. Their approach prioritized natural speech while maintaining parallel structure suitable for downstream modeling. In contrast, phonetically balanced corpus construction efforts (e.g., Gupta et al., 2016) emphasize systematic coverage of phonemes and triphones to improve acoustic robustness.

These approaches demonstrate that linguistic planning—whether targeting phonetic coverage or grammatical paradigms—is essential in low-resource settings. Yet relatively little work formalizes elicitation strategies that simultaneously target (i) morphological coverage, (ii) high-frequency functional vocabulary, and (iii) discourse-level naturalness within a unified speech-translation framework. For languages with complex alignment systems and agreement structures, corpus design must explicitly address paradigm diversity rather than rely solely on frequency-based sampling.

2.3 Speech-to-Text Translation in Low-Resource Setting

End-to-end speech-to-text translation (ST) systems have gained prominence as alternatives to cascaded ASR-plus-machine-translation pipelines. Multilingual encoders combined with Transformer decoders have demonstrated competitive performance in several low-resource scenarios when supported by parallel speech–text data. Multilingual pretraining has been shown to reduce supervision requirements, particularly when typologically related languages are included in the training distribution.

However, most low-resource ST studies assume the existence of standardized orthography and

curated parallel corpora. For predominantly oral languages, creating aligned speech–translation data is itself a nontrivial task. Orthographic instability complicates transcription consistency, while limited literacy may constrain text-based data collection workflows. As a result, methodologies for constructing parallel speech–translation datasets remain underdeveloped for languages lacking established textual norms.

In such cases, translation-aligned elicitation—where speech is collected directly with corresponding translations—offers an alternative pathway that bypasses dependence on large native-language text corpora. The effectiveness of this approach depends on careful prompt design and validation to ensure linguistic diversity and functional relevance.

2.4 Community-Embedded Data Collection and Crowdsourcing

Crowdsourcing has become a common strategy for scaling speech data collection. However, studies have shown that conventional online platforms systematically underrepresent speakers of low-resource and rural languages due to infrastructure, literacy, and accessibility constraints. Field-based and community-embedded approaches can mitigate these limitations by tailoring tools and workflows to local contexts.

Mobile data collection applications have been successfully used in documentation-oriented corpora to standardize prompts while preserving naturalistic speech production. Translation-based elicitation further enables the creation of parallel corpora without requiring extensive written materials in the target language. Nevertheless, ensuring speaker diversity, gender balance, and dialectal variation remains an ongoing challenge, particularly in small pilot datasets.

2.5 Positioning of the Present Work

Existing research demonstrates the value of self-supervised speech encoders, linguistically informed corpus design, and translation-aligned data collection. Yet there remains a methodological gap in integrating these strands for morphologically complex, predominantly oral languages that lack stable orthography and prior speech infrastructure.

The present study addresses this gap by formalizing an audio-first corpus design framework for Burushaski that:

1. Systematically targets morphological paradigms (ergative-absolutive alignment, agreement, tense–aspect–mood distinctions);
2. Incorporates high-frequency functional and formulaic language relevant to practical translation scenarios;
3. Integrates oral narratives to capture discourse-level phenomena and natural variation; and
4. Embeds data collection within a structured mobile application workflow enabling scalable community participation.

Rather than proposing a novel model architecture, this work treats dataset construction as the primary research contribution and evaluates baseline multilingual encoders within that framework. By foregrounding corpus design as a central methodological problem, the study contributes a replicable approach for developing speech-translation resources in similarly under-documented and typologically complex languages.

3. Dataset Collection

Developing speech resources for a predominantly oral and morphologically rich language requires deliberate corpus design rather than opportunistic data gathering. The dataset collection strategy for Burushaski was therefore structured around three principles: (i) systematic linguistic coverage, (ii) functional relevance for translation tasks, and (iii) scalable community participation through controlled prompting.

3.1 Corpus Design Framework

The dataset was constructed using a category-based elicitation protocol defined in a master spreadsheet of English–Burushaski sentence pairs. Our approach systematically targets Burushaski's morphological complexity through stimulus design informed by standardized linguistic questionnaires (Lahiri 2018, Max Planck Field Manuals), generating parallel text-audio alignments essential for textless pipeline training and evaluation. Sentences were grouped into linguistically motivated categories, including:

- Simple and complex sentence structures capturing the basic syntactic organization of the language
  (e.g., "The monkey is on the tree.")
- Case markers that target ergative, absolutive, and other case distinctions
  (e.g., "She (right here) hit the bear.")
- Tense, aspect, and mood contrasts that elicit inflectional variation across temporal and aspectual paradigms
  (e.g., "I am walking to the field now.")
- Negative constructions capturing negation morphology across sentence types
  (e.g., "There is no water in the river.")
- Reflexive constructions targeting self-referential agreement patterns
  (e.g., "We saw ourselves in the mirror.")
- Serial and converb verb constructions (ECV) capturing multi-predicate and chained verb structures
  (e.g., "She has gone outside the house.")

- Noun classifiers eliciting Burushaski's gender-class agreement system *(e.g., "People have gathered here.")*
- Interrogatives covering both content and polar question formation *(e.g., "Where are you going?")*
- Pronouns targeting person, number, and gender paradigms *(e.g., "You are late.")*
- Valency alternations contrasting transitive and intransitive verb frames *(e.g., "He opened the door." vs. "The door opened.")*
- Causative constructions eliciting morphological and periphrastic causation *(e.g., "Mother made me go outside.")*
- Possession constructions covering attributive and predicative possession *(e.g., "I have a book.")*
- Adpositions and postpositions capturing spatial and relational marking *(e.g., "She is in the room.")*
- Passive constructions targeting voice alternations and agent demotion *(e.g., "The door was opened by the boy.")*
- Imperative constructions eliciting directive speech acts *(e.g., "Sit down.")*
- Naturalistic narrative data capturing discourse-level phenomena and spontaneous speech variation *(e.g., "What is the climate like where you live?")*

Rather than collecting isolated lexical items, each category was designed to trigger controlled variation within grammatical paradigms. For example, minimal pairs such as:

- "She (right here) saw a man."
- "She (not here) saw a man."

These two sentences are identical in meaning, The only difference is whether the person, "she" is present near the speaker or not. In Burushaski, this distinction is grammatically marked, the verb and pronoun forms change depending on the location of the subject relative to the speaker. This small variation can reliably elicit and isolate the specific morphological forms Burushaski uses to encode deictic contrast. This controlled design ensures that these distinctions are captured across multiple sentence contexts rather than appearing only once in the corpus, reducing sparsity in the training data.

Similarly, parallel prompts varying subject gender or participant role target inflectional alternations without altering lexical content. This structured variation reduces morphological sparsity by ensuring that inflectional paradigms are represented across multiple lexical roots.

The design deliberately balances frequency-based vocabulary with morphosyntactic coverage. High-frequency functional language (e.g., kinship terms, classroom interactions, temporal expressions) ensures relevance for real-world translation scenarios, while targeted paradigm elicitation supports downstream modeling robustness. For further robustness, we have also used Common Voice's dataset for Burushaski[1], this dataset had only audios and transcript without English translation, so we asked volunteers to translate it for us. The source dialect for this dataset is that of Hunza.

3.2 Translation-Aligned Elicitation

Given the limited availability of standardized written Burushaski, the collection process followed a translation-aligned workflow. English prompts were used as structured stimuli which target frequency based everyday vocabulary across general contexts (Kilgarriff et al., 2014; Finlayson et al., 2024), and native speakers produced corresponding Burushaski speech recordings. This approach ensures parallel speech–English data without requiring large native-language textual corpora.

To increase ecological validity, prompts were written in simple, contextually plausible English rather than artificial grammatical templates. Where possible, multiple speakers recorded the same prompt to capture phonetic and prosodic variation while preserving semantic alignment.

In addition to structured elicitation, selected folktale excerpts and narrative prompts were included to capture discourse-level phenomena, natural prosody, and code-switching behavior. This hybrid design combines controlled grammatical targeting with naturalistic speech segments.

To ensure the validity of the translations, we assigned human validators to ensure quality and accuracy. Each Burushaski recording was reviewed by 3 independent native-speakers, who verified that the spoken output accurately and naturally conveyed the meaning of the original English prompt. This validation pipeline was applied consistently across both the structured elicitation material and the narrative/folktale segments, ensuring dataset-wide translation fidelity.

3.3 Mobile Application Infrastructure

To standardize collection and enable distributed participation, we developed a mobile recording

[1] Bsk Common Voice Dataset: https://datacollective.mozillafoundation.org/datasets/cmj8u3owx003xnxxb3f7n3m1l

application that presents prompts sequentially and records corresponding speech in similar fashion to Godard et al. (2017). Each prompt appears with:

- An English stimulus sentence
- Its Burushaski equivalent (where applicable)
- A recording interface
- Submission and validation feedback – volunteers and research assistants were part of the validation process to ensure that the given Burushaski equivalent is correct and that the community is involved at each step of the data collection process.

The application enforces uniform recording procedures, reducing variability introduced by heterogeneous recording setups. Each submission is logged with metadata including speaker identity (anonymized ID), sentence ID, and timestamp. Participants ranged in age from approximately 18 to 60 years, with the majority recruited from Hunza and Nagar. Quality assurance was ensured within the community workflow at two distinct stages rather than relying solely on post-hoc expert review.

Stage 1 – Prompt Validation (pre-recording) Before any prompt was made available to speakers, research assistants and community volunteers confirmed the Burushaski equivalent displayed alongside each English sentence was accurate and natural. Items flagged as incorrect or unidiomatic were withdrawn, corrected, and re-reviewed before re-release. This ensured no speaker was provided with erroneous or unnatural Burushaski text.

Stage 2 – Post-recording audio review. Submitted recordings were reviewed whether (a) the spoken utterance conveyed the meaning of the English prompt (b) was grammatically natural and fluent, and (c) audio quality free for noise, clipping or truncation.

Speaker representation across the three primary dialect regions - Hunza, Nagar, and Yasin was uneven in this initial phase, with Yasin speakers significantly underrepresented, broadening dialectal coverage is a priority for future collection rounds.

This structured interface serves two purposes:

1. Ensuring consistent prompt delivery across participants.
- Enabling scalable, community-driven data collection without requiring advanced technical literacy.

3.4 Ethics, Consent and Data Governance

All participants provided informed consent prior to recording. Consent and demographic information were collected digitally via a structured Google Form. The form captured participant age, gender, dialect region alongside explicit consent for the use of recordings in research. Participation was entirely voluntary and participants were informed they could withdraw at any time. Speaker identities are retained internally only as anonymized IDs; no personally identifiable information is included in the released dataset. The Google forms responses are stored securely and separately from the audio data, accessible only to the research team.

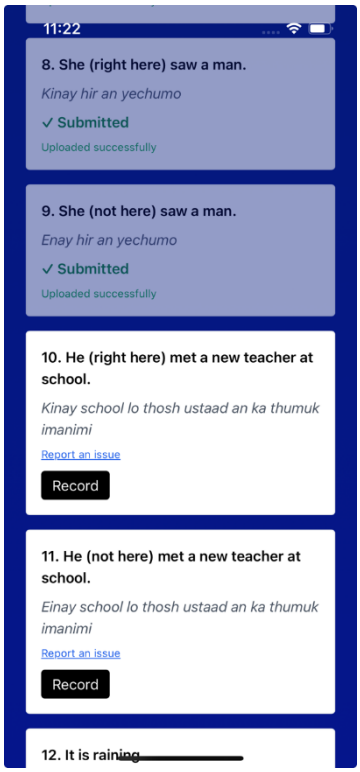

Figure 1: Screenshot of the mobile application for data collection

4. Translation Model – Whisper

Whisper uses a Transformer-based encoder to process log-Mel spectrogram inputs, and a Transformer decoder to autoregressively generate text tokens. The model supports both speech-to-source transcription and direct speech-to-English translation. In this work, we used direct speech-to-English translation mode, bypassing intermediate Burushaski transcription.

We fine-tuned three variants of Whisper: Whisper-Small (~244M parameters), Whisper-Medium (~769M parameters) and Whisper-Large-v2 (~1.55B parameters). All experiments used the same dataset of 7,115 examples drawn from a

combined pool of Mozilla Common Voice Burushaski recordings and app-collected data, which were merged into a single corpus prior to splitting. The dataset was partitioned using a fixed random seed (seed=42) into a training set of 6,403 examples (90%) and a validation set of 712 examples (10%), with a held-out test set of 1,779 samples used exclusively for final evaluation.

Pre processing and data preparation was the same for all three models. Audio was resampled at 16kHz and converted into 80-channel log-mel spectrograms using Whisper's default feature extractor. English target translations were lowercased and stripped of punctuation and excess whitespace before tokenization. Audio durations ranged from 0.5s to 14.3s (avg: 3.2s), and English reference lengths ranged from 1 to 41 words (avg: 5.4 words).

The training parameters for Whisper-Small and Whisper-Medium were the same, with only slight modifications made for Whisper-Large-v2-v2 because of GPU memory limitations. Every model was trained using a learning rate of 1e-5, a warmup of 200 steps, mixed precision training (fp16), a batch size of 8 per device with two gradient accumulation steps, and gradient checkpointing enabled.

Fine-tuning was performed using paired speech–English translation targets. No intermediate ASR step or external machine translation system was used; the model was trained end-to-end for direct translation. Given (a) lack of standardized Burushaski orthography, (b) limited transcription resources, (c) desire to avoid cascading ASR error propagation, direct speech-to-English translation was selected over ASR+MT pipelines. This approach removes dependence on consistent native-language text while aligning with Whisper's pretraining objective. The Mozilla Common Voice recordings and app-collected data were not treated as separate streams during training; both sources were merged into a single dataset and shuffled before the train/test/validation split was applied, ensuring that both data sources were represented across all splits.

5. Results

5.1 Dataset Summary

We have successfully started the process of gathering data that is representative of the language, the process will continue and deliver better quality data in the coming years. More sentence categories will be added, and more literature will be added. Table 1 summarises the pilot corpus collected to date. The 5,095 application recordings from 42 speakers cover 511 unique prompt sentences, yielding an average of approximately 10 recordings per sentence, sufficient for capturing phonetic and prosodic variation across speakers for most prompts. The gender distribution is notably skewed, with female speakers contributing 3,642 recordings (71%) compared to 1,453 from male speakers (29%), this imbalance reflects recruitment constraints in the pilot phase and is a known limitation of the current dataset. The Mozilla Common Voice component (~6 hours) contributes naturalistic speech, though its dialectal distribution is uncontrolled. Combined, the ~10-hour corpus represents the largest known Burushaski–English speech resource, though it remains small relative to the data requirements of standard neural translation systems

Public release is planned following completion of data validation and speaker consent review.

Total Number of Recordings (App)	5095
Total Duration (App)	~4 hours
Participants	42
Unique Sentences	511
Average Utterance Length	~3 seconds
Average Sentences per Category	~30
Dialect Distribution	Hunza: 4,867 (95.5%) Yasin: 228 (4.5%)
Gender distribution of number of recordings	Male: 1,443 (28%) Female: 3,642 (72%)
Average time per participant	121.3 seconds (~2 minutes)
Mozilla Dataset Translated	~6 hours
Total Data in hours	**~10 hours**

Table 1: Summary of data collected

5.2 Translation

Evaluation was conducted on 244 held-out test samples using standard speech translation and transcription metrics.

All three Whisper variants were evaluated on the same held-out test set of 1,779 samples. Decoding was performed using default greedy generation with a maximum output length of 225 tokens. Before metric computation, both predictions and references were normalized by lowercasing and removing punctuation and excess whitespace. BLEU and chrF++ were computed on normalized text; BERTScore F1 was computed on the original unnormalized text using bert-base-multilingual-cased; WER and CER were computed on normalized text using the jiwer library.

Metric	Whisper-Small	Whisper-Medium	Whisper-Large-v2
WER	34.33%	25.71%	24.54%
CER	25.94%	20.90%	19.64%
Word Accuracy	65.67%	74.29%	75.46%
BLEU	65.42	69.94	70.90
chrF++	71.19	77.15	78.24
BERTScore F1	0.9195	0.9375	0.9402

Table 2: *Evaluation results across Whisper model variants on the held-out test set (n=1,779). Lower is better for WER and CER, higher is better for all other metrics.*

The model scale consistently improves the results for all three variants. With a BLEU of 70.90, WER of 24.54%, and BERTScore F1 of 0.9402, Whisper-Large-v2-v2 performs the best. It is closely followed by Whisper-Medium (BLEU 69.94, WER 25.71%) and Whisper-Small (BLEU 65.42, WER 34.33%). Predictions are both lexically close to the references and semantically coherent, which is a significant improvement over the pilot baseline, as evidenced by the comparatively small difference between BERTScore and BLEU across all models.

Furthermore, we did a qualitative error analysis with the help of native Burushaski speakers and realized the reason why there is a discrepancy between BERTScore and BLEU.

1. We found that in several cases Whisper was attempting to get semantic approximation but fails to preserve culturally specific lexical items. For example:
   Reference: “It is the story of Pegasus and a jinn”
   Prediction: “It is the story of the Wind Horse and the Ghost”
2. In other cases, model demonstrated semantic compression, where meaning is preserved but structure differs—contributing to high BERTScore but elevated WER. For example:
   a. Reference: "The wind has been blowing since yesterday."
   b. Prediction: "It has been windy since yesterday."
3. There were several cases where named entities are replaced with semantically plausible but incorrect phrases, reflecting uncertainty in acoustic mapping. For example:
   a. Reference: "I am standing between Ammy and Mary."
   b. Prediction: "I am standing between the two rivers."

The results reflect the tension between large-scale multilingual supervision and language-specific adaptation:

- Whisper’s pretraining enables semantically coherent English output.
- However, absence of Burushaski exposure during original training leads to lexical instability and hallucination.
- Fine-tuning on ~10 hours give a strong translation performance, with the remaining errors concentrated in culturally specific expressions and named entities rather than general fluency.

These findings reinforce the central thesis of this study: strong performance is achievable with carefully structured parallel data, and remaining limitations point to the need for broader cultural and lexical coverage rather than architectural changes.

6. Future Work

The present work establishes a methodological foundation and baseline system for Burushaski speech-to-text translation; several directions for improvement are planned for future iterations.

On the data side, the corpus will be expanded substantially through continued community-embedded collection. Additional sentence categories will be added to the elicitation protocol, with particular attention to underrepresented morphological paradigms and domain-specific vocabulary. Natural conversational data and extended oral narratives , including informal dialogues and unscripted community exchanges, will be collected to complement the currently elicitation-heavy corpus. This will address the prosodic and discourse-level gaps inherent to prompted elicitation, following the approach of Godard et al. (2017), who demonstrated that integrating spontaneous and controlled speech yields more generalizable ASR performance. Speaker diversity will also be broadened to achieve better coverage across age groups, genders, and the three primary dialect regions of Hunza, Nagar, and Yasin. The translation of additional Burushaski literary and community texts will provide supplementary parallel data outside of the elicitation framework.

On the modeling side, several more advanced architectures will be explored as the dataset grows. The prosody-aware Generative Spoken Language Model (pGSLM) of Kharitonov et al. (2021) extends the GSLM framework by explicitly modeling duration and pitch alongside discrete unit sequences using a multi-stream Transformer, enabling richer and more natural speech generation. As our corpus expands, this architecture may offer a more expressive

alternative to the basic GSLM pipeline we initially tested. Additionally, the textless speech-to-speech translation framework of Lee et al. (2021) demonstrated that direct speech-to-speech translation is achievable without intermediate text representations, which is particularly relevant for Burushaski given its orthographic instability. Exploring this framework on our data could provide a fully text-free translation pathway. Finally, AudioChatLlama (Fathullah et al., 2023) offers a promising direction for integrating spoken Burushaski input directly with a large language model decoder, enabling zero-shot speech question answering and translation through audio-conditioned LLM inference. As multilingual LLMs continue to scale, this approach may offer meaningful gains even for isolate languages that remain absent from pretraining data, particularly given Burushaski's natural code-switching with Urdu and English.

Evaluation methodology will also be refined. Current assessment relies on standard metrics such as BLEU; future work will complement these with human evaluations by native Burushaski speakers and linguists and will explore character error rate (CER) metrics on any phonemic transcriptions produced as an intermediate step. Targeted evaluation of culturally specific expressions, including idioms and proverbs, will be conducted through structured human evaluation protocols involving native speaker judges. Taken together, these directions position the present corpus and baseline system as a scalable starting point for an ongoing Burushaski speech technology program.

7. Conclusion

This paper presented the first structured effort toward building a curated Burushaski–English parallel speech corpus for direct speech translation. Rather than positioning model performance as the primary contribution, we focused on documenting a reproducible data collection methodology tailored to a severely under-resourced, predominantly oral language. The resulting dataset, though modest in size, establishes an initial benchmark and experimental foundation for future work in Burushaski speech processing.

Our experiments whisper model demonstrates that large pretrained multilingual models can produce semantically coherent English output even when trained on approximately ten hours of parallel data. However, high error rates and frequent lexical substitutions indicate that current model performance remains unstable and far from deployment-ready. In particular, the discrepancy between surface-level metrics (e.g., WER, BLEU) and semantic similarity measures (e.g., BERTScore) suggests that while models capture partial meaning, they struggle with lexical fidelity and culturally specific expressions.

These findings highlight a central challenge in extremely low-resource speech translation: model architecture is not the primary bottleneck—data scale, consistency, and linguistic coverage are. Fine-tuning powerful multilingual models provides measurable gains, but without substantially larger and more diverse parallel corpora, performance will remain constrained.

The broader contribution of this work lies in establishing a practical framework for corpus creation in under-documented languages. Our structured elicitation approach, speaker management protocol, and alignment pipeline provide a template that can be replicated for other low-resource or isolate languages.

Ultimately, sustainable progress in Burushaski speech translation will depend less on incremental architectural modifications and more on systematic, community-centered data development. This work serves as a foundational step in that direction.

8. Bibliographical References

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv:2111.09296*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., Guo, J., Xiong, W., Mahadeokar, J., Kalinli, O., Fuegen, C., and Seltzer, M. L. (2023). AudioChatLlama: Towards general-purpose speech abilities for LLMs. *arXiv:2311.06753.*

Finlayson, N., Marsden, E., and Hawkes, R. (2024). Creating and evaluating corpus-informed word lists for adolescent, beginner-to-low-intermediate learners of French, German, and Spanish. *Language Teaching Research*. December 2024.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Schwartz, S., Yvon, F., and Zanon-Boito, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K.,

Salakhutdinov, R., and Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv:2106.07447*.

Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T.-A., Rivière, M., Mohamed, A., Dupoux, E., and Hsu, W.-N. (2021). Text-free prosody-aware generative spoken language modeling. *arXiv:2109.03264*.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. H., Khalil, S., Kokkinakis, S. J., Lew, R., Sharoff, S., Vadlapudi, R., and Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):123–140.

Lahiri, A. (2018). *Linguistic Fieldwork: A Student Guide*. Cambridge University Press.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., and Dupoux, E. (2021). On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Lee, A., Chen, P.-J., Wang, C., Gu, J., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., and Hsu, W.-N. (2021). Textless speech-to-speech translation on real data. *arXiv:2112.08352*.

Malviya, S., Mishra, R., and Tiwary, U. S. (2016). Structural analysis of Hindi phonetics and a method for extraction of phonetically rich sentences from a very large Hindi text corpus. In *Proceedings of the 2016 Conference of the Oriental Chapter of the International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Bali, Indonesia. IEEE. https://doi.org/10.1109/ICSDA.2016.7919009

Max Planck Institute for Evolutionary Anthropology. (n.d.). *Field Manuals and Stimulus Materials*. Available at: https://www.eva.mpg.de/lingua/resources/fieldmanuals.php