FieldMatters 2025

Field Matters. The Fourth Workshop on NLP Applications to Field Linguistics

Proceedings of the Workshop

August 1, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-282-4

Preface

Field Matters is a workshop focused on the various applications of NLP methods to field linguistics and the analysis of field data. The primary pursuit of linguistic fieldwork is to document and describe languages. The former typically involves building a corpus and other resources for the language community, the latter ideally aims to produce a reference grammar. Advances in technology have enabled vast quantities of media to be recorded. These recordings (sound and/or video) require annotation and analysis for further linguistic research or resource development. This is often done manually. This processing bottleneck can be significantly sped up with computational methods. NLP research focuses on developing methodology for different tasks that show significant performance in high-resource languages, allowing the automation of various routine tasks. The processing burdens faced by field linguists present a natural opportunity to marry NLP practices with the workflow of a field linguist. Similarly, the future development of NLP methods could gain from the linguistic diversity and unique tasks encountered during the description/documentation efforts.

With these in mind, *Field Matters* aims to provide a platform to deepen the dialogue between Computational and Field Linguists. Our workshop is hosted by the 63rd Annual Meeting of the Association for Computational Linguistics in Vienna, Austria.

Field Matters 2025 continued to provide field linguists expert reviews, a distinct feature of the review process introduced one year ago. Each paper was assigned a field linguist alongside minimally two computational linguists. Analyzing the difference in reviews of field linguists and NLP researchers, we have seen that reviewers provide different perspectives and give more diverse and fruitful feedback: while field linguists pay attention to how practical this application could be or how well it fits in the idea of the workshop, NLP specialists comment on how relevant and accurate chosen methods are.

This year, *Field Matters* shared a venue with *SigTyp*, a workshop dedicated to linguistic typology and multilingual NLP. Although the ultimate goals of *Field Matters* and *SigTyp* differ, the co-location provided a valuable opportunity for both communities to learn from one another. Careful consideration suggested we share our space while keeping the publication processes separate. This gave us twice as many keynotes and a tightly packed schedule of oral presentations. We anticipate twice as fruitful discussions in the hallways, though the dual load brings an intense workload for both organizers and participants of the one-day event, reflecting the growing audience of both workshops.

After the hard process of reviewing all submissions, the program committee chose nine papers for a poster or oral presentation at the workshop. Accepted papers illustrate the main idea of our workshop: how field linguistics may benefit from using contemporary methods of computational analysis and how the NLP community may evolve its methods with the help of under-resourced languages. More specifically, chosen papers cover the following topics:

- The creation of datasets and tools for field linguistics
- ASR and speech processing to address the transcription bottleneck
- Machine translation for very low resource languages

We are incredibly grateful to the *Field Matters* program committee, who worked on peer review to give meaningful comments for each submission and made this workshop possible. We want to thank the invited speakers, Alexis Michaud, researcher at LACITO-CNRS in Paris, France, and Eduardo Sanchez, research scientist at Meta. We would also like to acknowledge all the authors who submitted their papers to our workshop, and we hope to continue to serve as a link between NLP specialists and field linguists.

Keynote Talks

Alexis Michaud, LACITO CNRS
"Archives, Algorithms, and Alliances: Grounding NLP in the Realities of Language Documentation"

This talk offers a linguist's perspective on the evolving place of NLP in language documentation, focusing on the interplay between archives (as both legacy and infrastructure), algorithms (with ASR on the Na language as an example), and alliances (the human networks that sustain such work). Drawing on experience within "Computational Language Documentation" projects led by computer scientists, I reflect on shared goals, realistic expectations, and the practical conditions required to keep interdisciplinary teams motivated over time.

Eduardo Sanchez, Meta
"A few good texts: how small sets of high quality linguistic data power massive multilinguality in language models"

While scale remains a key driver of performance in multilingual language models, it's not always an option, especially for low-resource languages where data is scarce or noisy. We'll explore how small, high-quality datasets can play a surprisingly powerful role in enabling multilinguality, especially where coverage gaps exist. Beyond parallel corpora, we'll show how strategic use of linguistic resources can complement large-scale training, improve generalization, and unlock better performance for underserved languages. A few good texts, chosen well, may be worth billions of tokens, and for many languages, they may be the key to ensuring visibility, usability, and survival in the digital age.

Organizing Committee

General Chairs

Éric Le Ferrand, Boston College
Elena Klyachko
Anna Postnikova
Oleg Serikov
Tatiana Shavrina, Meta
Ekaterina Voloshina, University of Gothenburg, Chalmers University of Technology
Ekaterina Vylomova, University of Melbourne

Program Committee

Program Chairs

Eric Le Ferrand, Boston College
Elena Klyachko
Anna Postnikova
Oleg Serikov, King Abdullah University of Science and Technology
Tatiana Shavrina, Meta
Ekaterina Voloshina, Göteborg University and Chalmers University of Technology
Ekaterina Vylomova, The University of Melbourne

Reviewers

Angelina Aspra Aquino, Alexandre Arkhipov

James Bednall, Anton Buzanov

Michael Daniel

Harald Hammarström, William N. Havard

Elena Klyachko, Ezequiel Koile

Jordan Lachler, Eric Le Ferrand, Kate L Lindsey

Tessa Masis, Field Matters, Saliha Muradoglu

Shu Okabe

Anna Postnikova, Michael Proctor

Emmanuel Schang, Oleg Serikov, Tatiana Shavrina

Nick Thieberger

Alexey Vinyar, Ekaterina Voloshina, Ekaterina Vylomova

Table of Contents

Automatic Phone Alignment of Code-switched Urum–Russian Field Data

Emily P. Ahn
University of Washington
eahn@uw.edu

Eleanor Chodroff
University of Zurich
eleanor.chodroff@uzh.ch

Gina-Anne Levow
University of Washington
levow@uw.edu

Abstract

Code-switching, using multiple languages in a single utterance, is a common means of communication. In the language documentation process, speakers may code-switch between the target language and a language of broader communication; however, how to handle this mixed speech data is not always clearly addressed for speech research and specifically for a corpus phonetics pipeline. This paper investigates best practices for conducting phone-level forced alignment of code-switched field data using the Urum speech dataset from DoReCo. This dataset comprises 117 minutes of narrative utterances, of which 42% contain code-switched Urum–Russian speech. We demonstrate that the inclusion of Russian speech and Russian pretrained acoustic models can aid the alignment of Urum phones. Beyond using boundary alignment precision and accuracy metrics, we also discovered that the method of acoustic modeling impacted a downstream corpus phonetics investigation of code-switched Urum–Russian.

1 Introduction

Code-switching is a phenomenon where multilingual speakers communicate in more than one language, often within a single utterance.[1] Speakers of languages that are not widely spoken may also speak a *language of broader communication*, or *lingua franca*, in order to communicate with people in the same region or in contact settings. In the language documentation and analysis pipeline, recordings of the target language can be found to be mixed with a language of broader communication. Yet this other language is often overlooked or explicitly ignored if the goal of the fieldwork is to document the language of interest. On the other hand, it may be useful to include the mixture of languages in the analyzed data for methodological or scientific purposes. The extra data could add robustness to the performance of models, or the code-switched speech could better reflect actual usage of the target language.

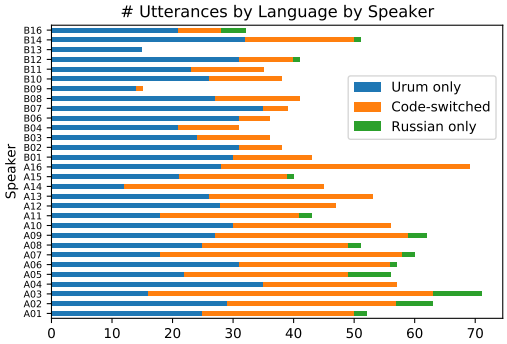

Figure 1: Across the 30 speakers in the DoReCo Urum field repository (Skopeteas et al., 2022), almost all produced code-switched utterances (orange, middle) in addition to monolingual Urum (blue, left) and monolingual Russian utterances (green, right).

Regardless of the analytical use, inclusion of the code-switched language data may benefit processes within the corpus phonetics pipeline. A critical part of this pipeline is phonetic forced alignment, in which a time-aligned phone sequence is identified from the input speech and corresponding transcript, typically using acoustic models of the language-specific phones. Generally, automatic alignment quality correlates with the amount of training data used for the acoustic model (Chodroff et al., 2024). In the case of code-switched speech, there is a question, however, of whether to use *only* the target language data—or to use *all* of the linguistic data—for training the acoustic models. Including code-switched speech during training would result in more data per speaker, which could help build more robust phone-specific acoustic models (as hy-

[1]While *code-switching* can refer to mixing languages or dialects within a whole conversation, we use it to mean switching languages within a single utterance. This finer-grained mixing is also called *code-mixing* in the literature.

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 1–14
August 1, 2025 ©2025 Association for Computational Linguistics

pothesized by Chodroff et al., 2024).

Very limited research has included code-switched speech in forced alignment studies, and our work is the first to examine this type of speech in a field data setting. We ask the following research questions (RQs):

1. Does the inclusion of Russian code-switched data in acoustic model training help the alignment of target Urum data?

2. Does the method of acoustic modeling impact a downstream corpus phonetics investigation of code-switched Urum–Russian?

In this paper, we summarize prior work and introduce the Urum language (Section 2), then discuss the methodology of data preparation, acoustic modeling and forced alignment, evaluation and analysis (Section 3). We used the Montreal Forced Aligner (McAuliffe et al., 2017) to train acoustic models from scratch as well as adapt pretrained Russian and English models to our data. With respect to RQ1, we found that the inclusion of code-switched speech and Russian pretrained models improved alignments of Urum (Section 4). To answer RQ2, we tested the impact of acoustic modeling strategies in a bilingual phonetics investigation (Section 5): Are vowels in Urum words pronounced differently in monolingual Urum utterances than in code-switched utterances? After discussion, we conclude with methodological recommendations and areas for future work (Section 6). All code for replicating this work is publicly available.[2]

2 Background

2.1 Phonetic forced alignment

For phonetics research, it can be extremely useful to know the temporal locations of phones within a speech recording. While this can be achieved manually, an automated process can greatly facilitate this, speeding up annotation and enabling analysis of substantially larger speech corpora (Labov et al., 2013). Popular forced alignment tools include the Montreal Forced Aligner (MFA, used in this work; McAuliffe et al., 2017), EasyAlign (Goldman, 2011), and WebMAUS (Kisler et al., 2017). Research has explored a range of strategies to force align low-resource data, including cross-language alignment and manipulation of phone categories (e.g., Ahn et al., 2024; Coto-Solano et al., 2018; DiCanio et al., 2013). However, forced alignment work on low-resource languages that are code-switched has been limited.

2.2 Research on the nature of code-switching

Much of the linguistic literature on code-switching has focused on the syntactic and sociopragmatic aspects of engaging multiple languages at once (Bullock and Toribio, 2009; Muysken, 2000). With respect to the phonetics of code-switching, research has focused on how acoustic properties shift when speakers activate multiple languages in their mind. For example, stop consonant voice onset time and speech rate changed noticeably near language switch boundaries between Spanish and English (Fricke et al., 2016). Relevant to our case study, Seo and Olson (2024) recorded read sentences from Korean–English bilinguals to investigate vowel quality across different syntactic structures. They found that English vowels in code-switched Korean–English utterances were more Korean-like in intra-sentential rather than in inter-sentential code-switched structures. We similarly investigated vowel quality in Urum–Russian code-switched utterances for this paper.

It has been observed that a language of broader communication, usually a high-resource language, is often used during the elicitation of a low-resource, target field language. In an overview of methods to bridge language documentation and speech processing technologies, Levow et al. (2017) proposed a language identification task between a high-resource language and a low-resource target language when both are present in field recordings. San et al. (2022) addressed the mixing of high- and low-resource languages by applying state-of-the-art language technologies to detect and transcribe English portions of speech in a dataset documented for the field language Muruwari. In this case, English was largely used in meta-linguistic commentary and questions, such as, *"What is the word for tree?"* This approach helped the annotation process, where authorized people could scan the meta-linguistic content and triage the recordings for later annotators who had more limited access to the corpora. These studies demonstrate that (1) language mixing is prevalent, and (2) applying technology to the higher-resource language can benefit the documentation process.

Developing technologies for code-switching is still a challenging area of research in the Natural Language Processing (NLP) and speech commu-

[2] https://github.com/emilyahn/align_cs

nities. Winata et al. (2023) found over 400 public research papers on code-switching from the ACL anthology and ISCA proceedings over the past few decades. These works focused on tasks ranging from language identification to sentiment analysis to automatic speech recognition (ASR). Among these papers, English mixed with a non-English variety such as Hindi, Chinese, and Spanish, was overrepresented. The authors highlighted a need for work to be done on more diverse non-English language pairs, for which this paper fills a gap.

Forced alignment with code-switched data Two studies incorporated a language of broader communication when training forced alignment systems on field data, though the impacts of mixed language speech input were not explicitly studied. Ahn et al. (2024) included Portuguese speech when developing acoustic models for Panãra, an Amazonian language of Brazil. Chodroff et al. (2024) retained the Russian speech content in the acoustic modeling of Evenki, a Tungusic language, and Urum, a Turkic language, which is also used in this work (Kazakevich and Klyachko, 2022; Skopeteas et al., 2022).

More relevant to the present study is work by Pandey et al. (2020) who compared methods of training and aligning code-switched Hindi–English read speech. Three acoustic models were trained with MFA: Hindi-only, English-only, and Hindi–English mixed, and they discovered that the combined model best aligned English-only speech. It was unclear, however, if the high performance from the Hindi–English mixed acoustic models was due to that model simply having more tokens in its training data than the other models. Our work extends these findings to a low-resource field data scenario with spontaneous speech, and we carefully controlled the variable of training data quantity. We investigated whether including code-switched data could improve the alignment performance of a target low-resource language.

2.3 Urum language

Urum (ISO: uum) is a Turkic language spoken by ethnic Greeks in the Lesser Caucasus of Georgia and in Ukraine. Also known as Caucasian Urum, it is a variety of Anatolian Turkish that is classified as endangered (Campbell et al., 2022). The language variety documented by Skopeteas et al. (2022) and analyzed in this paper has been strongly influenced by Russian since the group's arrival in Georgia in the early 19th century. Notably, most Urum speakers are bilingual in Russian and code-switch often between the two languages (Skopeteas, 2014). Unlike the examples of code-switching being used in purely meta-linguistic commentary, Russian portions of speech in this dataset were part of the narrative content by the speakers. The following shows an example of an Urum–Russian utterance with transliterated Russian displayed in brackets:

> äp halhımız egıler kissäya **[muzıka]** ederıh **[maladež]** **[tantsuet]** oinamah etmäh
>
> *"All the people get together at the church, we organise **[music]**, and the **[youth]** is **[dancing]**."* (Skopeteas et al., 2022)

3 Methodology

3.1 Data source

We utilized the Urum dataset from the DoReCo corpus, which is a field data repository that contains manual word-level and automatic phone-level alignments of speech (Paschen et al., 2020). Traditional and personal Urum narratives were recorded across 30 speakers (16 female, 14 male) and spanned 117 minutes[3] of speech (Skopeteas et al., 2022). Figure 1 presents the distribution of Urum-only, Russian-only, and code-switched utterances among speakers. All but one speaker code-switched. Table 1 reveals that while 42% of the utterances were code-switched, they represent 53% of the repository in minutes.[4] Code-switched utterances averaged 6.5 seconds, which was on average longer than non-codeswitched utterances (Urum-only: 4.5 seconds; Russian-only: 2 seconds). Among the code-switched utterances, Urum word tokens were more frequent than Russian word tokens, as shown in Figure 2.

3.2 Data preparation

Data from the field repository included long-form audio recordings (wav format, sampling rate of 44.1 kHz) and time-aligned transcriptions of the utterances, words, and phones. The audio files were first segmented into utterances using Praat (Boersma and Weenink, 2022), with the corresponding utterance transcripts as Praat TextGrids. Four utterances were removed due to encoding issues.

[3]Time was calculated by summing utterance durations, not file or word durations.

[4]If all utterances with "foreign material" were excluded, as was the protocol in Zhu et al. (2024) over the full DoReCo corpus, we would miss out on half the data.

Utt	Count	Time (min)	Avg (sec)
All	1373	117.6	5.1
Urum	752 (55%)	53.5 (45%)	4.3
CS	581 (42%)	62.9 (53%)	6.5
Russ	40 (3%)	1.3 (1%)	2.0

Table 1: Distribution of utterances across language usage by count and time. Notably, code-switched (CS) utterances had longer durations.

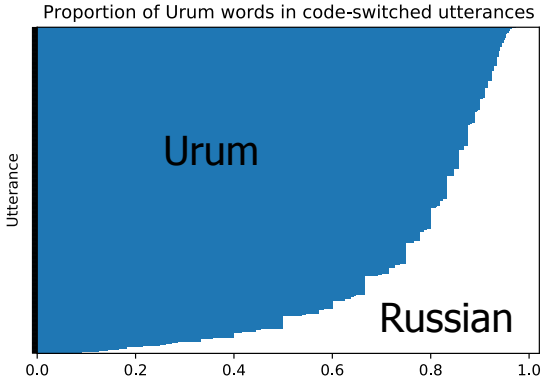

Figure 2: Proportion of Urum (blue, shaded) to Russian (white) word tokens in all 581 code-switched utterances, sorted highest to lowest. The majority of these utterances had more Urum than Russian tokens.

Urum phone sequences were derived automatically by the repository contributors, so our lexicons (two-column text files with words and their corresponding phone sequences) were gathered from these existing phone sequences. Most of the Russian words had been transliterated into Latin script at the word-level, so we used a simple mapping script to build the lexicon. The Urum phone set from DoReCo included nine vowels and 30 consonants while the transliterated Russian phone set included six vowels and 19 consonants. Only four Russian phones did not exist in the Urum set, as seen in Table 2, and we used the PanPhon tool (Mortensen et al., 2016) to map them to their nearest neighboring Urum phones in the lexicons: ɨ → ɯ, tɕ → tʃ, ʂ → ʃ, ʐ → ʒ. Partially-tagged words such as filled pauses and prolongations were assigned phone sequences in the lexicon and were marked as Urum words.[5]

[5]The repository contributors used tags to transcribe content such as filled pauses, prolongations, and false starts. When a tagged word was partially transcribed (such as in this example of a false start, "<fs>ba"), we manually assigned it a phone sequence ("[b a]") and classified it as an Urum word.

Urum-only	Both	Russian-only
y, æ, œ, ɯ	a, e, i, o, u	ɨ
ɟ, c, dː, tː	b, p, d, t, g, k	
sː, ʃ, ʒ, ɣ, dʒ, tʃ	v, f, z, s, x	tɕ, ʂ, ʐ
l, lː, ɾ, mː	j, r, ɬ, m, n	

Table 2: The phone sets present in the DoReCo transcriptions across Urum and Russian with the middle column representing their overlap.

3.3 Acoustic modeling

We used the Montreal Forced Aligner (MFA version 2.2.17; McAuliffe et al., 2017) to train acoustic models and conduct forced alignment on our data in its default unsupervised manner. Acoustic models learn the probability distributions for all given phone states and their transitions. We split the DoReCo files into mutually exclusive train and test partitions following the same split as Chodroff et al. (2024): 1,097 utterances (100 minutes) in the train set and 273 utterances (16 minutes) in the test set. For this study, we created further subsets of the training data to answer our first research question. First, we summed the minutes of utterances of each language type and found 47 minutes of monolingual Urum utterances and 52 minutes of code-switched utterances. To keep the quantity of Urum-only and code-switched training data the same, we reduced the number of code-switched utterances to 47 minutes, which would equal the Urum-only speech. Our first results compared the alignment performance of a model trained on 47 minutes of purely Urum speech to a model trained on 47 minutes of Urum–Russian speech.[6] Our third training data partition combined both sets to include 94 minutes of Urum-only and code-switched speech. All Russian-only utterances were excluded from training and evaluation.

Since it has been shown to be advantageous to use larger, pretrained models for aligning low-resource languages (e.g., Ahn et al., 2024), we chose two relevant MFA models to continue the experiments. The Russian MFA v3.1.0 model was trained on over 400 hours of data from over 3,000 speakers; this model was selected since Russian was frequently spoken in our dataset (McAuliffe and Sonderegger, 2024). The Global English MFA

[6]While the minutes across the two partitions were the same, the number of utterances was 618 for Urum and 414 for code-switched. However, the number of phones in each partition was roughly 27,000.

v2.2.1 model was trained on over 3,700 hours of data from over 79,000 speakers across the world (McAuliffe and Sonderegger, 2023). This model has previously proven to be effective in aligning the Urum dataset in Chodroff et al. (2024). For cross-language modeling and alignment, we developed the lexicons by applying the PanPhon tool for determining the nearest neighboring phones in cases where the target phone did not exist in the model (Mortensen et al., 2016). Appendix A displays these phone mappings. Each pretrained model was adapted to the same three training data partitions as described in the train-from-scratch settings above.

3.4 Forced alignment evaluation

Our "gold standard" phone alignments for evaluation of the system outputs were obtained from the manually annotated phone boundaries of Urum words in the test partition from Chodroff et al. (2024). For *precision*, we calculated the percent for which the model onset boundary was within 20 ms (the selected threshold) of the manually aligned onset boundary (McAuliffe et al., 2017; MacKenzie and Turton, 2020). For *accuracy*, we utilized a measure that calculated the proportion of model-aligned intervals whose midpoints lay within the respective gold intervals (a similar measure is used in Knowles et al., 2018; Mahr et al., 2021). All evaluation was conducted on the test partition which consisted of 132 Urum utterances and 119 code-switched utterances. The evaluation was conducted only on phones from Urum words and ignored all phones from Russian words.

3.5 Analysis

We conducted mixed-effects regressions in R using the lme4 package to analyze the variables that contributed to both the precision and the accuracy metrics (Bates et al., 2015). We ran two models: the first was a linear model with the dependent variable as log seconds of onset boundary differences, with 0 seconds mapped to 0.001 prior to the log transformation. The second model was a logistic regression with the binary dependent variable being accuracy. Main effects were the language of the test utterance (Urum or CS), the natural class of the current phone, the natural class of the previous phone, the interaction of these two natural classes, the proportion of contaminated (tagged) tokens,[7] the utterance duration (in hectoseconds, seconds / 100, for model convergence), the interactions of model configuration with utterance language, and whether or not the speaker of the test utterance was present in the training set. Random effects were the speaker ID and the file ID of the utterances. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. The eight classes analyzed were vowels, approximants, taps/trills, nasals, fricatives, affricates, and stops. Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model.

[7] Contamination in an utterance was calculated as the number of tagged tokens, such as false starts or prolongations, divided by the total number of tokens.

4 Results

4.1 Alignment precision and accuracy

The following results answer our first research question: Does including Russian code-switched data in acoustic model training help the alignment of target Urum data? The different acoustic model configurations were trained or adapted on subsets of the DoReCo dataset, and they were all tested on the held-out test utterances that included both Urum-only and CS utterances. In the scenario where we trained MFA models from scratch (i.e., no pretrained model was used—note the None column in Table 3), we have two findings. When we kept the training data quantity equal at 47 minutes for both Urum-only speech and code-switched speech, the Urum-only model (47m) outperformed the purely code-switched one (47m). This was expected given that we evaluated only on phones from Urum words. However, combining these two training sets in the Urum + CS (94m) model substantially improved upon either smaller model. This also conforms to expectations given that the combined training set included more Urum tokens and also more data overall.

For the experiments using pretrained models adapted on the various Urum/CS partitions, the Russian MFA model adapted on Urum + CS (94m) produced the best results. Even though the Global English MFA model was trained on nearly 4,000 hours of diverse speech, its alignments did not outperform the smaller Russian MFA model. This is perhaps due to the language similarity of Russian to Urum, or the history of Urum being influenced by Russian contact. All models trained or adapted on the different DoReCo subsets patterned the same where the ranking of best to worst subset was Urum + CS (94m) > Urum (47m) > CS (47m), with the

Train/Adapt Partition	Pretrained model None	Eng	Russ
Precision % ↑			
Urum (47m)	63.2	70.4	71.2
CS (47m)	58.2	70.0	70.4
Urum + CS (94m)	70.9	70.6	**71.3**
Accuracy % ↑			
Urum (47m)	80.6	83.7	84.9
CS (47m)	77.2	83.1	84.4
Urum + CS (94m)	**85.1**	83.6	**85.1**

Table 3: Results revealed that the Russian MFA model adapted on all 94 minutes of Urum and code-switched (CS) data performed the best, with maximal training-from-scratch (i.e., Urum + CS (94m)) on par in terms of accuracy. Highest scores are bolded and shaded.

slight exception of accuracy from the Global English MFA with Urum (47m) > Urum + CS (94m).

4.2 Regression analysis

The mixed-effects regression analysis revealed several factors that influenced alignment performance. We report all significant findings of $p < 0.05$, and full output tables are provided in Appendix B. Except for the train-from-scratch CS (47m) model which performed significantly worse, all other models performed significantly better than the Urum (47m) model. Longer utterance durations and higher contamination amounts were correlated with worse performance. The speaker appearing in the training data had no significant effect. The language of the test utterance also had no effect, with a slight exception of the CS (47m) model performing slightly worse on Urum-only test utterances.

In terms of precision, boundaries around taps/trills were displaced more significantly, while boundaries around fricatives showed higher precision. Boundaries preceding vowels also performed better. Significantly better precision was found for vowel–tap/trill, fricative–vowel, affricate–vowel, affricate–nasal, stop–vowel, and stop–tap/trill sequences. Significantly worse precision was found for vowel–vowel, vowel–approximant, tap/trill–vowel, nasal–nasal, and fricative–approximant sequences.

As for accuracy, which used a logistic mixed-effects regression model, significantly better performance was found for phone intervals preceded by nasals, fricatives, affricates, and stops, as well as for targeted phone intervals that were fricatives and affricates. Significantly worse accuracy was found for phone intervals preceded by vowels, approximants, and taps/trills, as well as targeted phone intervals of these three classes. These results are largely comparable to the mixed-effects regression results from Chodroff et al. (2024).

5 Case Study

Following Babinski et al. (2019), we asked a general phonetics question and observed whether there were significant differences between the outputs of the different model configurations above. In other words, to what degree are we comfortable substituting an automatic alignment for manual alignment, in our quest to answer a question about code-switching phonetics? We investigated the following: Are vowels in Urum words pronounced differently in monolingual Urum utterances compared to in CS utterances? First, we answered this with the manually-annotated "gold" test data.

5.1 Methodology

The Pillai–Bartlett trace, or Pillai score, is a useful metric to measure overlap in vowel category qualities. It takes output from a MANOVA test, which is used for measuring overlap between two distributions across two dependent variables—in our case, the first two formant values. Among four commonly used metrics for vowel overlap, Kelley and Tucker (2020) showed that Pillai scores are among the most reliable. Stanley and Sneller (2023) additionally provided a formula to derive a threshold for determining overlap vs separation based on the exact sample size of the tokens. We followed these recommendations and calculated Pillai scores for formant values extracted from the gold test set. Formants were first extracted with the Linear Predictive Coding (LPC) tool in Praat (Boersma and Weenink, 2022), searching for five formants under 5000 Hz for reported male speakers and 5500 Hz for reported female speakers. The formant value analyzed per vowel was an average of the values extracted from the interval midpoint and ten milliseconds before and after the midpoint.

5.2 Results from manual alignments

The gold test data revealed several instances of within-speaker differences in pronouncing certain Urum vowels. Table 4 shows four instances of a particular vowel being marked as significantly non-overlapping across two conditions. For example, the cell for male speaker A03 /a/ marked with

	Spkr	VOWELS a	e	i	o	u	y	œ	æ	ɯ
Male	A01									
	A03	n=189			n=57					
Female	A02									
	A07	n=13								
	B08									
	B11									
	B16	n=20								

Table 4: Our case study revealed that from the gold data, /a/ for 3 speakers and /o/ for 1 speaker (marked with shaded cells and token counts) in Urum words were pronounced significantly differently in monolingual Urum vs code-switched utterances. For these four instances, Pillai scores indicated that the vowel formants for the two groups in question (Urum vs CS) were significantly non-overlapping.

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03	X		X	X					
A02									
A07	X								
B08								X	
B11									
B16									

Table 5: The *best-performing* acoustic model (Russian MFA adapted on Urum + CS 94m) yielded 3 true positives (shaded X), 3 false positives (unshaded X), and 1 false negative (shaded empty cell).

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03			X	X				X	X
A02									
A07	X								
B08									
B11									
B16									

Table 6: The *worst-performing* acoustic model (trained on the CS 47m partition) yielded 2 true positives (shaded X), 4 false positives (unshaded X), and 2 false negatives (shaded empty cells).

$n = 189$ indicates that A03 uttered 189 /a/ vowels, and his F1×F2 values for /a/ in Urum words from monolingual Urum utterances were significantly different than values for /a/ in Urum words from code-switched utterances. The same can be said for speaker A03's /o/ ($n = 57$), speaker A07's /a/ ($n = 13$), and speaker B16's /a/ ($n = 20$).

5.3 Results from automatic alignments

Second, we calculated Pillai scores from the output of the best-performing and worst-performing models and compared these to the gold scores (Tables 5 and 6). From the best-performing model, the Russian MFA model adapted on the Urum + CS (94m) data, it found six instances of significant non-overlap. Three out of the four gold instances were correctly identified (i.e., three true positives and one false negative), while producing three spurious significant findings (i.e., three false positives). From the worst-performing model, trained on the CS (47m) partition, it produced less congruent findings: only two out of the four gold instances were correctly identified (i.e., two true positives and two false negatives), with four spurious significant findings (i.e., four false positives). We used the phonR package in R (McCloy, 2012) to plot vowel ellipses for /i, a, o/ for male speaker A03, over the two language conditions, and across the three types of output (Figure 3). The gold plot reflects our findings that /a/ and /o/ were significantly different between Urum and CS environments while /i/ was not. The ellipses from the best and worst models show divergence from the gold ellipses. Both models found spurious differences for /i/, and although /a/ visually appears significantly different for the worst model, /a/ was a false negative.

Essentially, the automatic alignments did not yield the same findings as those from the gold alignments in our vowel overlap analysis. Output from the best- and worst-performing models tended to

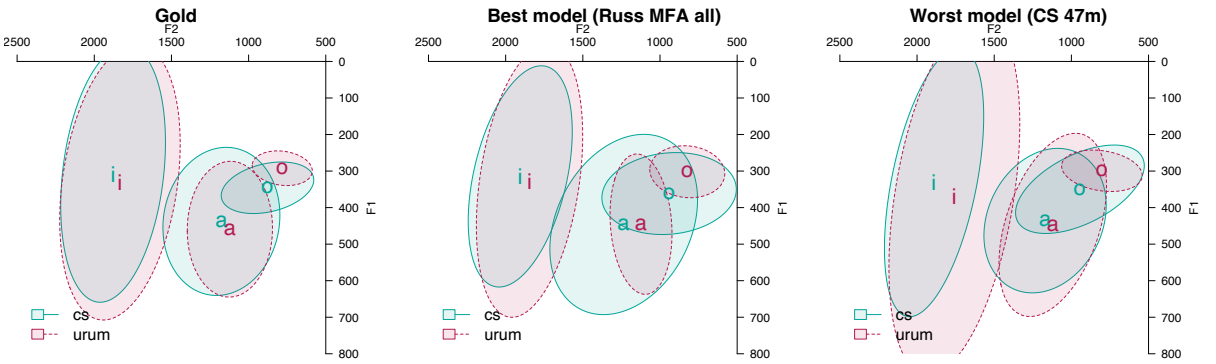

Figure 3: These plots reflect the first two formants (in Hz) for three of the nine Urum vowels, /a, i, o/, for male speaker A03. From left-to-right are formants extracted from the gold alignments, the best model (Russian MFA adapted on Urum + CS 94m) output, and the worst model (CS 47m) output. Vowel labels are positioned at means, and ellipses cover one standard deviation away from the mean.

hallucinate more vowel disparities than the gold output suggested, though the best model's vowel disparity predictions more closely aligned to the gold findings than the worst model's. While the best model's alignments were 11 percentage points higher than the worst model's alignments for precision (and seven percentage points higher for accuracy), these differences can be hard to interpret. This case study allowed us to reveal the nuances of alignment performance, as the downstream output yielded different conclusions.

6 Conclusion

This work tested methodologies of incorporating code-switched data in acoustic model training and alignment in a low-resource, field data scenario. We tested the inclusion of Urum–Russian code-switched utterances in training acoustic models to align Urum phones and found that it was helpful to keep the code-switching to produce a larger train set.[8] The maximally trained-from-scratch model performed roughly on-par with a pretrained Russian model adapted to the same field data. If one is fortunate enough to have 90-some minutes of transcribed data, it should be sufficient to train models (see also the recommendations in Chodroff et al., 2024). Otherwise, utilizing a large, pretrained model performed reasonably, particularly when adapted on target data.

In order to functionally assess the quality of the systems, we tested our best and worst systems' alignment outputs against the gold alignments to answer a bilingual phonetics question (RQ2). Calculating Pillai scores across formant values for individual speakers, we discovered that several speakers pronounced certain Urum vowels significantly differently in monolingual Urum utterances than in code-switched utterances. While not matching the gold alignment results exactly, the best acoustic model yielded more similar results to the gold than the worst acoustic model. We recommend manual adjustment of phone boundaries when conducting phonetic analyses, particularly those involving smaller datasets and temporally sensitive phonetic measurements (e.g., analysis of duration or cases where the boundary determines the measurement location such as onset f0).

As future work, it would be beneficial to conduct a survey study with qualitative and quantitative statistics on the prevalence of code-switching across field data repositories. How are multiple languages used by the elicitors and by the community members of the language being documented?

Further research could also aim to extend the study of phonetics and phonology for code-switched language more broadly. Our case study only scratched the surface to discover the nature of shifting Urum vowel qualities depending on the languages present in an utterance. It would be interesting to discover if the significantly different Urum vowel formants were becoming more Russian-like when surrounded by Russian context, similar to findings on Korean–English by Seo and Olson (2024). Cross-linguistic interference or transfer could be in effect and is worth investigating.

[8] Our findings echo similar cross-language modeling experiments from other domains such as speech recognition and text-based NLP research, where the inclusion of data from a higher-resource language improved model performance on low-resource language data (e.g., Downey et al., 2024; Farooq and Hain, 2023; Fujinuma et al., 2022).

Limitations

When conducting our regression analyses or case study, we did not take into account code-switching properties at the syntactic or prosodic level. It would be interesting to factor into account whether the code-switched utterance was inter-sentential or intra-sentential (i.e., mixing languages at phrase boundaries or within phrases). When calculating boundary differences, examining how close an Urum word was to a Russian word could have provided useful information. Prosodic factors such as speech rate and pitch would also add insight as, anecdotally, prosody was at times visibly different near the language switch points. Additionally, code-switched words can be confused with loanwords that have a legitimate place in a language's lexicon. All of the Russian words in this repository were explicitly tagged as Russian by the field linguists, but there may be disagreement to the classification of language at the token-level.

The Urum dataset from the DoReCo repository used in this work was particularly well-annotated for both Urum and Russian. However, the quality and quantity of transcriptions here may not be comparable to that in other field data repositories, and replication of our findings on other datasets may be challenging.

Ethics Statement

The dataset in this study has been made publicly available for download and research use. Speech data that is public carries inherent potential harms for misuse in downstream tasks.

Particularly for our methodological approach of including code-switched speech or the language of broader communication in the analysis of field data, we advise some caution. The speech from the non-target language may have been meant to be ignored and not recorded. If sections of the speech data were not explicitly transcribed, they may not have been intended to be used for analysis.

References

Emily P. Ahn, Eleanor Chodroff, Myriam Lapierre, and Gina-Anne Levow. 2024. The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra. In *Interspeech 2024*, pages 1505–1509.

Sarah Babinski, Rikker Dockum, J. Hunter Craft, Anelisa Fergus, Dolly Goldenberg, and Claire Bowern. 2019. A Robin Hood Approach to Forced Alignment: English-trained Algorithms and their Use on Australian Languages. *Proceedings of the Linguistic Society of America*, 4:3:1–12.

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

Paul Boersma and David Weenink. 2022. Praat: Doing Phonetics by Computer (Version 6.0.3).

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge Handbook of Linguistic Code-switching*. Cambridge University Press.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, and Kaori Ueki. 2022. The Catalogue of Endangered Languages (ElCat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Eleanor Chodroff, Emily P. Ahn, and Hossep Dolatian. 2024. Comparing Language-specific and Cross-language Acoustic Models for Low-resource Phonetic Forced Alignment. *Language Documentation & Conservation*.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33, Dunedin, New Zealand.

Christian DiCanio, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, and Rey Castillo García. 2013. Using Automatic Alignment to Analyze Endangered Language Data: Testing the Viability of Untrained Alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted Multilingual Adaptation for Low-resource Language Families. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663, Miami, Florida, USA. Association for Computational Linguistics.

Muhammad Umar Farooq and Thomas Hain. 2023. Learning Cross-lingual Mappings for Data Augmentation to Improve Low-resource Speech Recognition. In *Interspeech 2023*, pages 5072–5076.

Melinda Fricke, Judith F. Kroll, and Paola E. Dussias. 2016. Phonetic Variation in Bilingual Speech: A Lens for Studying the Production–comprehension Link. *Journal of Memory and Language*, 89:110–137. Speaking and Listening: Relationships Between Language Production and Comprehension.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Jean-Philippe Goldman. 2011. EasyAlign: an Automatic Phonetic Alignment Tool under Praat. In *Interspeech 2011*, pages 3233–3236.

Olga Kazakevich and Elena Klyachko. 2022. Evenki DoReCo Dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Matthew C. Kelley and Benjamin V. Tucker. 2020. A Comparison of Four Vowel Overlap Measures. *The Journal of the Acoustical Society of America*, 147(1):137–145.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual Processing of Speech via Web Services. *Computer Speech & Language*, 45:326–347.

Thea Knowles, Meghan Clayards, and Morgan Sonderegger. 2018. Examining Factors Influencing the Viability of Automatic Acoustic Analysis of Child Speech. *Journal of Speech, Language, and Hearing Research*, 61(10):2487–2501.

William Labov, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language*, 89(1):30–65.

Gina-Anne Levow, Emily M. Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, Michael Maxwell, Michael Tjalve, and Fei Xia. 2017. STREAMLInED Challenges: Aligning Research Interests with Shared Tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47.

Laurel MacKenzie and Danielle Turton. 2020. Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English. *Linguistics Vanguard*, 6(s1):20180061.

Tristan J. Mahr, Visar Berisha, Kan Kawabata, Julie Liss, and Katherine C. Hustad. 2021. Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research*, 64(6S):2213–2222.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2023. English MFA acoustic model v2.2.1. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20MFA%20acoustic%20model%20v2_2_1.html.

Michael McAuliffe and Morgan Sonderegger. 2024. Russian MFA acoustic model v3.1.0. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/Russian/Russian%20MFA%20acoustic%20model%20v3_1_0.html.

Daniel R McCloy. 2012. Vowel Normalization and Plotting with the phonR Package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, 1:1–8.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Ayushi Pandey, Pamir Gogoi, and Kevin Tang. 2020. Understanding Forced Alignment Errors in Hindi-English Code-Mixed Speech–a Feature Analysis. In *Proceedings of the First Workshop on Speech Technologies for Code-Switching in Multilingual Communities*, pages 13–17.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a Time-aligned Cross-linguistic Reference Corpus from Language Documentation Data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France. European Language Resources Association.

Nay San, Martijn Bartelds, Tolúlopé Ògúnrèmí, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. 2022. Automated Speech Tools for Helping Communities Process Restricted-access Corpora for Language Revival Efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41–51.

Yuhyeon Seo and Daniel J. Olson. 2024. Phonetic Shifts in Bilingual Vowels: Evidence from Intersentential and Intrasentential Code-switching. *International Journal of Bilingualism*, 0(0):1–16.

Stavros Skopeteas. 2014. Caucasian Urums and Urum Language. *Journal of Endangered Turkish Languages*, 3(1):333–364.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2022. Urum DoReCo Dataset. In Frank Seifart, Ludger Paschen,

and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Joseph A. Stanley and Betsy Sneller. 2023. Sample Size Matters in Calculating Pillai Scores. *The Journal of the Acoustical Society of America*, 153(1):54–67.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The Taste of IPA: Towards Open-Vocabulary Keyword Spotting and Forced Alignment in Any Language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750–772. Association for Computational Linguistics.

A Cross-language Phone Mappings

Table 7 shows the mappings from Urum or Russian to either English or Russian pretrained MFA models.

B Regression Results

Tables 8 and 9 display the output from the mixed-effects regression models.

Russ (CS) to Eng MFA	Urum to Eng MFA	Urum to Russ MFA
tɕ → tʃ	dː → d	ɾ → r
ɨ → ɯ	lː → l	œ→ ɛ
ʂ → ʃ	mː → m	ɯ → ɨ
ʐ → ʒ	r → ɾ	ʃ → ʂ
	sː → s	ʒ → ʐ
	tː → t	d → d̪
	x → ç	dː → d̪ː
	ʏ → ʉ̟	dʒ → dʐ̪
	œ→ ɛ	l → ɬ
	ɣ → ç	lː → ɬː
	ɯ → ə	n → n̪
		s → s̪
		sː → s̪ː
		t → t̪
		tː → t̪ː
		tʃ → tʂ̪
		ʏ → ʉ̟
		z → z̪

Table 7: Urum and Russian (code-switched) phones from DoReCo that did not exist in the pretrained English or Russian MFA model lexicons were mapped to their nearest neighboring phones, calculated with the PanPhon tool (Mortensen et al., 2016).

Predictors	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-4.39	0.18	-24.37	**<0.001**
CS (47m)	0.11	0.03	3.35	**<0.001**
Urum + CS (94m)	-0.28	0.03	-8.56	**<0.001**
English + Urum (47m)	-0.22	0.03	-6.83	**<0.001**
English + CS (47m)	-0.20	0.03	-6.02	**<0.001**
English + Urum/CS (94m)	-0.22	0.03	-6.67	**<0.001**
Russian + Urum (47m)	-0.20	0.03	-6.17	**<0.001**
Russian + CS (47m)	-0.18	0.03	-5.42	**<0.001**
Russian + Urum/CS (94m)	-0.21	0.03	-6.46	**<0.001**
Utterance duration	4.03	0.25	16.39	**<0.001**
Contamimation amount	0.63	0.05	13.14	**<0.001**
Speaker seen in training	0.20	0.20	0.98	0.360
Utt is Urum-only	0.01	0.03	0.18	0.859
Prec vowel	-0.11	0.08	-1.39	0.166
Prec approx	0.31	0.55	0.57	0.567
Prec tap/trill	0.30	0.04	7.00	**<0.001**
Prec nasal	-0.17	0.11	-1.56	0.118
Prec fric	-0.22	0.10	-2.28	**<0.05**
Prec affr	0.63	0.39	1.63	0.103
Prec stop	-0.12	0.10	-1.19	0.236
Vowel	-0.39	0.08	-4.79	**<0.001**
Approximant	-0.03	0.09	-0.37	0.712
Tap/trill	0.75	0.38	1.97	**<0.05**
Nasal	-0.18	0.09	-1.95	0.052
Fricative	-0.20	0.08	-2.56	**<0.05**
Affricate	-0.01	0.12	-0.10	0.921
CS (47m) x Utt is Urum-only	0.09	0.05	2.06	**<0.05**
Urum + CS (94m) x Utt is Urum-only	0.04	0.05	0.86	0.390
English + Urum (47m) x Utt is Urum-only	0.01	0.05	0.22	0.828
English + CS (47m) x Utt is Urum-only	0.00	0.05	-0.03	0.974
English + Urum/CS (94m) x Utt is Urum-only	-0.01	0.05	-0.15	0.880
Russian + Urum (47m) x Utt is Urum-only	-0.02	0.05	-0.45	0.651
Russian + CS (47m) x Utt is Urum-only	-0.02	0.05	-0.42	0.674
Russian + Urum/CS (94m) x Utt is Urum-only	-0.01	0.05	-0.23	0.822
Prec vowel x vowel	0.79	0.08	9.31	**<0.001**
Prec vowel x approx	0.39	0.09	4.35	**<0.001**
Prec vowel x tap/trill	-0.83	0.38	-2.17	**<0.05**
Prec vowel x nasal	-0.18	0.09	-1.89	0.059
Prec vowel x fric	-0.10	0.08	-1.24	0.215
Prec vowel x affr	0.24	0.13	1.84	0.066
Prec approx x vowel	0.14	0.55	0.25	0.802
Prec approx x approx	-0.86	0.56	-1.54	0.125
Prec approx x tap/trill	0.68	3.26	0.21	0.836
Prec approx x nasal	0.45	0.56	0.81	0.419
Prec approx x fric	-0.16	0.64	-0.25	0.800
Prec approx x affr	-0.14	0.56	-0.25	0.801
Prec tap/trill x vowel	0.15	0.05	3.06	**<0.01**
Prec tap/trill x approx	0.05	0.07	0.66	0.510
Prec tap/trill x nasal	-0.17	0.11	-1.61	0.108
Prec tap/trill x fric	0.06	0.14	0.40	0.693
Prec tap/trill x affr	-0.13	0.13	-1.05	0.296
Prec nasal x vowel	0.21	0.11	1.94	0.052
Prec nasal x approx	0.09	0.16	0.57	0.572
Prec nasal x tap/trill	-0.24	0.54	-0.45	0.651
Prec nasal x nasal	0.52	0.13	4.03	**<0.001**
Prec nasal x fric	-0.24	0.13	-1.94	0.053
Prec nasal x affr	-0.14	0.19	-0.76	0.445
Prec fric x vowel	-0.25	0.10	-2.58	**<0.01**
Prec fric x approx	0.25	0.11	2.21	**<0.05**
Prec fric x tap/trill	-0.54	0.47	-1.16	0.245
Prec fric x nasal	-0.05	0.15	-0.34	0.736
Prec fric x fric	0.16	0.12	1.36	0.173
Prec fric x affr	0.20	0.20	0.99	0.321
Prec affr x vowel	-1.05	0.39	-2.68	**<0.01**
Prec affr x approx	-0.49	0.43	-1.13	0.257
Prec affr x tap/trill	3.15	2.14	1.48	0.140
Prec affr x nasal	-1.04	0.48	-2.20	**<0.05**
Prec affr x fric	-0.01	0.94	-0.01	0.993
Prec stop x vowel	-0.32	0.10	-3.11	**<0.01**
Prec stop x approx	-0.01	0.12	-0.08	0.933
Prec stop x tap/trill	-0.80	0.38	-2.12	**<0.05**
Prec stop x nasal	0.22	0.14	1.60	0.109

Table 8: Linear mixed-effects regression results for phone onset boundary difference (in log seconds, with 0 seconds mapped to 0.001 prior to the log transformation). Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. Utterance duration was entered as hectoseconds (seconds / 100).

Predictors	Estimate	Std. Error	z-value	Pr(>|z|)
Intercept	2.07	0.27	7.68	**<0.001**
CS (47m)	-0.22	0.04	-5.05	**<0.001**
Urum + CS (94m)	0.34	0.05	7.10	**<0.001**
English + Urum (47m)	0.23	0.05	4.94	**<0.001**
English + CS (47m)	0.18	0.05	3.93	**<0.001**
English + Urum/CS (94m)	0.22	0.05	4.67	**<0.001**
Russian + Urum (47m)	0.33	0.05	6.87	**<0.001**
Russian + CS (47m)	0.29	0.05	6.08	**<0.001**
Russian + Urum/CS (94m)	0.35	0.05	7.17	**<0.001**
Utterance duration	-2.13	0.47	-4.56	**<0.001**
Contamimation amount	-0.94	0.10	-9.46	**<0.001**
Speaker seen in training	-0.08	0.34	-0.24	0.814
Utt is Urum-only	0.06	0.03	1.94	0.053
Prec vowel	-0.31	0.03	-10.18	**<0.001**
Prec approx	-0.16	0.04	-3.94	**<0.001**
Prec tap/trill	-0.26	0.04	-6.28	**<0.001**
Prec nasal	0.10	0.04	2.26	**<0.05**
Prec fric	0.24	0.04	6.12	**<0.001**
Prec affr	0.33	0.10	3.34	**<0.001**
Prec stop	0.15	0.03	4.64	**<0.001**
Vowel	-0.23	0.04	-5.90	**<0.001**
Approximant	-0.96	0.04	-23.66	**<0.001**
Tap/trill	-1.11	0.04	-27.73	**<0.001**
Nasal	0.03	0.04	0.60	0.547
Fricative	0.45	0.05	9.97	**<0.001**
Affricate	1.62	0.17	9.72	**<0.001**

Table 9: Logistic mixed-effects regression results for accuracy. Accuracy is 1 if the midpoint of the system interval lies within the corresponding gold interval. Models were treatment-coded, each compared to the train-from-scratch Urum-only (47m) model. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. Utterance duration was entered as hectoseconds (seconds / 100).

What Causes Knowledge Loss in Multilingual Language Models?

Maria Khelli[1], **Samuel Cahyawijaya**[2], **Ayu Purwarianti**[1], **Genta Indra Winata**[3]
[1]Institut Teknologi Bandung [2]Cohere [3]Capital One
khelli07.id@gmail.com, samuelcahyawijaya@cohere.com
ayu@informatika.org, genta.winata@capitalone.com

Abstract

Cross-lingual transfer in natural language processing (NLP) models enhances multilingual performance by leveraging shared linguistic knowledge. However, traditional methods that process all data simultaneously often fail to mimic real-world scenarios, leading to challenges like catastrophic forgetting, where fine-tuning on new tasks degrades performance on previously learned ones. Our study explores this issue in multilingual contexts, focusing on linguistic differences affecting representational learning rather than just model parameters. We experiment with 52 languages using LoRA adapters of varying ranks to evaluate non-shared, partially shared, and fully shared parameters. Our aim is to see if parameter sharing through adapters can mitigate forgetting while preserving prior knowledge. We find that languages using non-Latin scripts are more susceptible to catastrophic forgetting, whereas those written in Latin script facilitate more effective cross-lingual transfer.

1 Introduction

Cross-lingual transfer in natural language processing (NLP) models has demonstrated enhanced performance in multilingual contexts compared to monolingual settings, largely due to the advantages of leveraging cross-lingual knowledge (Hu et al., 2020; FitzGerald et al., 2023; Winata et al., 2023b, 2024). Typically, training occurs only once simultaneously, where all available data is processed in a single training run. However, in real-world applications, data is often received sequentially over time, highlighting the importance of continuous model updates to maintain performance (Rolnick et al., 2019). Unlike humans, who can retain prior knowledge while acquiring new skills, neural network models often struggle to preserve previously learned information when fine-tuned on new tasks, which is known as catastrophic forgetting, a decline in performance on earlier tasks after the model is exposed to new data (Winata et al., 2023a). To mitigate this issue, several studies have investigated continual learning strategies and the implementation of adapters (Badola et al., 2023) as viable solutions. This limitation poses a significant challenge for multilingual NLP, as models must adapt to new languages while retaining previously acquired linguistic knowledge. Without an effective learning strategy, models risk performance degradation, rendering them less suitable for long-term deployment.

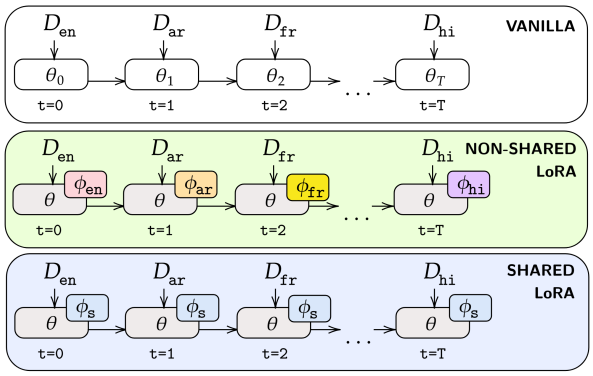

Figure 1: Pipeline for various approaches in lifelong learning. In our lifelong learning framework, we employ a LoRA-based approach where the parameters of the base model, denoted as θ, remain fixed, and for VANILLA, the model parameters are updated at all times. We explore the phenomenon of multilingual knowledge loss by comparing the effects of training with both shared and non-shared parameters.

Lifelong learning is essential for integrating new annotated data across languages without requiring full retraining of systems. As language changes and new data becomes available, models must adapt incrementally to minimize computational costs. This approach helps maintain efficiency and scalability, while addressing the challenge of catastrophic forgetting, which has been explored in various studies (Liu et al., 2021; Winata et al., 2023a; Badola et al., 2023; M'hamdi et al., 2023). However, there is a lack of systematic analysis on this issue in multi-

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 15–25
August 1, 2025 ©2025 Association for Computational Linguistics

lingual contexts. This study aims to fill that gap by investigating factors contributing to catastrophic forgetting beyond model parameters, including how linguistic differences can affect representational learning and lead to knowledge erosion when learning multiple languages sequentially.

In this study, we investigate the effects of non-shared, partially shared, and fully shared parameters in a multilingual context, examining 52 languages through the use of LoRA adapters with varying ranks and different sharing model parameter settings as shown in Figure 1. Our primary focus is to assess the impact of parameter sharing on model performance, while also conducting a comprehensive analysis of the role that different languages play in catastrophic forgetting. Additionally, we explore sequential learning to identify when performance drops occur and whether these declines are influenced by the introduction of newly learned languages or the cumulative number of previously learned languages. Our contributions can be summarized as follows:

- We examine the factors contributing to knowledge loss in multilingual language models, focusing on aspects such as language diversity, parameter sharing strategies, and base model selection within a lifelong learning framework for massively multilingual learning.
- We assess cross-lingual transferability and introduce multi-hop metrics to better understand the impact of language skills on model performance.
- We analyze model parameter adaptation to investigate trends in the model's ability to learn languages in a lifelong learning context.

2 Methodology

2.1 Task Setup

A sequence of T tasks is structured as an ordered set of datasets $\mathcal{D} = \{D_1, D_2, \ldots, D_t, \ldots, D_T\}$, where each dataset D_t corresponds to a specific task t, representing a distinct language. The model, parameterized by θ_t, undergoes iterative updates, with parameters at step $t+1$ being derived from those at step t through the function $f(\theta_t, D_t)$. These updates are performed using gradient-based optimization to maximize the log-likelihood over dataset D_t. In this paper, task T is interchangeable with language L.

2.2 Training Methods

We use XLM-R_{BASE} (Conneau et al., 2020) as our base model and compare key methods with E5 instruct (Wang et al., 2024) for evaluating the consistency of the findings. A classification layer is added on top of the encoder model, tailored sequence label of the slot filling. For adapter-based approaches, only the parameters within the adapter modules are updated during training.

MULTI. A single model (or LoRA adapter) is trained on all languages simultaneously, optimizing over the entire dataset $\mathcal{D}$:

$$\theta_{\text{MULTI}} = \arg\max_{\theta} \sum_{t=1}^{T} \log p(D_t \mid \theta). \quad (1)$$

MONO. Each language/task has its own independently trained model θ_t:

$$\theta_t = \arg\max_{\theta} \log p(D_t \mid \theta), \quad (2)$$

$$\forall t \in \{1, \ldots, T\}. \quad (3)$$

VANILLA. A single model is trained incrementally, updating parameters sequentially:

$$\theta_{t+1} \leftarrow f(\theta_t, D_t), \forall t \in \{1, \ldots, T-1\}. \quad (4)$$

SHARED LoRA. A single LoRA adapter ϕ is trained while keeping the base model θ_0 frozen:

$$\phi_s \leftarrow f(\phi'_t, D_t), \theta = \theta_0, \forall t \in \{1, \ldots, T-1\}. \quad (5)$$

NON-SHARED LoRA. Each language has its own separate LoRA adapter ϕ_t, while keeping the base model θ_0 frozen:

$$\phi_t = \arg\max_{\phi} \log p(D_t \mid \theta_0, \phi), \forall t \in \{1, \ldots, T\}. \quad (6)$$

The specific ordering of languages used in the VANILLA is specified in Appendix Table 3.

2.3 Model Parameters Adaptation

We utilize low-rank adapters LoRA (Hu et al., 2021) for training parameters to analyze the effectiveness to have sharing parameters. It is a parameter-efficient fine-tuning method for large pre-trained models leveraging the intrinsic low-dimensionality of parameter updates, reducing the need for full model adaptation. Instead of modifying dense layers directly, it freezes the pre-trained

weights and introduces trainable low-rank matrices, significantly minimizing the number of learnable parameters and enhancing fine-tuning efficiency.

Formally, given a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA constrains the update ΔW to a low-rank decomposition:

$$\Delta W = BA, \tag{7}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll \min(d, k)$. This decomposition ensures that only A and B are updated while W_0 remains fixed. Consequently, the forward pass is expressed as:

$$h = W_0 x + \Delta W x = W_0 x + BAx, \tag{8}$$

where x is the input vector, and h is the output. The low-rank update $\Delta W x$ is scaled by a constant factor $\frac{\alpha}{r}$, analogous to a learning rate, to regulate the magnitude of the update. LoRA offers key advantages: it enhances *memory* and *computational efficiency* by limiting trainable parameters, reducing resource requirements, and enabling modular fine-tuning. Its linear structure ensures *no additional inference latency* and allows seamless integration. By leveraging low-rank adaptation, LoRA enables scalable and efficient model adaptation without compromising previously learned tasks.

3 Experimental Setup

3.1 Datasets

We utilize the MASSIVE, multilingual slot filling dataset (FitzGerald et al., 2023), which encompasses 52 languages and provides structured information, including scenarios, intents, utterances, and annotated utterances. Each language is uniformly represented, with 11.5K training samples, 2.03K validation samples, and 2.97K test samples.

3.2 Hyper-parameters

The training setup employed different configurations depending on whether LoRA was used. For models trained with LoRA, a learning rate of 5×10^{-6} was applied, whereas models without LoRA used a higher learning rate of 5×10^{-5}. The number of training epochs is 100 for models with LoRA, and 50 for those without. Early stopping was implemented in both settings, with a patience of 15 epochs for LoRA and 5 epochs for non-LoRA models, based on the F1-score on validation data. The LoRA configuration included a dropout rate of 0.1, and the scaling factor α was set equal to the rank (32, 64, 256 respectively).

3.3 Evaluation Metrics

We evaluate the performance of the model using average F1 score for the learned tasks and visualized its progression over number of learned languages, as illustrated in Figure 2. Besides that, there are additional metrics, particularly for sequential methods such as `VANILLA` and `SHARED LoRA`.

3.3.1 Performance Shift

This metrics is used to measure the average performance shift, which quantifies the change in a previously learned language performance after training in a new language. Formally, we define the average performance change as follows:

$$\mathcal{P}_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N}(\mathcal{P}_t - \mathcal{P}_{t+1}), \tag{9}$$

where $\mathcal{P}_t$ and $\mathcal{P}_{t+1}$ represent the average F1 score over all previously encountered tasks at time steps t and $t+1$, respectively. To account for variability in task sequences, the performance changes are averaged over five times ($N = 5$).

3.4 Cross-lingual Transfer

We assess cross-lingual transfer effectiveness using Cross-lingual Forward Transfer (CFT) and Cross-lingual Backward Transfer (CBT) metrics from Winata et al. (2023a) and we introduce a new metric, Multi-Hop Forward Transfer (MFT), and Multi-Hop Backward Transfer (MBT) to measure the multi-hop transfer for each language. Let $R \in \mathbb{R}^{T\times T}$ be a matrix where $R_{i,j}$ represents the test score performance on task t_j after training on the last sample from task t_i. The two types of metrics are defined as follows.

Cross-lingual Forward Transfer (CFT). The metric evaluates the model's ability to generalize to unseen languages by assessing test performance on tasks not encountered during training. It is defined as:

$$CFT = \frac{1}{T-1}\sum_{i=1}^{T-1}\bar{X}_i, \tag{10}$$

where

$$\bar{X}_i = \frac{1}{T-i}\sum_{j=i+1}^{T} R_{i,j}. \tag{11}$$

Here, $\bar{X}_i$ represents the average performance across future tasks ($t_j > t_i$).

Cross-lingual Backward Transfer (CBT). The metric measures the impact of learning a new task t_i on the performance of previously learned tasks. It is formally defined as:

$$CBT = \frac{1}{T-1} \sum_{i=1}^{T-1} (R_{T-1,i} - R_{i,i}) . \quad (12)$$

This metric quantifies the extent of catastrophic forgetting, where adding a new task may negatively impact the performance of past tasks.

Multi-Hop Forward Transfer (MFT). The metric measures the knowledge transfer effect between tasks separated by multiple learning steps. For a hop distance h, MFT is defined as:

$$MFT_h = \frac{1}{|L|} \sum_{l \in L} (\mathcal{P}_{i+h} - \mathcal{P}_{i-1}), \quad (13)$$

where $\mathcal{P}_i$ represents the average performance on tasks seen up to step i. This metric quantifies how learning a language affects performance on another language that will be encountered h steps later in the training sequence.

Multi-Hop Backward Transfer (MBT). The metric similarly evaluates the effect of learning a new task on the performance of tasks encountered several steps earlier. For a hop distance h, MBT is defined as:

$$MBT_h = \frac{1}{|L|} \sum_{l \in L} (\mathcal{P}_i - \mathcal{P}_{i-h-1}). \quad (14)$$

This metric measures how training on a language affects the performance on languages that were learned h steps before in the training sequence. The term *multi-hop* refers to our evaluation across multiple hops, as illustrated in Figure 5. A hop distance of zero corresponds to the performance change metric.

4 Results

Figure 3 illustrates the impact of training different languages sequentially on model performance towards learned language, measured by the average F1 change across 5 different orders.

Performance vs. Model Parameters. Table 1 presents a comparison of training methods in terms of average F1 score and trainable parameters. The `MULTI` method achieves the best overall performance (75.03%) with a much less parameter footprint (278.04M) compared to `MONO`'s, offering an excellent balance between effectiveness and efficiency. On the opposite end, `MONO`, which trains an entirely separate model per language, consumes an enormous parameter budget (14,458.27M) while yielding only moderate performance (72.98%), highlighting the inefficiency of isolated training.

Among parameter-efficient alternatives, LoRA-based approaches exhibit varying trade-offs. `NON-SHARED LoRA` performs competitively (up to 73.16% at rank 256), benefiting from task-specific specialization, albeit with moderate parameter cost (2,228.75M). In contrast, `SHARED LoRA`'s best result dramatically reduces the number of trainable parameters (e.g., 10.72M at rank 256) but suffers heavily in performance, dropping to as low as 61.26%.

Method	Params (M)	F1 (%)	Language Vitality		
			Low	Mid	High
MULTI	278.04	**75.03**	75.42	**75.84**	72.63
$r = 32$	5.36	74.19	74.27	**75.17**	71.83
$r = 64$	10.72	73.79	74.00	**74.56**	71.73
$r = 256$	42.86	74.11	74.16	**74.83**	72.41
MONO	14,458.27	72.98	73.66	**74.11**	69.43
VANILLA	278.04	66.16	65.70	**67.65**	63.46
SHARED LoRA					
$r = 32$	5.36	60.24	59.35	**62.34**	56.75
$r = 64$	10.72	**61.26**	60.55	**63.37**	57.48
$r = 256$	42.86	60.16	59.06	**62.15**	57.22
NON-SHARED LoRA					
$r = 32$	278.04	72.14	72.42	**73.39**	68.89
$r = 64$	557.19	72.38	72.55	**73.48**	69.65
$r = 256$	2,228.75	**73.16**	73.82	**74.26**	69.73

Table 1: Comparison of methods based on trainable parameters (in million parameters) and averaged F1 (%). Lower trainable parameters is better, higher average performance is better.

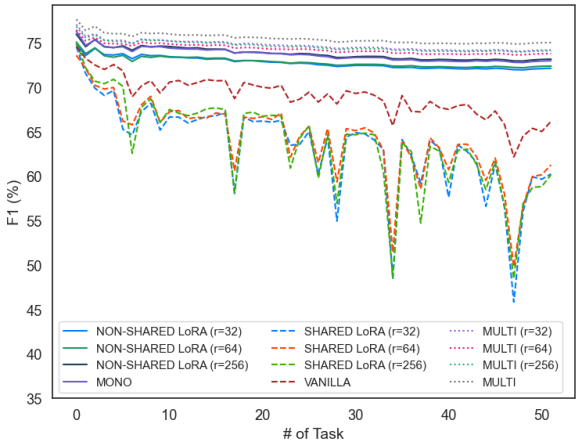

Figure 2: Performance results after training each language over the time.

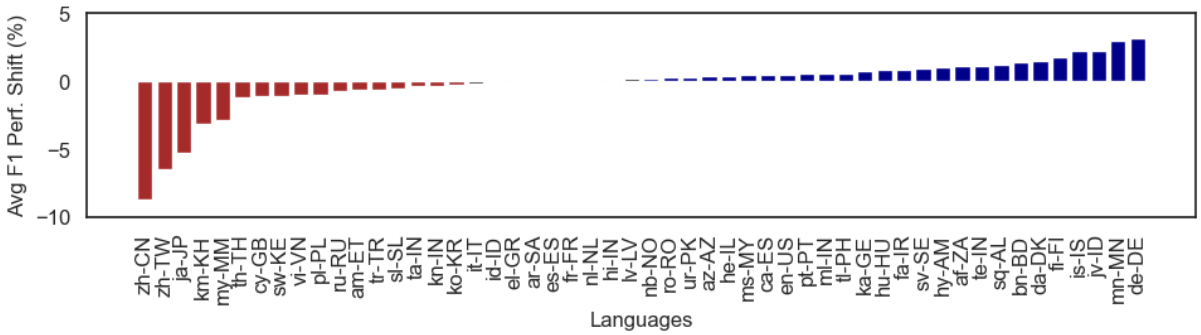

Figure 3: Performance change after training a certain language on x-axis in sequential training (VANILLA). Chinese (zh-CN) exhibits the most significant performance decline, while German (de-DE) serves as the most effective donor language, enhancing overall performance.

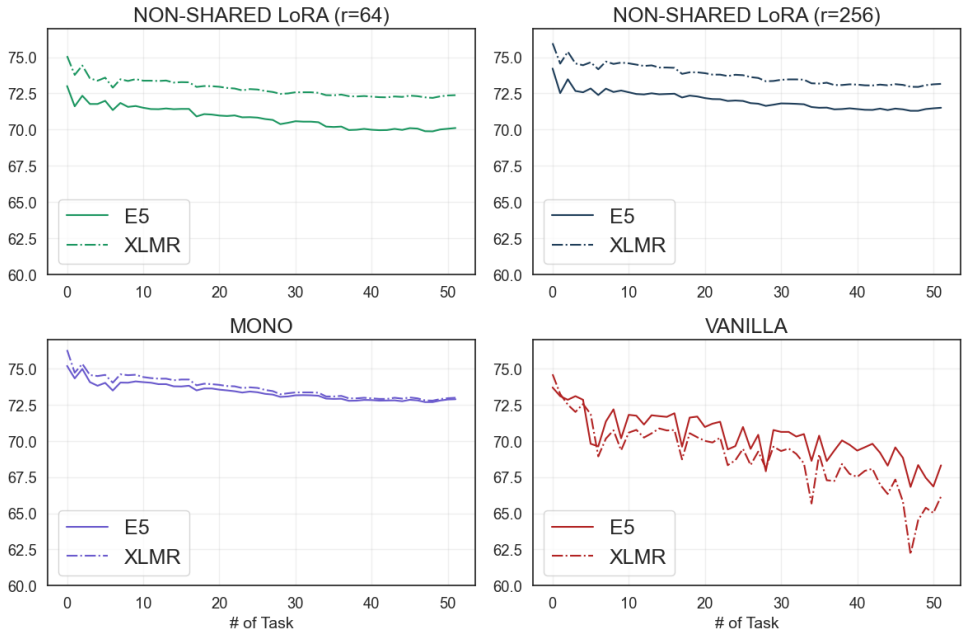

Figure 4: Comparison results between XLM-R and E5 models.

Crucially, increasing the LoRA rank—while expanding the model's capacity—does not substantially improve performance. For instance, MULTI with rank 32 (74.19%) performs nearly as well as at rank 256 (74.11%), and similar diminishing returns are observed across both SHARED and NON-SHARED LoRA. This trend extends to transfer metrics: Table 2 shows that higher rank under SHARED LoRA does not significantly improve forward transfer—CFT remains within the narrow band of 0.51–0.53. These results highlight a key trade-off: higher trainable parameters generally improve performance, but the efficiency of parameter usage varies across methods. The MULTI method provides the best balance between parameter efficiency and performance, while LoRA-based approaches demonstrate potential for parameter-efficient training at the cost of reduced performance. However, it should be noted that the MULTI method might not be trainable in parallel like the NON-SHARED LoRA method. Hence, in some scenarios, the NON-SHARED LoRA method should be considered.

Trends Between Models. Figure 2 illustrates how different training strategies affect performance over time. A key trend is that MULTI method (dotted lines), trained jointly on all languages, exhibit consistent performance, maintaining F1-scores above 73% throughout training. In contrast, sequential learning models show clear signs of degradation as training progresses. The VANILLA model suffers from moderate catastrophic forgetting, with F1-score reductions of 10–15 points. SHARED LoRA

fares worse, degrading by as much as 15–30 points across tasks. Meanwhile, NON-SHARED LoRA offers more stable performance across steps, ranging between 70–73% and demonstrating greater resilience to forgetting.

These observations are further supported by Table 2, which reports backward and forward transfer scores. The VANILLA model achieves a CBT of $-\mathbf{0.08}$ and CFT of $\mathbf{0.55}$, suggesting that while it suffers from forgetting, it still generalizes reasonably well to future tasks. SHARED LoRA, however, shows consistently more negative CBT scores (-0.13 to -0.14), confirming its vulnerability to catastrophic forgetting. This performance is also reflected in CFT, where the scores are also lower than VANILLA method. Together, these findings underscore the importance of balancing task generalization and knowledge retention, particularly in continual cross-lingual setups.

Method	CBT	CFT
VANILLA	**-0.08**	**0.55**
SHARED LoRA		
$r = 32$	-0.13	0.52
$r = 64$	-0.12	0.53
$r = 256$	-0.14	0.51

Table 2: CBT and CFT metrics for VANILLA and SHARED LoRA models — higher values indicate better performance.

Comparison XLM-R and E5 Models. Figure 4 presents a comparison of XLM-R and E5 models across different training methods. Despite variations in methodology, the general pattern of results remains consistent across models. Overall, XLM-R performs better than E5, except in VANILLA method where E5 tends to outperform XLM-R_{BASE}, though performance degradation due to forgetting is still evident. The results suggest that while different methods and model architectures influence the degree of forgetting, the overall trend of performance degradation remains a common characteristic across all settings.

5 Analysis on Languages

To frame our analysis, we interpret MFT as measuring a language's ability to **donate** knowledge to subsequent languages, while MBT reflects how well a language **receives** and retains knowledge after subsequent training steps. This donor-receiver perspective allows us to reason about asymmetries in cross-lingual transfer.

5.1 Languages Affect Forgetting

The results reveal that certain languages significantly impact the model's capacity to retain prior knowledge. Training on Chinese (zh-CN), Japanese (ja-JP), and Traditional Chinese (zh-TW) consistently leads to the most pronounced cases of catastrophic forgetting. This is evidenced by their strongly negative MBT values in Figure 5 and severe performance degradation in Figure 3, particularly when these languages are introduced later in the training sequence. As receivers, these languages appear highly vulnerable to interference from prior tasks. More detailed explanation can be seen in Appendix A. In contrast, languages such as Norwegian (nb-NO), Catalan (ca-ES), Portuguese (pt-PT), and Greek (el-GR) show some of the highest MBT scores across hop distances. These languages maintain stability when trained after others and also preserve prior task performance, indicating they are robust receivers. Interestingly, they may also act as indirect donors by not interfering with earlier knowledge.

However, not all performance trends align perfectly with MBT. For example, German (de-DE) appears beneficial in performance drop metrics (Figure 3), yet does not rank highly in MBT. This suggests that its apparent advantage may be due to its position in the training sequence—e.g., being trained before high-forgetting languages—rather than any inherent ability to preserve earlier knowledge. This underscores an important point: interpreting language influence solely through performance drop can be misleading. MBT offers a more principled, sequence-agnostic perspective on which languages genuinely aid in preserving prior knowledge and resisting catastrophic forgetting.

5.2 Latin vs. Non-Latin Scripts

Script similarity plays a significant role in cross-lingual knowledge transfer. In both MFT and MBT heatmaps, we observe that languages using Latin scripts—such as es-ES, fr-FR, and de-DE—tend to be strong donors and stable receivers. They benefit more from training on other languages and also suffer less from catastrophic forgetting. This likely reflects greater subword token overlap and lexical similarity, which help preserve learned representations under shared tokenization.

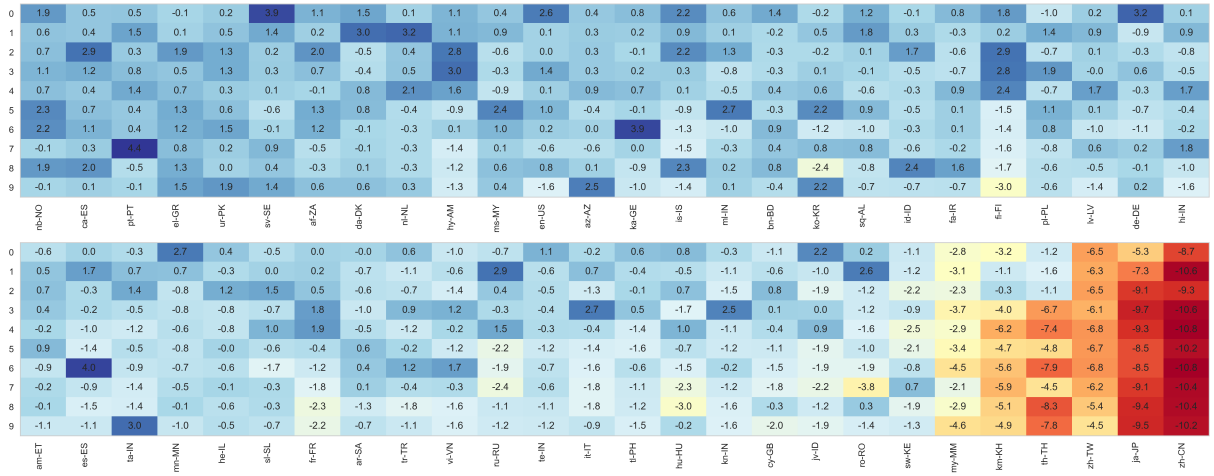

Figure 5: Heatmap of Multi-hop Backward Transfer (MBT), illustrates how training on later languages affects earlier ones over increasing hop distances (y-axis: 0–9). Cooler colors indicate positive backward transfer, while warmer colors reflect degradation in performance. Orders of the language is sorted descending (read from top-left to bottom-right) based on its average over all hops.

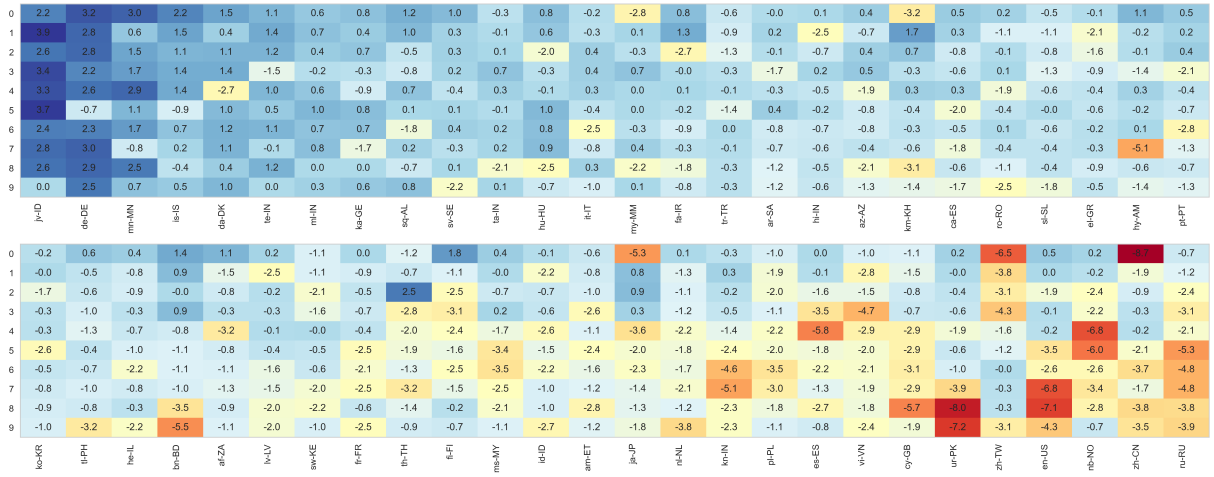

Figure 6: Heatmap of Multi-hop Forward Transfer (MFT), represents each language's ability to donate knowledge to subsequent tasks over increasing hop distances (y-axis: 0–9). Cooler colors indicate stronger positive transfer, while warmer colors reflect limited or negative influence on future learning. Orders of the language is sorted descending (read from top-left to bottom-right) based on its average over all hops.

In contrast, non-Latin script languages, especially those using logographic (e.g., `zh-CN`) or abugida scripts (e.g., `th-TH`, `hi-IN`), tend to be weak donors and vulnerable receivers. These languages show low MFT—suggesting limited forward transfer to other tasks—and highly negative MBT, indicating susceptibility to forgetting. The subword tokenizer, likely optimized for Latin-based alphabets, aggravates this imbalance. This highlights a fundamental challenge for multilingual continual learning: shared vocabulary spaces can lead to representational dominance of Latin-script languages, marginalizing others.

5.3 Language Family

While language family information is not explicitly modeled, typologically or lexically similar languages often demonstrate mutual reinforcement in transfer. Under the donor-receiver lens, we observe that Romance languages such as `es-ES`, `pt-PT`, and `fr-FR` frequently act as strong donors (high MFT) and reliable receivers (stable MBT), especially when positioned near each other in the training sequence. Similarly, Germanic languages like `nl-NL`, `sv-SE`, and `de-DE` show stable transfer interactions.

However, these patterns are not universal. The apparent family-related benefits may arise from

shared scripts and vocabulary rather than deep structural similarity. For instance, several Indo-European languages from different branches perform well together, likely due to orthographic overlap. Conversely, languages from distant families—such as Sino-Tibetan (zh-CN), Austroasiatic (km-KH), or Afro-Asiatic (ar-SA)—often act as poor receivers (low MBT) and limited donors (low MFT), especially when sequenced after typologically dissimilar languages. Future work could explicitly incorporate phylogenetic distances to better disentangle the impact of language family on multilingual continual learning.

5.4 Language Vitality

Language vitality—encompassing speaker population, data availability, and digital presence—also plays a nuanced role in continual learning dynamics. As receivers, high-vitality languages such as zh-CN, ja-JP, and hi-IN (Joshi et al., 2020) show some of the most negative MBT scores, indicating that they are especially vulnerable to forgetting. Surprisingly, they also make relatively poor donors, as reflected in lower MFT scores compared to more typologically compatible mid-vitality languages.

This counterintuitive trend is clarified in Table 1, where mid-vitality languages (Joshi et al., 2020) consistently achieve the highest F1 scores across model variants. These languages appear to strike a balance: they share enough structure with other languages to act as effective donors, while remaining resilient as receivers under sequential training. In contrast, high-vitality languages—despite abundant resources—struggle under parameter-efficient continual learning setups. Their unique token distributions and structural divergence make them harder to adapt to and easier to overwrite. These findings suggest that vitality-aware scheduling or modularization may be critical for improving cross-lingual robustness in continual learning scenarios.

6 Related Work

Catastrophic forgetting is a significant challenge in neural networks, where models lose previously acquired knowledge when fine-tuned on new tasks (McCloskey and Cohen, 1989). This issue is particularly pronounced in multilingual contexts, as models must adapt to new languages without degrading performance on previously learned ones (Winata et al., 2023a). To mitigate this, various strategies have been proposed, including memory replay (Rolnick et al., 2019), regularization techniques (Kirkpatrick et al., 2017), and architectural innovations like progressive networks (Rusu et al., 2016).

Lifelong learning also known as continual learning, is an emerging approach that enables models—particularly LLMs and their agents—to continuously acquire new knowledge while retaining prior capabilities. This knowledge can be integrated into LLMs either by updating model parameters through training or adapters, or by leveraging external sources like Wikipedia or tools without modifying the model itself or knowledge base (Zheng et al., 2024). Recent work extends lifelong learning to agent-based settings, decomposing it into perception, memory, and action modules that together support continuous adaptation (Zheng et al., 2025).

For internal knowledge updates, adapters have proven to be a lightweight and effective solution, introducing small, task-specific modules that can be fine-tuned independently, reducing interference across tasks (Houlsby et al., 2019; Winata et al., 2021; Hu et al., 2021). The MAD-X framework (Pfeiffer et al., 2020b) enhances cross-lingual transfer by separating language and task adaptation, while language-specific adapters balance specialization and sharing (Badola et al., 2023). Additionally, methods like AdapterFusion (Pfeiffer et al., 2020a) combines task-specific adapters through a learned composition layer, promoting parameter sharing and effective transfer learning while minimizing forgetting.

7 Conclusion

Our paper highlights the critical challenges of catastrophic forgetting in cross-lingual transfer for multilingual NLP models with 52 languages. We provide insights into how various parameter-sharing strategies can influence knowledge retention and overall model performance. Our findings indicate that partial parameter sharing can effectively mitigate forgetting while maintaining performance, presenting a promising approach for developing more robust multilingual NLP systems. Additionally, we identify that certain languages during training can negatively impact performance, contributing to catastrophic forgetting. Overall, this research enhances the ongoing efforts to improve the adaptability and efficiency of NLP models in real-world NLP applications.

Limitations

In this paper, we concentrate our investigation on XLM-R model and use E5, rather than exhaustively evaluating every possible model due to resource constraints. This focused approach allows us to provide a more in-depth analysis of these models and their performance in cross-lingual contexts.

Ethical Considerations

In our evaluation of language models for multilingual tasks, we place strong emphasis on transparency and fairness. We carefully design and document our data collection and evaluation methodologies to ensure they are consistent, unbiased, and reproducible. By applying uniform assessment criteria across models, we aim to enable meaningful and equitable comparisons.

References

Kartikeya Badola, Shachi Dave, and Partha Talukdar. 2023. Parameter-efficient finetuning for robust continual multilingual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9763–9780.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2023. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Bangkok, Thailand (Online). Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier.

Meryem M'hamdi, Xiang Ren, and Jonathan May. 2023. Cross-lingual continual learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3908–3943, July 9-14, 2023. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, et al. 2016. Progressive neural networks. In *arXiv preprint arXiv:1606.04671*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2023a. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 768–777.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, et al. 2023b. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi. 2021. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, page 361.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 2024. Towards lifelong learning of large language models: A survey. *arXiv preprint arXiv:2406.06391*.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. 2025. Lifelong learning of large language model based agents: A roadmap. *arXiv preprint arXiv:2501.07278*.

A Detailed Results

A.1 Language Order

Table 3 presents the language orders used in the sequential training experiments. These orders are used to train models in a step-by-step fashion, where each iteration introduces a new language. The results from these training sequences are subsequently used to compute aggregate metrics, as shown in Figure 2 and Figure 4.

The first order is derived based on the amount of language resources available in the XLM-R model (Conneau et al., 2020). This order reflects the relative training data size used during XLM-R's pretraining, with high-resource languages appearing earlier in the sequence. The remaining orders (2 through 5) are randomly shuffled variants to introduce diversity and reduce potential order bias. However, in the fifth order, languages that are found to be particularly destructive—i.e., those that tend to cause performance degradation on previously learned languages—are deliberately placed toward the end of the sequence. This design allows us to analyze how the position of destructive languages affects knowledge retention and transfer in sequential multilingual training.

A.2 Heatmap on `VANILLA` method for first language order

The heatmap on Figure 7 provides a detailed visualization of the model's performance across training iterations (represented by rows) and evaluated languages (represented by columns). In each iteration, the model is trained on a new language. For instance, as shown in the figure, the first iteration trains on `en-US`, the second on `ru-RU`, the third on `id-ID`, and so forth. After training on a language, the model's performance on that language typically improves. This trend is reflected in the heatmap: the lower-left triangle (below the diagonal), corresponding to previously learned languages, tends to display cooler colors, indicating better performance; in contrast, the upper-right triangle (unlearned languages) often exhibits warmer colors, reflecting performance degradation.

This visualization clearly highlights cross-lingual interactions—specifically, how training on a new language can either benefit or harm performance on other languages. For example, in row 18, where the model is trained on `zh-CN`, the corresponding row becomes noticeably warmer compared to previous iterations, suggesting a general decline in performance across many languages. However, for linguistically related languages such as `ja-JP`, where many Kanji characters overlap with Chinese characters (hence vocabulary overlap), performance actually improves. This suggests that while `zh-CN` introduces interference for many languages, it serves as a helpful donor for `ja-JP`—likely due to shared orthographic features, such as the incorporation of Chinese characters in the Japanese writing system.

Order	Languages in ISO 639-1
1	en-US, ru-RU, id-ID, vi-VN, fa-IR, th-TH, ja-JP, de-DE, ro-RO, hu-HU, fr-FR, fi-FI, ko-KR, es-ES, pt-PT, nb-NO, el-GR, zh-CN, da-DK, pl-PL, he-IL, it-IT, nl-NL, ar-SA, tr-TR, hi-IN, zh-TW, ta-IN, sv-SE, sl-SL, ca-ES, ka-GE, lv-LV, ms-MY, bn-BD, ml-IN, az-AZ, ur-PK, hy-AM, sq-AL, te-IN, kn-IN, is-IS, tl-PH, mn-MN, my-MM, sw-KE, km-KH, af-ZA, am-ET, cy-GB, jv-ID
2	tr-TR, ro-RO, ur-PK, es-ES, hi-IN, pl-PL, hy-AM, sv-SE, sl-SL, ta-IN, te-IN, ml-IN, id-ID, ka-GE, el-GR, ko-KR, de-DE, fa-IR, ms-MY, ca-ES, az-AZ, nl-NL, pt-PT, fr-FR, hu-HU, sw-KE, mn-MN, he-IL, zh-CN, fi-FI, ru-RU, is-IS, cy-GB, ja-JP, sq-AL, vi-VN, th-TH, jv-ID, it-IT, my-MM, kn-IN, lv-LV, am-ET, nb-NO, ar-SA, en-US, af-ZA, zh-TW, bn-BD, da-DK, km-KH, tl-PH
3	sv-SE, nl-NL, fi-FI, kn-IN, hu-HU, ms-MY, es-ES, my-MM, is-IS, ko-KR, af-ZA, vi-VN, bn-BD, tr-TR, tl-PH, lv-LV, ru-RU, fr-FR, en-US, ro-RO, am-ET, he-IL, hi-IN, ja-JP, te-IN, id-ID, ta-IN, it-IT, jv-ID, nb-NO, ka-GE, sq-AL, ca-ES, az-AZ, zh-TW, fa-IR, mn-MN, zh-CN, de-DE, da-DK, ml-IN, sw-KE, sl-SL, km-KH, ar-SA, pt-PT, cy-GB, ur-PK, hy-AM, el-GR, pl-PL, th-TH
4	nb-NO, ta-IN, th-TH, fi-FI, ru-RU, af-ZA, vi-VN, ko-KR, ro-RO, km-KH, is-IS, ms-MY, sl-SL, en-US, hi-IN, he-IL, bn-BD, pt-PT, fa-IR, sv-SE, am-ET, kn-IN, az-AZ, tl-PH, ar-SA, nl-NL, cy-GB, hy-AM, it-IT, de-DE, da-DK, te-IN, hu-HU, lv-LV, zh-CN, mn-MN, es-ES, ca-ES, pl-PL, fr-FR, ja-JP, ka-GE, sw-KE, id-ID, zh-TW, jv-ID, sq-AL, el-GR, tr-TR, my-MM, ml-IN, ur-PK
5	mn-MN, ml-IN, is-IS, fa-IR, az-AZ, pl-PL, de-DE, ko-KR, ar-SA, sw-KE, jv-ID, sq-AL, tl-PH, ru-RU, lv-LV, fr-FR, ro-RO, ka-GE, cy-GB, tr-TR, he-IL, sl-SL, af-ZA, nl-NL, my-MM, hu-HU, hi-IN, vi-VN, it-IT, pt-PT, da-DK, ca-ES, am-ET, el-GR, ta-IN, id-ID, te-IN, sv-SE, bn-BD, ur-PK, en-US, kn-IN, ms-MY, nb-NO, es-ES, fi-FI, zh-TW, zh-CN, ja-JP, th-TH, km-KH, hy-AM

Table 3: Language orders in the sequential training experiments.

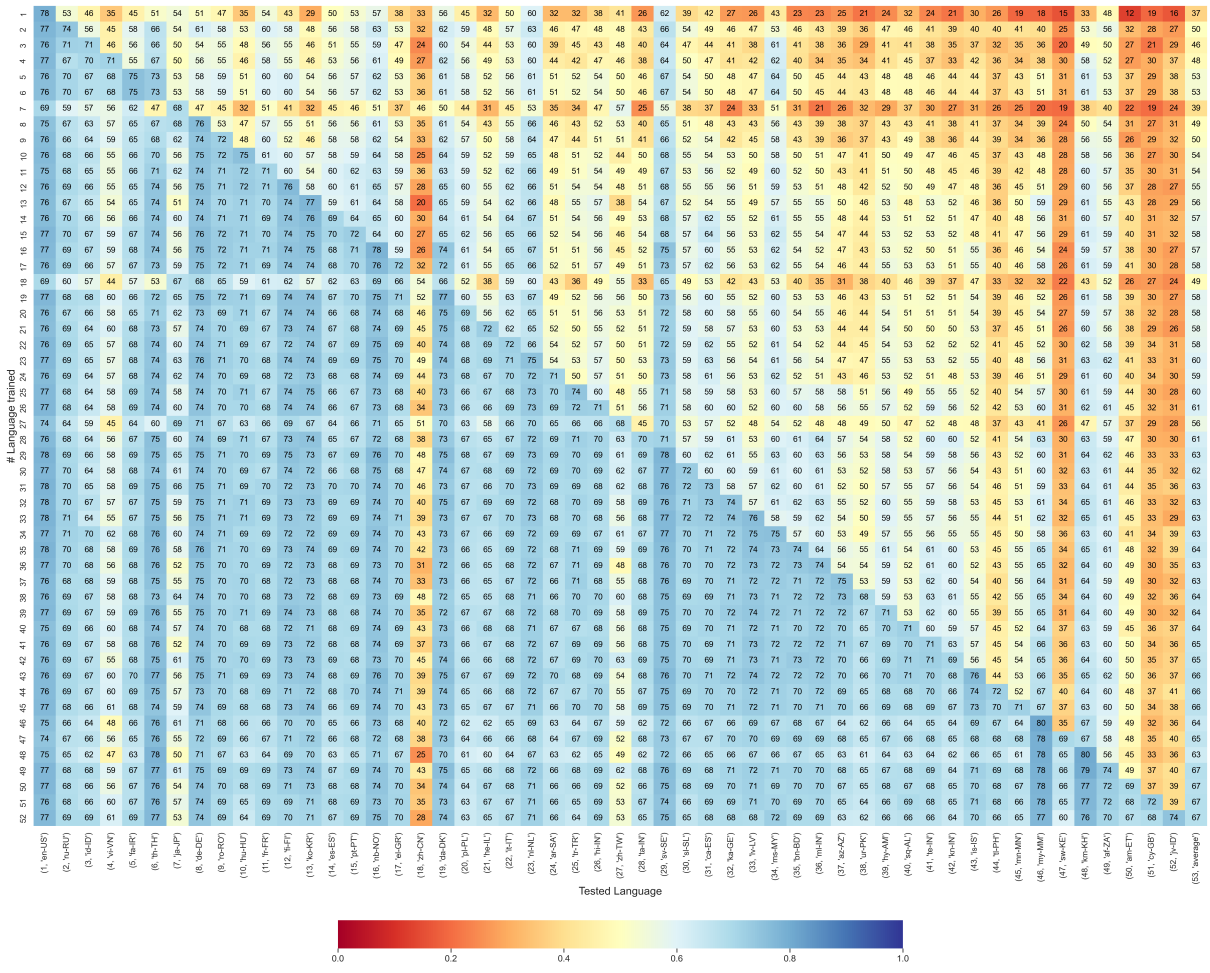

Figure 7: Heatmap on VANILLA method for first language order.

Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages

Siyu Liang and **Gina-Anne Levow**
University of Washington
liangsy, levow@uw.edu

Abstract

The development of Automatic Speech Recognition (ASR) has yielded impressive results, but its use in linguistic fieldwork remains limited. Recordings collected in fieldwork contexts present unique challenges, including spontaneous speech, environmental noise, and severely constrained datasets from underdocumented languages. In this paper, we benchmark the performance of two fine-tuned multilingual ASR models, MMS and XLS-R, on five typologically diverse low-resource languages with control of training data duration. Our findings show that MMS is best suited when extremely small amounts of training data are available, whereas XLS-R shows parity performance once training data exceed one hour. We provide linguistically grounded analysis for further provide insights towards practical guidelines for field linguists, highlighting reproducible ASR adaptation approaches to mitigate the transcription bottleneck in language documentation.

1 Introduction

Automatic Speech Recognition (ASR) has achieved significant breakthroughs in recent years, with deep learning-based models reported to reach near-human word error rates for high-resource languages (Radford et al., 2023; Baevski et al., 2020). However, these advancements have largely been driven by massive transcribed datasets (e.g. Chang et al. (2022); Panayotov et al. (2015); Godfrey et al. (1992)), leaving a substantial performance gap for low-resource languages, particularly those encountered in linguistic fieldwork (Guillaume et al., 2022a). Fieldwork speech data presents distinct challenges, including spontaneous speech, varied recording setups, and typologically diverse linguistic features, all of which could degrade the performance of ASR models trained on standardized speech corpora.

Linguistic fieldwork plays a critical role in preserving endangered languages and documenting linguistic diversity. These recordings capture not only the linguistic structures of a language, but also oral traditions, discourse patterns, and sociolinguistic variations (Himmelmann, 1998; Austin and Sallabank, 2011). However, while some well-researched low-resource languages have substantial datasets (Guillaume et al., 2022b; Przezdziak, 2024), there is usually limited data to bootstrap an ASR model for most field linguists. Evaluations of the ASR approaches usually tend to focus on one language (Jones et al., 2024; Rijal et al., 2024; Guillaume et al., 2022b; Mainzinger and Levow, 2024) or are inconsistent regarding data size, genre, etc. in the sample (Jimerson et al., 2023). Evaluations of models for low-resource languages also tend to favor clean, good quality, read speech (Rijal et al., 2024; Mainzinger and Levow, 2024; Jimerson et al., 2023), compared with the noisier and more spontaneous speech of fieldwork recordings.

1.1 The transcription bottleneck

Linguistic fieldwork plays a crucial role in documenting endangered and under-researched languages, yet the process of manually transcribing recordings remains a significant barrier: transcribing a single hour of audio in a newly documented language can require up to 50 hours of work (Shi et al., 2021). Moreover, many of the fieldwork languages also lack standardized orthographies, requiring a handful of trained linguists to make discerning decisions during transcription. As a result, the volume of untranscribed linguistic data continues to grow, creating a severe bottleneck in language documentation, analysis, and distribution (Anastasopoulos and Chiang, 2018; Bird, 2020; Thieberger, 2012).

The dependence on large transcribed datasets for training ASR models exacerbates this issue, as most endangered and low-resource languages

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 26–37
August 1, 2025 ©2025 Association for Computational Linguistics

lack sufficient annotated speech data to support model development (Levow et al., 2021). Without adequate transcriptions, traditional supervised ASR methods remain ineffective, requiring alternative approaches that can leverage limited data more efficiently (Dunbar et al., 2019; Baevski et al., 2020, 2021).

1.2 ASR for Low-resource Data

The application of ASR in linguistic fieldwork closely parallels the development of low-resource ASR research. Since the 2010s, much of this work has focused on languages from the IARPA Babel project, which served as a cornerstone for ASR development in low-resource settings (Miao et al., 2013; Cui et al., 2014; Grézl et al., 2014). Research leveraging Babel datasets introduced key techniques such as transfer learning, multilingual adaptation, and data augmentation, which have since become fundamental to ASR advancements in underdocumented languages (Zhang et al., 2014; Khare et al., 2021; Vanderreydt et al., 2022; Guillaume et al., 2022a).

The widespread adoption of the Kaldi toolkit (Povey et al., 2011) further propelled ASR research in these domains, enabling the development of reproducible pipelines and fostering the open distribution of Kaldi-compatible datasets (Yadava and Jayanna, 2017; Milde and Köhn, 2018; Adams et al., 2021; Zhang et al., 2022). Concurrently, researchers have explored approaches such as transfer learning and fine-tuning from multilingual pre-trained models (Guillaume et al., 2022a; Sikasote and Anastasopoulos, 2021) or adapting English-centric models to new linguistic domains (Kim et al., 2021; Thai et al., 2020). Additionally, self-supervised and semi-supervised learning approaches have gained traction as viable solutions for overcoming transcription scarcity, further bridging the gap between ASR and field linguistics (Babu et al., 2021; Baevski et al., 2021).

1.3 Fine-tuning Pre-trained ASR Models

Fine-tuning pre-trained ASR models has emerged as a key approach for improving recognition accuracy in low-resource settings, particularly for linguistic fieldwork recordings (Guillaume et al., 2022a; Pillai et al., 2024; Nowakowski et al., 2023). Self-supervised learning models, such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), MMS (Massive Multilingual Speech) Model (Pratap et al., 2024), and XLS-R (Babu et al., 2021), have demonstrated the ability to learn generalized speech representations from large-scale multilingual datasets, significantly reducing the need for extensive transcriptions in under-documented languages. Studies on specific low-resource languages, such as Bribri (Coto-Solano, 2021), Japhug (Guillaume et al., 2022b), Mvskoke (Mainzinger and Levow, 2024), and the Čakavian dialect of Croatian (Jones et al., 2024), underscore the benefits of adapting large multilingual models and report considerable reductions in error rates even with very limited data.

Nevertheless, recent work suggests that no single architecture or end-to-end approach consistently outperforms others under extremely low-resource conditions (Jimerson et al., 2023). Some studies advocate experimenting with multiple toolkits and hyperparameter configurations to identify solutions best suited to the language at hand. Indeed, while fully fine-tuning massive models can be effective, it often requires large amounts of computational resources, can risk overfitting with very small datasets, and demands updating millions of parameters.

Instead of modifying all model parameters, adapters introduce small trainable layers while keeping the base pre-trained model frozen, thereby reducing memory requirements and improving efficiency (Houlsby et al., 2019). This makes them particularly useful for linguistic fieldwork applications, where data is scarce and computational resources are limited. The MMS model developed by Meta (Pratap et al., 2024) integrates adapter layers specifically designed for ASR, enabling efficient adaptation to new languages with minimal training data. Studies in low-resource settings (Bai et al., 2024; Mainzinger and Levow, 2024) have shown that adapter-based fine-tuning can achieve performance comparable to full fine-tuning while requiring significantly fewer trainable parameters. By avoiding overfitting on small datasets and focusing on the most relevant parameters for language adaptation, adapter-based methods offer an attractive balance between accuracy and efficiency, an approach increasingly vital to sustaining language documentation efforts in the face of extremely sparse resources.

In contrast to earlier studies that often focus on a single language or on clean, scripted corpora, our work systematically evaluates both MMS and XLS-R in truly low-resource fieldwork conditions spanning multiple typologically diverse languages.

By examining noise-heavy, spontaneous recordings rather than controlled speech, we test model adaptability in settings that more accurately reflect real-world linguistic documentation. Further, our fine-grained error analysis explores how each model handles the nuanced phonological features that typify endangered and under-documented languages—details that have often been overlooked in prior research. Together, these innovations provide a clearer roadmap for linguists seeking practical ASR solutions under extreme data scarcity and diverse orthographic conventions.

2 Data

We test the performance of fine-tuned ASR models on five typologically varied low-resource languages: Cicipu (ISO639-3: awc, McGill (2012)), Mocho' (ISO639-3: mhc, Pérez González (2018)), Toratán (ISO639-3: rth, (Jukes, 2010)), Ulwa (ISO639-3: yla, (Barlow, 2018a)), and Upper Napo Kichwa (ISO639-3: quw, (Grzech, 2020)). These languages span multiple language families and exhibit distinct phonetic, phonological, and morphological features. The data is drawn from the Endangered Languages Archive (ELAR)[1], where gold-standard transcriptions can be derived from the recordings and the corresponding time aligned transcriptions in the ELAN (Brugman and Russel, 2004) format. The dataset encompasses a variety of genres, such as greetings, narratives, ritual discourse, interviews, elicitation sessions, folktales, and cultural practices, the details of which are given in Table 5 of Appendix B. Table 1 provides an overview of key linguistic features, including vowel and consonant inventories as well as tonal systems.

2.1 Data details

Given that the recordings were made in naturalistic fieldwork environments, they exhibit acoustic idiosyncrasies that could pose significant challenges for ASR. There's background noise from outdoor settings, such as wind, animals, and community sounds, in many of the recordings. We also observe code-switching, usually in the regional dominant languages. For example, Toratán speakers are also speakers of Manado Malay (with various degrees of fluency), and loans from Malay are generally not adapted to Toratán phonology (Himmelmann and Wolff, 1999). We also observe significant Spanish code-mixing in Mocho', as well as some English content in Cicipu.

The dataset sizes vary, with archived speech ranging from approximately 2 to 22 hours per language. However, not all archived data have been transcribed. This variation reflects real-world constraints in linguistic fieldwork, where some languages have more extensive documentation than others.

2.2 Dataset Pre-processing

All recordings were resampled to 16 kHz (the training sampling rate for both MMS and XLS-R), converted to mono-channel WAV format, and aligned with their corresponding transcriptions. During transcription pre-processing, we referenced the phonological description of each language to ensure that punctuation marks or special characters used to denote phonological features were retained (see Table 5 of Appendix B). Audio segments explicitly transcribed as non-linguistic sounds, such as laughter, were excluded from the dataset. We also removed utterances that contain only filler words, such as 'mhm', 'aaa', etc. A breakdown of the audio lengths before and after pre-processing is given in Table 6 of Appendix B

We created four total train+dev duration configurations for each language—10, 30, 60, and 120 minutes—before splitting the data into training (90%) and development (10%) sets. In addition, we set aside a fixed 10-minute test set for final evaluation. To maintain consistency and facilitate interpretation, larger dataset splits were structured as supersets of smaller ones. We did not designate a held-out speaker, as field linguists typically work with a limited number of consultants and would prioritize consistent model performance across familiar speakers (Liu et al., 2023). Details of the cleaned dataset are shown in Table 2. The last column of the table lists the number of unique characters used in the language. Due to different transcription conventions, features such as nasalization, vowel length and voicing could be indicated with diacritics or extra letters. Therefore, although transcriptions are meant to be phonemic, the number of unique characters might not match the number of contrastive vowels and consonants. In the case of Cicipu, its unusually large inventory of 93 characters is due to the number of all possible combinations of nasality, tone, and vowel quality marking.

[1] https://www.elararchive.org/

Language	Family	Region	#V	#C	#T	Features
Cicipu	Niger-Congo	Nigeria	28	27	4	Ṽ, Vː, Cː
Mocho'	Mayan	Mexico	10	27	2	Vː
Toratán	Austronesian	Indonesia	5	21	0	
Ulwa	Keram	Papua New Guinea	8	13	0	[-voice, +son]
Upper Napo Kichwa	Quechuan	Ecuador	8	20	0	Vː

Table 1: Linguistic data used for the study, showing language family, region spoken, phoneme inventory size, tones, and phonological features such as nasality, vowel length, consonant gemination, etc.

Language	#Spk	Avg. Leng.	#Char
Cicipu	33	2.1	93
Mocho'	6	2.0	29
Toratán	13	2.35	27
Ulwa	6	3.65	25
U.N. Kichwa	16	3.79	33

Table 2: Dataset statistics for different languages, including the number of speakers, average utterance length in seconds, and character inventory.

3 Methodology

This study investigates the effectiveness of fine-tuning multilingual ASR models to address the unique challenges posed by low-resource linguistic fieldwork recordings. By evaluating the performance of two state-of-the-art models, we aim to determine how fine-tuning can enhance recognition accuracy on typologically diverse, low-resource languages. In addition, we discuss the impact of key factors, including training data size, model choice, and pre-trained model features, to provide practical insights for ASR adaptation in fieldwork contexts.

3.1 Models

Our goal is to fine-tune state-of-the-art multilingual ASR models that have been pre-trained on large-scale speech corpora. Specifically, we evaluate models from Meta's Massively Multilingual Speech (MMS) project (Pratap et al., 2024) alongside XLS-R (Babu et al., 2021), a widely used multilingual ASR model.

MMS is based on the wav2vec 2.0 framework (Baevski et al., 2020), which employs self-supervised learning to extract generalized speech representations from vast amounts of unlabeled audio. The model has been trained on 1,406 languages, making it one of the most comprehensive ASR models for multilingual speech recognition. Specifically, we choose MMS-1B-l1107, a 1-billion parameter model fine-tuned specifically for ASR with an additional 2-million parameter adapter, which supports ASR in 1107 languages out of the box (Houlsby et al., 2019). The adapter facilitates efficient language-specific fine-tuning while preserving the generalized multilingual knowledge encoded in the base model.

XLS-R (Babu et al., 2021) is a multilingual ASR model pre-trained on 128 languages using the wav2vec 2.0 framework. It has been extensively used for low-resource ASR and cross-lingual transfer learning, making it a strong baseline for evaluating ASR performance in linguistic fieldwork settings. Unlike MMS, which is trained on over 1,000 languages, XLS-R has been optimized for a balanced selection of 128 languages, with a strong focus on phonetic diversity. This makes it particularly useful for comparison against MMS models to assess the effectiveness of scaling multilingual pre-training to extremely low-resource languages. The XLS-R-300m with 300 million parameters is chosen for the study.

None of the languages used in the study, with the exception of Upper Napo Kichwa, is represented in the training data of the two models. A discussion of the possible effects is included in Section 4.2.

3.2 Implementation

Following the fine-tuning procedure outlined by von Platen[2] and using the Hugging Face Transformers library (Wolf et al., 2020), we implement model-specific strategies. For MMS-1B-l1107, only the adapter layers are fine-tuned, with the base model frozen. For XLS-R, the entire model is fine-tuned. To assess the effect of data size, models are trained on subsets of 10, 30, 60, and 120 minutes of transcribed fieldwork data per language. Early stopping, based on the development set Character Error Rate (CER), mitigates overfitting. Hyperparameter details, including batch size, learning rate, opti-

[2] https://huggingface.co/blog/mms_adapters

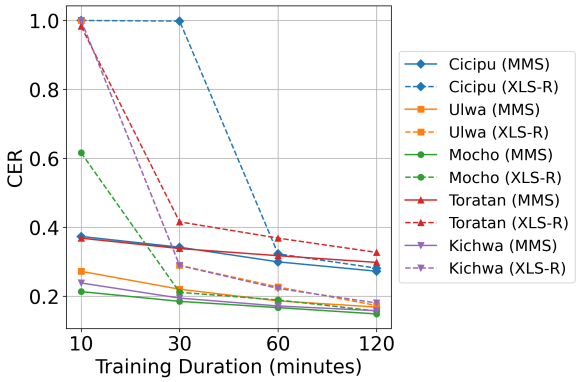

Figure 1: CER comparison for MMS-1b1107 and XLS-R-300m models across five languages. The MMS model performs markedly better under extremely low-resource settings (less than 1 hour), but XLS-R performs similarly well with 2 hours of data. Points are connected to aid trend reading and do not imply performance at intermediate durations.

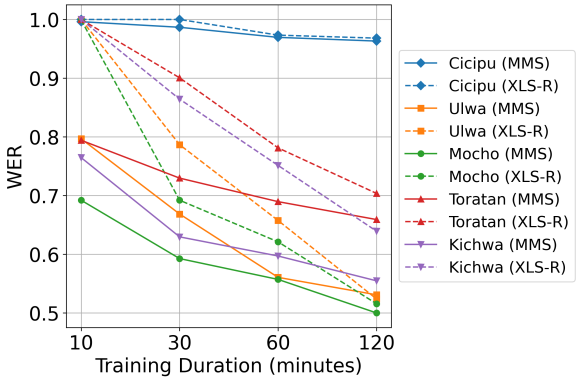

Figure 2: WER comparison for MMS-1b1107 and XLS-R-300m models across five languages.

mization strategy, and training details are provided in Appendix A.

4 Results

Overall, the performance of the MMS and XLS-R models in automatic speech recognition (ASR) tasks on low-resource language data is comparable, though nuanced differences emerge depending on the availability of training data. In general, the MMS model outperforms XLS-R under extremely low-resource conditions (i.e., less than one hour of transcribed data). However, XLS-R demonstrates a marked improvement as the size of the training data scales beyond this threshold, ultimately becoming on par with MMS. Figure 1 and Figure 2 illustrate these trends in Character Error Rate (CER) and Word Error Rate (WER) metrics across the five languages in this study.

In our analysis, CER serves as the primary evaluation metric. Unlike WER, which operates at the word level and often presupposes well-defined word boundaries and stable orthographic forms, CER better reflects the needs of field linguists, who often lack enough data to be able to use language models. In fieldwork contexts, the primary goal of transcription is to capture phonetic accuracy, particularly in languages without standardized orthographies. During model training, the best model metric is set to CER to ensure that model performance aligns with the core task of producing reliable phoneme-level transcriptions from spontaneous and noisy speech data.

4.1 Data size effect

The relationship between training data size and model performance is critical in understanding model suitability for field linguistics applications. For both MMS and XLS-R, performance improvements start to plateau when training data exceeds approximately one hour. This finding suggests steady although diminishing returns beyond this point, aligning with previous observations in low-resource ASR research (Guillaume et al., 2022a).

In scenarios where less than one hour of data is available, MMS consistently achieves lower error rates, likely due to its extensive multilingual pre-training on over 1,000 languages. For field linguists dealing with extremely limited resources, we thus recommend fine-tuning MMS to achieve acceptable ASR performance. However, when approximately one hour of data or more is obtainable, XLS-R becomes a more effective option due to its improving performance with increasing data volumes. This suggests that one hour of transcribed data serves as a practical threshold for developing a robust fine-tuned ASR system in fieldwork contexts.

It is worth noting that Cicipu exhibits particularly high error rates under extreme data scarcity, including a character error rate approaching 1.0 at 10 minutes of data for both models and at 30 minutes for XLS-R. Cicipu's unusually large orthographic inventory (93 unique characters reflecting combinations of nasality, tone, and vowel quality) requires more training examples to accurately learn the mapping from acoustics to graphemes. Consequently, with only a few minutes of labeled data, neither model can fully learn Cicipu's complex phonological and orthographical features.

4.2 MMS vs. XLS-R

The MMS model excels in settings with minimal data due to its multilingual pretraining on a vast corpus that includes low-resource languages. Moreover, unlike XLS-R—whose core version is primarily self-supervised and not initially fine-tuned for ASR—MMS has already undergone a large-scale ASR fine-tuning step. This means that MMS starts off with more task-specific parameters, making it more effective than XLS-R in extremely low-data regimes. However, MMS's reliance on primarily read speech data (e.g. Bible translations) may limit its adaptability to spontaneous speech environments, which are common in linguistic fieldwork recordings.

In contrast, XLS-R benefits from a more diverse training corpus that encompasses conversational and spontaneous speech, allowing it to generalize better once sufficient data becomes available. Indeed, Mainzinger and Levow (2024) reported superior performance of MMS over XLS-R when fine-tuning Mvskoke—likely due to both the advantage of MMS's ASR fine-tuning and the fact that much of the Mvskoke training material was similarly read or scripted speech.

Since several related dialects of Kichwa (Eberhard et al., 2024, 2025) were included in the MMS pre-training dataset, we investigated whether the performance gap between MMS and XLS-R would be *larger* for Kichwa than for the other languages in our study. Specifically, we fit a linear mixed-effects model (with random intercepts for each language) to our character error rate (CER) data, using *model* (MMS vs. XLS-R), *time*, and an indicator *similar* (1 = Kichwa, 0 = other languages) as fixed effects. If Kichwa had benefitted disproportionately from MMS's pre-training, we would have observed a significant positive interaction in the model. However, the interaction term (*model* $\times$ *similar*) was small and not statistically significant ($\beta = 0.021$, $p = 0.896$), indicating that while MMS outperforms XLS-R overall, the additional advantage for Upper Napo Kichwa is not discernibly greater than for the other languages.

Further research is needed to evaluate whether the performance trend continues with larger datasets, particularly for languages with similar phonological and morphological complexity as those in this study. Additionally, the effectiveness of adapter-based fine-tuning for MMS suggests that optimizing model architecture for scalable adaptation could yield further improvements.

Lang	Model	Tone	Nas.	V-Len	C-Len
Cicipu	MMS	0.199	0.337	0.239	0.174
	XLS-R	0.215	0.337	0.248	0.183
Mocho'	MMS	—	—	0.154	—
	XLS-R	—	—	0.160	—

Table 3: Phonological error rates (0–1 scale) for Cicipu and Mocho'. Cicipu shows higher confusion in tone, nasality, and consonant length, while Mocho' displays more issues with vowel length. Both models (MMS vs. XLS-R) yield broadly similar error patterns.

4.3 Error analysis

We performed a phonologically informed error analysis on two of the languages in our dataset, Cicipu and Mocho', both of which exhibit segmental contrasts that could be challenging for ASR models. Cicipu's orthography explicitly marks tone and nasality with diacritics and differentiates both vowel and consonant length with doubled letters (e.g., 'aa' for long vowels, 'tt' for geminate consonants), making it well suited for evaluating the system's performance on these phonological categories. Mocho' similarly features a vowel length distinction encoded with doubled vowel letters. Other languages in our dataset, such as Ulwa and Upper Napo Kichwa, contain too few instances of long vowels or voiceless sonorants for a robust category-based analysis.

4.3.1 Error rates

To quantify performance on these features, we leverage character-level alignments (details in Appendix C) and calculate phonological segment error rates (Table 3). For each category $C \in \{\text{Tone}, \text{Nasality}, \text{V\_length}, \text{C\_length}\}$, we sum all substitutions, deletions, and insertions that affect that category and normalize by the total number of reference tokens L_C exhibiting C.

As shown in Table 3, both MMS and XLS-R struggle with Cicipu's tone, nasality, and consonant length, each exhibiting error rates in the range of 30–38%. By contrast, vowel-length confusion is comparatively low (7–9%). For Mocho', the long–short vowel distinction remains problematic, with error rates around 35–38%. These findings suggest that neither model has a strong advantage for these particular phonological categories; even after fine-tuning, nuanced contrasts such as nasality and tone remain challenging.

Lang	Category	S%	D%	I%
Cicipu	Tone	47.62	32.11	20.27
	Nasality	0.00	70.00	30.00
	V-Len	55.58	27.86	16.56
	C-Len	44.18	33.90	21.92
Mocho'	V-Len	58.29	26.86	14.86

Table 4: Percentage of substitutions (S%), deletions (D%), and insertions (I%) in XLS-R 120 min output, for each phonological category in Cicipu and Mocho'.

4.3.2 Error distribution

For further investigation, we performed a more detailed error analysis on the output from the 120-minute model of XLS-R for Cicipu and Mocho'. Table 4 breaks down each category by the percentage of substitutions (*S*), deletions (*D*), and insertions (*I*).

In Cicipu, nearly half of the tone errors ($\sim 48\%$) are substitutions, indicating confusion over which tone mark to apply, while around a third are deletions and the remainder insertions. Nasality errors, by contrast, skew heavily toward deletions ($\sim 70\%$), suggesting the model often fails to detect nasal vowel features. Substitutions are rare for nasality, indicating that the system either omits it entirely or adds it spuriously rather than confusing it with another tone diacritic. Vowel-length errors ($\sim 44\%$ for deletions and insertions) and consonant-length errors ($\sim 55\%$ for deletions and insertions) reflect a high level of segment-level confusion, whereas confusion with a different vowel ($\sim 56\%$ substitutions) or consonant ($\sim 44\%$ substitutions) is also frequent. For Mocho', vowel-length confusion is likewise dominated by substitutions ($\sim 58\%$), a pattern similar to Cicipu which reveals that XLS-R often misidentifies one segment in the long vowel. The distributions point out that while the model does capture some acoustic correlates of nasality, tone, and length, it nevertheless struggles to map them consistently to diacritics and extended graphemes in low-resource scenarios.

Overall, these patterns highlight the persistent challenge of representing languages with complex orthographies and rich phonological inventories. Even after multilingual pre-training and fine-tuning, contrasts such as tone or nasality may be overlooked when the amount of transcribed data is minimal. Addressing these gaps may require linguistically informed data augmentation, specialized adapter modules, or loss functions that explicitly emphasize distinct phonological categories. In extremely low-resource settings, such targeted methods could provide the additional examples and acoustic cues needed for more accurate transcription of endangered languages.

5 Conclusion

Our experiments show that fine-tuned multilingual ASR models can substantially reduce the transcription burden for endangered and low-resource languages. Across five typologically diverse languages, MMS proved more effective with extremely limited labeled data, whereas XLS-R caught up once approximately one hour of transcribed material was available. By using Character Error Rate (CER) rather than Word Error Rate (WER), we focus on phoneme-level accuracy—a more direct measure for languages without standardized orthographies. Despite improvements in overall accuracy, both models struggled with challenging phonological categories in Cicipu, such as tone and consonant length, and exhibited a high rate of vowel-length confusion in Mocho'. These findings confirm that current multilingual ASR systems are indeed helpful for language documentation but still require targeted adaptations to handle nuanced phonological contrasts in under-resourced settings.

6 Future Work

Although our study confirms that fine-tuning multilingual ASR models can substantially reduce transcription overhead for low-resource languages, several research directions remain promising for further performance gains. One compelling approach is continued pre-training (CoPT) on unlabeled in-language audio. DeHaven and Billa (2022) show that CoPT on a wav2vec 2.0–based multilingual model can match or outperform pseudo-labeling techniques while being more computationally efficient. Similarly, Nowakowski et al. (2023) demonstrate that CoPT on about 234 hours of Sakhalin Ainu audio yields a considerable reduction in error, beyond what standard multilingual fine-tuning achieves. CoPT has also proven effective in domain adaptation, especially for noisy data or new speaker types (Attia et al., 2024). While the scarcity of fieldwork data could limit the scale of CoPT, even incremental benefits may substantially ease the manual transcription effort.

A second avenue is leveraging diverse augmentation methods to enlarge the effective training set. Self-training (pseudo-labeling) uses an initial

ASR model to generate transcripts for unlabeled audio, which can then be added to the training pool. This method has consistently boosted low-resource ASR performance (Bartelds et al., 2023), particularly when coupled with filtering or iterative refinement. TTS-based augmentation offers another option: if a target-language text-to-speech system is available, synthesizing speech from text yields additional "perfectly labeled" data, potentially improving recognition robustness (ibid). Finally, common audio perturbations, speed/pitch changes, SpecAugment, and noise injection, remain valuable for avoiding overfitting and preparing the model for real-world variability.

A final challenge involves better capturing difficult phonological categories, such as tone, nasality, and consonant length. Adapters in MMS could be extended or reconfigured to emphasize language-specific features, while training regimes could incorporate acoustic or phonological priors explicitly. Future work might integrate fine-grained linguistic annotations (if available) or employ specialized masking strategies during CoPT to boost the model's sensitivity to subtle contrasts. Combining these techniques into user-friendly toolkits will be essential for widespread adoption by field linguists, who often have limited computational resources yet require high-accuracy, phoneme-level transcriptions for documenting and revitalizing endangered languages.

7 Limitations

Despite promising results, several specific limitations affect the generalizability and applicability of our approach. The most critical limitation is related to the data size and representativeness of the linguistic diversity considered. Our study focused on a small number of typologically diverse languages, each with relatively limited datasets ranging from just a few minutes to two hours. As such, the models' performances may not generalize to other endangered or low-resource languages with distinct phonological or orthographic features.

Additionally, due to constraints inherent in linguistic fieldwork, the training and test datasets often contained data from the same speakers, potentially inflating model accuracy estimates. Future research should validate these findings with genuinely held-out speakers to better gauge model robustness to speaker variability.

Moreover, the orthographic inconsistencies and the absence of standardized orthographies in our datasets likely influenced model performance, especially for phonologically complex categories like tone, vowel length, and nasality. This issue highlights a broader limitation: ASR models trained under these conditions may struggle to generalize to spontaneous and noisy field recordings, especially when orthographic conventions vary within and across datasets.

Finally, computational resource limitations (training on a single NVIDIA T4 GPU with constrained runtimes) restrict our ability to fine-tune larger models or extensively optimize hyperparameters, which may have further improved performance. Addressing these limitations would require additional computational resources and potentially more extensive data augmentation strategies tailored explicitly to low-resource linguistic contexts.

Acknowledgments

We are grateful to the depositors and leaders of the Endangered Languages Archive (ELAR, http://elararchive.org) for sharing their invaluable resources which made this project possible.

References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly Automatic Transcription of Low-resource Languages: Plugging ESPnet into Elpis. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 51–62, Online. Association for Computational Linguistics.

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *arXiv preprint*. ArXiv:1803.08991 [cs].

Ahmed Adel Attia, Dorottya Demszky, Tolulope Ogunremi, Jing Liu, and Carol Espy-Wilson. 2024. Continued Pretraining for Domain Adaptation of Wav2vec2.0 in Automatic Speech Recognition for Elementary Math Classroom Settings. *arXiv preprint*. ArXiv:2405.13018 [cs] version: 1.

Peter K. Austin and Julia Sallabank. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press. Google-Books-ID: 0XZRauYgO6AC.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika

Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs].

Alexei Baevski, Wei-Ning Hsu, Alexis CONNEAU, and Michael Auli. 2021. Unsupervised Speech Recognition. In *Advances in Neural Information Processing Systems*, volume 34, pages 27826–27839. Curran Associates, Inc.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Junwen Bai, Bo Li, Qiujia Li, Tara N. Sainath, and Trevor Strohman. 2024. Efficient Adapter Finetuning for Tail Languages in Streaming Multilingual ASR. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10841–10845. ISSN: 2379-190X.

Russell Barlow. 2018a. Documentation of Ulwa, an endangered language of Papua New Guinea.

Russell Barlow. 2018b. A Grammar of Ulwa.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *arXiv preprint*. ArXiv:2305.10951 [cs].

Steven Bird. 2020. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Hennie Brugman and Albert Russel. 2004. Annotating Multi-media / Multi-modal resources with ELAN.

Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. 2022. End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation. *arXiv preprint*. ArXiv:2204.00540 [cs].

Rolando Coto-Solano. 2021. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Xiaodong Cui, Brian Kingsbury, Jia Cui, Bhuvana Ramabhadran, Andrew Rosenberg, Mohammad Sadegh Rasooli, Owen Rambow, Nizar Habash, and Vaibhava Goel. 2014. Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA babel program. In *Interspeech 2014*, pages 2103–2107. ISCA.

Mitchell DeHaven and Jayadev Billa. 2022. Improving Low-Resource Speech Recognition with Pretrained Speech Models: Continued Pretraining vs. Semi-Supervised Training. *arXiv preprint*. ArXiv:2207.00659 [cs].

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The Zero Resource Speech Challenge 2019: TTS without T. *arXiv preprint*. ArXiv:1904.11469 [cs].

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Napo Quichua. Edition: 27 Publisher: SIL International.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Tena Lowland Quichua. Edition: 26 Publisher: SIL International.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. ISSN: 1520-6149.

Karolina Grzech. 2020. Upper Napo Kichwa: a documentation of linguistic and cultural practices.

Frantisek Grézl, Martin Karafiát, and Karel Veselý. 2014. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7654–7658. ISSN: 2379-190X.

Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyen, Maxime Fily, Guillaume Jacques, and Alexis Michaud. 2022a. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. pages 4905–4909. International Speech Communication Association.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyen, and Maxime Fily. 2022b. 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.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. 36(1):161–196. Publisher: De Gruyter Mouton Section: Linguistics.

Nikolaus P Himmelmann and John U Wolff. 1999. *Toratán (Ratahan)*, volume 130. Lincom Europa.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR. ISSN: 2640-3498.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

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

Austin Jones, Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzic, and Keith Langston. 2024. Comparing Kaldi-Based Pipeline Elpis and Whisper for Čakavian Transcription. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 61–68, Bangkok, Thailand. Association for Computational Linguistics.

Anthony Jukes. 2010. Documentation of Toratán (Ratahan).

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Interspeech 2021*, pages 1529–1533. ISCA.

Jiyeon Kim, Mehul Kumar, Dhananjaya Gowda, Abhinav Garg, and Chanwoo Kim. 2021. Semi-supervised transfer learning for language expansion of end-to-end speech recognition models to low-resource languages. *arXiv preprint*. ArXiv:2111.10047 [eess].

Gina-Anne Levow, Emily P. Ahn, and Emily M. Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:96–106.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Stuart McGill. 2012. Cicipu documentation.

Stuart McGill. 2014. Cicipu. *Journal of the International Phonetic Association*, 44(3):303–318.

Yajie Miao, Florian Metze, and Shourabh Rawat. 2013. Deep maxout networks for low-resource speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 398–403.

Benjamin Milde and Arne Köhn. 2018. Open Source Automatic Speech Recognition for German. *arXiv preprint*. ArXiv:1807.10311 [cs].

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting Multilingual Speech Representation Model for a New, Underresourced Language through Multilingual Fine-tuning and Continued Pretraining. *Information Processing & Management*, 60(2):103148. ArXiv:2301.07295 [cs].

Erin O'Rourke and Tod D. Swanson. 2013. Tena Quichua. *Journal of the International Phonetic Association*, 43(1):107–120. Publisher: Cambridge University Press.

Naomi Elizabeth Palosaari. 2011. *Topics in Mocho' phonology and morphology*. Ph.D., The University of Utah, United States – Utah. ISBN: 9781124576213.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. ISSN: 2379-190X.

Leena G. Pillai, Kavya Manohar, Basil K. Raju, and Elizabeth Sherly. 2024. Multistage Fine-tuning Strategies for Automatic Speech Recognition in Low-resource Languages. *arXiv preprint*. ArXiv:2411.04573 [cs].

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Agnieszka Przezdziak. 2024. *Optimizing Speech Recognition for Low-Resource Languages: Northern Sotho*.

Jaime Pérez González. 2018. Documentation of Mocho' (Mayan): Language Preservation through Community Awareness and Engagement.

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. PMLR. ISSN: 2640-3498.

Sanjay Rijal, Shital Adhikari, Manish Dahal, Manish Awale, and Vaghawan Ojha. 2024. Whisper Finetuning on Nepali Language. *arXiv preprint*. ArXiv:2411.12587 [cs].

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint*. ArXiv:2101.10877 [eess].

Claytone Sikasote and Antonios Anastasopoulos. 2021. BembaSpeech: A Speech Recognition Corpus for the Bemba Language. *arXiv preprint*. ArXiv:2102.04889 [cs].

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully Convolutional ASR for Less-Resourced Endangered Languages. 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 126–130, Marseille, France. European Language Resources association.

Nick Thieberger. 2012. *The Oxford Handbook of Linguistic Fieldwork*. OUP Oxford. Google-Books-ID: 86AE2_0nPbkC.

Geoffroy Vanderreydt, François Remy, and Kris Demuynck. 2022. Transfer Learning from Multi-Lingual Speech Translation Benefits Low-Resource Speech Recognition. In *Interspeech 2022*, pages 3053–3057. ISCA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Michael Wroblewski. 2012. Amazonian Kichwa Proper: Ethnolinguistic Domain in Pan-Indian Ecuador. *Journal of Linguistic Anthropology*, 22(1):64–86. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1548-1395.2012.01134.x.

G. Thimmaraja Yadava and H S Jayanna. 2017. Development and comparison of ASR models using kaldi for noisy and enhanced kannada speech data. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1832–1838. Conference Name: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) ISBN: 9781509063673 Place: Udupi Publisher: IEEE.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. *arXiv preprint*. ArXiv:2110.03370 [cs].

Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur. 2014. Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 215–219. ISSN: 2379-190X.

A Training Details

The models were trained on an NVIDIA T4 GPU, with training times ranging from approximately 1 to 60 minutes per model. The hyper-parameters were defined as follows:

- **Learning rate:** MMS: 1e-3; XLS-R: 3e-4
- **Maximum epochs:** 30
- **Best model metric:** Character Error Rate (CER)
- **Early stopping:** 3 epochs
- **Early stopping threshold:** 0.003

B Dataset Details

Table 5 provides detailed information on the genres and types of content present in the datasets used in this study, along with key linguistic references and citations to the original documentation archives. Table 6 summarizes the total archived hours of audio recordings for each language and the amount of data remaining after the cleaning and preprocessing steps described earlier in the manuscript.

Language	Genres of content	Phonological Description	Documentation
Cicipu	Greetings, conversations, hortative discourse, narratives, procedural discourse, ritual discourse, elicitation activities	McGill (2014)	McGill (2012)
Mocho'	Biographical and non-biographical narratives (historical events, myths, local beliefs, traditional building, witchcraft), prayer, conversation, elicitation sessions, text translation	Palosaari (2011)	Pérez González (2018)
Toratán	Conversational data, elicitation sessions, narratives (personal history, folk tales)	Himmelmann and Wolff (1999)	Jukes (2010)
Ulwa	Conversational data, traditional stories, personal stories, traditional singing and dancing video	Barlow (2018b)	Barlow (2018a)
Upper Napo Kichwa	Grammatical elicitation, life interviews	Wroblewski (2012), O'Rourke and Swanson (2013)	Grzech (2020)

Table 5: Details of depository content for languages used in this paper, related linguistic work referenced, and original documentation citations.

Language	Total Hours	Cleaned Hours
Cicipu	5.66	3.09
Mocho'	7.26	4.21
Toratán	22.84	11.15
Ulwa	3.25	2.83
Upper Napo Kichwa	13.19	6.97

Table 6: Total archived and cleaned hours of audio for all languages used in the study.

C Error Rates

We consider four phonological categories:

$$C \in \{\text{Tone, Nasality, V\_length, C\_length}\}.$$

Over the entire dataset, we record:

- S_C: the total *substitution* errors for category C,
- D_C: the total *deletion* errors for category C,
- I_C: the total *insertion* errors for category C,
- L_C: the total *reference tokens* exhibiting category C (e.g. `tone_labels` for tone, `total_vowels` for vowel length, etc.).

We then define the total errors and error rate for each category C as follows:

$$E_C = S_C + D_C + I_C \quad \text{and} \quad \text{ErrorRate}_C = \frac{E_C}{L_C}.$$

For example, if C = tone then L_C = tone_labels (the number of reference tokens with at least one tone diacritic). Similarly, if C = vowel_length then L_C = total_vowels (the total vowel tokens in the reference).

KazBench-KK: A Cultural-Knowledge Benchmark for Kazakh

Sanzhar Umbet[♡] **Sanzhar Murzakhmetov**[♡†] **Beksultan Sagyndyk**[♡†]
Kirill Yakunin[♡] **Timur Akishev**[♠] **Pavel Zubitskii**[♡]
[♡] Horde Research [†] Yieldmo [♠] KIMEP University
{sanzhar_u,sanzhar}@horderesearch.com

Abstract

We introduce **KazBench-KK**, a comprehensive 7,111-question multiple-choice benchmark designed to assess large language models' understanding of culturally grounded Kazakh knowledge. By combining expert-curated topics with LLM-assisted web mining, we create a diverse dataset spanning 17 culturally salient domains, including pastoral traditions, social hierarchies, and contemporary politics. Beyond evaluation, KazBench-KK serves as a practical tool for field linguists, enabling rapid lexical elicitation, glossing, and topic prioritization. Our benchmarking of various open-source LLMs reveals that reinforcement-tuned models outperform others, but smaller, domain-focused fine-tunes can rival larger models in specific cultural contexts.

1 Introduction

Kazakh reflects a web of pastoral traditions, kinship rules, and post-Soviet social change content that is almost invisible in the English-dominated web. Kazakh is a language that is primarily used and spoken in Kazakhstan and some neighboring regions, but mainstream language models rarely handle it well.

In the NLP landscape, Kazakh is considered a low-resource language due to the scarcity of openly available datasets. This consequently leads to poor performance of LLMs comprehending Kazakh speech and texts, and significantly makes them lack the culturally-specific knowledge of Kazakh traditions, customs and cultural context that are essential for creating inclusive and locally relevant AI systems. While recent efforts have produced datasets for tasks like named entity recognition, sentiment analysis and translation, these are often limited in scope and do not reflect the deep cultural grounding necessary to evaluate how well language models truly understand Kazakh society.

In this paper, we present a semi-automated pipeline designed to generate a benchmark focused on culturally significant knowledge in the Kazakh language. Our approach combines manual topic curation with LLM-assisted keyword generation, automated web retrieval and preprocessing, and context-driven QA generation, followed by both automatic filtering and human validation.

Beyond evaluation, our benchmark opens up practical use cases for linguists working with underrepresented languages. A culturally aware LLM can offer significant advantages to field linguists by connecting language and culture in efficient and innovative ways. Field linguists, who have traditionally relied on the manual collection of linguistic data, can now use LLMs to obtain quick summaries of culture-specific linguistic phenomena and determine which topics are worth further investigation.

Furthermore, both traditional data preparation tasks, including glossing, elicitation prompt construction, and other background research in general and situational decision-making procedures during fieldwork can benefit from these improvements. It is also possible to compare manually collected field data with AI-generated data.

A culturally aware LLM offers field linguists an efficient bridge between language and culture. Instead of relying solely on labor-intensive manual collection, they can query KazBench-KK-tuned models for rapid overviews of culture-specific phenomena, pinpoint promising domains for deeper elicitation, and automatically generate glosses or prompts. Moreover, the benchmark's hierarchical taxonomy reveals how Kazakh speakers organise concepts, turning traditional fieldwork into a more quantified and streamlined endeavour. The accompanying league table allows practitioners to quickly see which publicly available models consistently demonstrate culturally accurate and context-aware responses.

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 38–57
August 1, 2025 ©2025 Association for Computational Linguistics

Our contributions are as follows:

- **Introduction of Cultural Benchmark:** We introduce KazBench-KK, a 7111-question multiple-choice benchmark specifically designed to evaluate large language models' understanding of culturally grounded knowledge in the Kazakh language. This benchmark fills a critical gap in resources for evaluating how well AI models understand the nuances of Kazakh culture.

- **Culturally Salient Domain Coverage:** The benchmark covers 17 culturally significant domains, including pastoral traditions, social hierarchies, and contemporary politics. These domains were carefully selected, combining expert-curated topics with LLM-assisted web mining, ensuring a comprehensive and relevant assessment of cultural understanding.

- **Semi-Automated Pipeline for Data Generation:** We present a novel, semi-automated pipeline for the efficient generation of high-quality, culturally relevant data. This pipeline combines the strengths of both human expertise and machine automation, addressing the challenges of data scarcity for low-resource languages.

- **Benchmarking of Open-Source LLMs:** The paper includes a thorough benchmarking of several open-source large language models. This provides a valuable resource for linguists and practitioners seeking to choose the most appropriate models for tasks that involve the Kazakh language and its cultural context.

2 Related Work

Prior work on evaluating cultural knowledge falls into three strands: general English benchmarks, multilingual suits, and recent Kazakh-specific sets. They effortlessly handle multiple languages, generate text with human-like fluency, and are useful in many contexts. However, despite their global reach, these models remain heavily "westernized", and predominantly understand and reflect Western cultural norms and traditions (Naous et al., 2024; Wang et al., 2024; Cao et al., 2023). This western-centric bias inevitably creates a gap when it comes to accurately interpreting and engaging with non-Western, particularly Central Asian, cultures.

Multiple studies have analyzed the performance of language models to generate culturally relevant responses in diverse cultural settings. However, most of these evaluations are centered around high-resource languages, or rely mainly on translation-based approaches that fail to capture deep cultural context. To situate our work, we first review existing English language benchmarks, then discuss recent efforts to extend such benchmarks to multilingual or indigenous settings. Finally, we highlight the current limitations of Kazakh language resources and demonstrate how our work addresses this critical gap.

2.1 General-purpose English Benchmarks

Currently, there are multiple benchmarks in English that try to assess models' different aspects of knowledge. For example, the general language understanding evaluation (GLUE; Wang et al., 2018) and SuperGlue (Wang et al., 2019) benchmarks are aimed to evaluate language models on multiple tasks, including: sentiment analysis, lexical entailment, coordination scope and many more. Moreover, HellaSwag (Zellers et al., 2019) and CosmoQA (Huang et al., 2019) benchmarks are also commonly used to evaluate commonsense reasoning. Nevertheless, as the development of language models progress, it became more common for them to perform on these benchmarks on the human-like level. Therefore, to make better assessments of more advanced language models new challenging benchmarks were developed. They include: MMLU (Hendrycks et al., 2021b,a), AGIEval (Zhong et al., 2023) and BIG-bench (Srivastava et al., 2022), each introducing more complex questions on different topics.

2.2 Multilingual & Cross-cultural Benchmarks

The evaluation of LLMs across different languages has led to the creation of several multilingual benchmarks. Notable examples include XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), and MEGA (Ahuja et al., 2023), which are designed to test language models' performance on a range of tasks in multiple languages, from high-resource to low-resource ones. Additionally, efforts have been made to build datasets tailored to specific language families (Huang et al., 2023; Doddapaneni et al., 2023; Adebara et al., 2023).

These benchmarks mainly assess syntactic and semantic capabilities such as translation, question answering, and classification.

Beyond general linguistic evaluation, more recent research has focused on cultural benchmarks that aim to measure LLMs' understanding of sociocultural knowledge. These include datasets like GeoLAMA (Yin et al., 2022), which evaluates geo-diverse commonsense reasoning, and CulturalAtlas (Fung et al., 2024), which compiles social norms from over 193 countries. Other works, such as CREHate (Lee et al., 2024) and StereoKG (Deshpande et al., 2022), examine cultural stereotypes and bias across regions using social media and crowd-sourced data. However, none of these suits addresses the cultural fabric of Kazakh life.

2.3 Kazakh-specific Benchmarks

Despite recent advancements in multilingual NLP, Kazakh remains significantly underrepresented in benchmark development. While foundational datasets have been introduced for core NLP tasks, such as KazNERD for named entity recognition (Yeshpanov et al., 2022), KazSAnDRA for sentiment analysis (Yeshpanov and Varol, 2024), and KazParC for machine translation (Yeshpanov et al., 2024) - most of these are narrow in scope and task-specific. They offer valuable building blocks, but do not capture the broader reasoning capabilities or cultural depth needed to evaluate how well LLMs understand Kazakh society.

To help address this, a few benchmark-style datasets have recently emerged. One example is the Kazakh Unified National Testing MC dataset, which contains nearly 15,000 multiple-choice questions pulled from Kazakhstan's national standardized exams (Sagyndyk et al., 2024b). These questions span subjects such as Kazakh literature, history, geography, and biology, providing a realistic and academically grounded way to test the grasp of a model of school-level Kazakh knowledge.

Another effort is the Kazakh Constitution MC dataset, which includes more than 400 multiple-choice questions based on Kazakhstan's constitution (Sagyndyk et al., 2024a). This benchmark is more civic in nature, offering a way to evaluate how well a model understands the legal and governmental concepts that are specific to Kazakhstan.

There is also a Kazakh-translated version of the popular MMLU benchmark, containing around 15,900 multiple-choice questions across a wide range of topics (Sagyndyk et al., 2024c). While helpful for assessing general reasoning in a low-resource setting, this benchmark is entirely translation-based and may not fully preserve Kazakh-specific cultural or contextual nuances.

From a field-linguist perspective, an LLM that handles such culturally grounded content could accelerate tasks like domain word-list expansion or contextual translation checks. However, no public benchmarks let practitioners compare models on these abilities.

All of these benchmarks represent important steps forward. But they still focus mostly on academic or formal domains, and none are designed to test a model's ability to reason about everyday Kazakh customs, values or culturally embedded practices. In other words, we still do not know how well LLMs can engage with the lived experience of Kazakh speakers.

3 Methods

The creation of culturally aware NLP models requires considerable effort, particularly for low-resource languages, where even regular data is limited. Data acquisition methods generally fall into three categories, manual, automatic, and semi-automatic (Liu et al., 2025). Manual data acquisition involves hiring native speakers or professional translators to annotate or culturally adapt textual resources. Additionally, crowdsourcing platforms, university mailing lists, and Slack or Discord channels of relevant organizations regularly serve as sources for gathering culturally rich textual data through user interaction, conversations, and public messaging (Liu et al., 2021).

Another promising method for data collection leverages LLMs to extract cultural knowledge. For instance, Nguyen et al. (2023) proposes a workflow that identifies culturally significant information in texts by using named entity recognition, culturally trained classification models, and information retrieval and ranking algorithms to create culturally aware datasets. However, as highlighted by Putri et al. (2024), fully automating dataset creation using LLMs remains challenging, as the generated texts typically lack deep cultural understanding and may exhibit fluency errors. A potential solution to balance automation and quality is to adapt a semi-automatic approach, merging manual annotations with automated processes. Studies by

Liu et al. (2024) and Bhutani et al. (2024) demonstrated the effectiveness of using prompting techniques for initial data generation, followed by human evaluation to verify and refine cultural relevance.

To address the scarcity of culturally grounded Kazakh benchmarks, we developed a semi-automated data generation pipeline that uses LLMs and web-scale retrieval to synthesize high-quality data. The core goal of the system is to generate multiple choice questions centered on culturally and contextually significant topics in Kazakhstan, which are currently absent from existing benchmarks.

3.1 Linguistic & cultural categories

Our selection of categories and concepts was guided by the goal of capturing Kazakh culture in various forms of its representation. We primarily focused on those aspects of culture that can be expressed, preserved, or transmitted through language and text, whether spoken or written. The inherently textual categories that we added to the dataset are related to (1) creativity (literature, song lyrics, and films) and (2) formulaic language (proverbs, sayings, prayers, and spiritual expressions). Other categories selected for the dataset were not inherently textual in nature, but have been recorded and can be described using text: (3) traditions and customs, as they form the core of any culture, (4) social relations and hierarchies, as they reflect the organization of the society, (5) daily life (names of traditional foods and clothing and terminology used to refer to traditional household objects, architecture, and agriculture), and (6) arts and crafts (tools, materials, and techniques).

Criteria	**Description**
Traditions	Family events; holidays, rituals and ceremonies
History	Crucial historical events; historical figures
Social relationships	Family members; relatives; polite terms for strangers; endearments for loved ones
Politics and social strata	Historical terms (e.g., khans, bis); zhuzes and rus
Proverbs, spirituality	Sayings, spiritual terms (e.g., *bata*); superstitions, mythology
Humor	Jokes, *aitys*, humorous figures (e.g., Aldar Kose); wordplay
Cuisine	Recipes; names for food and beverages
Sports and games	Names and rules of traditional games and sports
Films	Classic and contemporary Kazakh cinema; landmark films, directors, actors, and culturally significant storylines
Literature	Poetry and fiction with cultural relevance
Song lyrics	Traditional songs, *kuys*
Instruments	Names of instruments and parts
Arts and crafts	Crafts, decorative and performing arts
Clothing	Names of traditional garments
Named entities	Names of people/places and their meanings (onomastics)
Agriculture	Terms related to farming and herding
Architecture	Yurt structure and home elements

Table 1: Cultural Knowledge Categories

3.2 Semi-Supervised Benchmark-generation pipeline

Our data generation pipeline consists of several key stages

Topic initialization. Initially, we manually curated a comprehensive list of general topics, organizing them into clearly defined knowledge categories relevant to Kazakh society, such as: Media, Politics, Traditions, and so on. Within each general category, we further identified distinct subcategories to cover diverse perspectives and deepen contextual relevance. For instance, under 'Current social life', we explored subcategories like the scandalous 'Bishimbayev case', ecological issues in Almaty or negligence in the Thermal Plant in Ekibastuz.

LLM-based keyword generation. For each category–subcategory pair, our linguists and sociologists first compiled a concise seed list of culturally salient terms. We then used *GPT-4o* to expand these expert-provided seeds, instructing the models to propose roughly ten additional, culturally anchored keywords (i.e., sub-subcategories) that captured dialectal variation, idiomatic usage, and other nuanced linguistic forms. This human-in-the-loop procedure ensured that domain knowledge grounded the process while the LLM broadened the lexical scope. The resulting keyword sets were subsequently transformed into natural-language search queries, reflecting how a native speaker might phrase them in a typical Google search.

Algorithm 1: KazBench-KK data-generation pipeline

Input: Manually curated category list C with seed keywords
Output: Multiple-choice question set $\mathcal{Q}$

foreach $(c, \mathit{sub}) \in C$ **do**
  `/* Step 1: keyword expansion */`
  $\mathit{Seeds} \leftarrow$ linguist/sociologist seed list ;
  $\mathit{Expanded} \leftarrow$ LLM_Expand($\mathit{Seeds}, n{=}10$) ;
  $\mathit{Queries} \leftarrow$ MakeQueries($\mathit{Seeds} \cup \mathit{Expanded}$) ;
  `/* Step 2: content retrieval */`
  $\mathit{Docs} \leftarrow$ WebSearch($\mathit{Queries}$) ;
  `/* Step 3: preprocessing */`
  $\mathit{Clean} \leftarrow$ ParseAndClean(Docs) ;
  $\mathit{Corpus} \leftarrow$ Deduplicate(Clean) ;
  `/* Step 4: MCQ generation */`
  foreach $d \in \mathit{Corpus}$ **do**
    $\mathit{mcq} \leftarrow$ LLM_MCQ(d) ;
    if IsCultureSpecific*(mcq)* **then**
      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\mathit{mcq}\}$;
return $\mathcal{Q}$

Content retrieval. With the search queries generated, we then performed automated web retrieval. We integrated external API services to execute extensive searches on websites and platforms such as Wikipedia, local Kazakh news outlets, and blog posts.

Webparsing and Preprocessing. The retrieved website URLs underwent an automated custom parsing and clearing process. We utilized the open-sourced HTML parsing scripts to scrape textual data from the websites, and implemented preprocessing techniques to remove HTML tags, navigation elements, and redundant information. Additionally, we employed a deduplication approach to ensure data quality and consistency.

LLM-based question generation. After preprocessing, the cleaned text corpus was fed into a large language model to generate structured multiple-choice questions (MCQ). For each content chunk, the LLM was prompted to produce context-based MCQs along with four answer options, with three being distractors and one correct answer, grounded in the specific cultural or historical context. We adopted a four-option format to align with common standardized practices in Kazakhstan and global MCQ benchmarks, ensuring compatibility with existing evaluation tools. To support better dataset usability and analysis, each question was also tagged with a binary annotation indicating whether it required context-specific knowledge, and whether a generated question was Kazakh-culture-specific. This allowed us to later filter and categorize the dataset based on its cultural relevance and reasoning complexity.

3.3 Data Filtering

We developed a set of criteria to ensure the high quality of our data. These criteria applied to both the questions and the answer options, focusing on their overall structure, logic, coherence, grammatical correctness, and the relevance of the options to the questions. We aimed to avoid absurd or overly obvious items and ensure that the answer options, including distractors, were appropriate and justifiable. Additionally, we wanted our data to be balanced in terms of general quality, difficulty, and diversity. Finally, we evaluated the overall relevance of the question–answer pairs to the categories and subcategories constituting the notion of culture. Applying these criteria helped us refine the dataset and eliminate any major illogical, incoherent, absurd, or otherwise irrelevant items.

3.3.1 Automated pre-filtering

To reduce annotator load, we translated the above rules into a binary "keep vs. discard" classifier implemented as a `gemini-2.0-flash-lite` agent in LangChain. The model embeds each MCQ with its answer set, applies chain-of-thought self-critique, and filters out items whose risk score exceeds 0.5 prior to human review. Table 2 presents the classifier's performance on a held-out set of 97 examples; the macro F_1-score is 0.87.

Class	Precision	Recall	F_1	Support
Discard (noise)	0.84	0.88	0.86	42
Retain (good)	0.91	0.87	0.89	55
Accuracy			0.88	97
Macro avg	0.87	0.88	0.87	97
Weighted avg	0.88	0.88	0.88	97

Table 2: Metrics for the binary filter

3.3.2 Human curation

To complement the automatic filter, we collaborated with four native-speaker linguists who manually reviewed and refined the remaining items. Following the same rubric used by the automated filtering agent, the annotators could also correct the wording, swap distractors, or flag entire MCQs for removal; no overlapping assignments or majority voting was required.

Annotator profile. All four annotators are Kazakh women of Asian ethnicity. Three are aged 18–24, and one falls within the 35–44 age range. Two hold undergraduate degrees in Language studies, while the other two have completed master's programs. As a qualification check, each annotator answered ten control questions from Kazakhstan's national standardized exams (Sagyndyk et al., 2024b) for Kazakh language and all scored a perfect 10/10.

ID	Gender	Age	Education	Ethnicity / Nationality
A1	Woman	18–24	B.A. Linguistics	Asian / Kazakh
A2	Woman	18–24	B.A. Linguistics	Asian / Kazakh
A3	Woman	18–24	M.A.	Asian / Kazakh
A4	Woman	35–44	M.A.	Asian / Kazakh

Table 3: Demographic profile of human annotators.

4 Dataset Description

4.1 Overview and format

Statistic	A	B	C	D	question
Tokens (total)	22 425	25 056	24 045	22 193	63 059
Tokens (avg.)	3.154	3.524	3.381	3.121	8.868
Unique tokens	8 997	10 565	10 048	9 297	15 282
Sentences (avg.)	1.009	1.011	1.010	1.009	1.013
Kk-char ratio	0.0907	0.0906	0.0895	0.0874	0.1020

Table 4: Descriptive statistics for answer options (A–D) and question stems (Q).

KazBench–KK consists of **7,111** multiple-choice questions (MCQs).[1] Each JSON record contains a single-sentence stem in Cyrillic Kazakh, four answer options (`A-D`), a field indicating the correct answer, and three metadata fields (`category`, `subcategory`, `keyword`).

4.2 Quantitative characteristics

Category distribution. Figure 2 shows that cultural topics are highly uneven on the web and the dataset mirrors this reality: *History* is the largest class with *1 103* items, followed by *Onomatopoeia* (621) and *Agriculture* (579). The smallest bar belongs to *Swearing* category with slightly over 50 questions. Despite the long tail, every category contains dozens of samples, enabling per-domain evaluation.

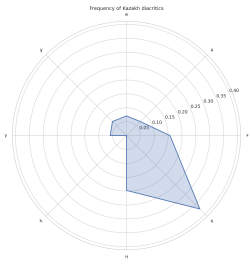

Figure 1: Diacritics distribution

Sub-category coverage. The finer-grained view (Fig. 3) contains 70-plus sub-categories. Counts range from roughly 450 questions at the top to around 30 at the bottom, implying that no single niche dominates the benchmark.

Question length. Box plots in Fig. 4 reveal a tight span: the median stem length is **7 tokens** across all domains, with the middle 50 % of examples falling between 6 and 9 tokens. Only a handful of outliers exceed 14 tokens.

Lexical diversity. Token-type ratios by column are plotted in Fig. 6. Stems have the lowest variety, reflecting repeated use of interrogatives (*қандай, қай*). Answer options are markedly richer (TTR $\approx$ 0.60–0.70), and some specialised domains (e.g. *Swearing expressions*) push the ratio beyond 0.95.

Orthographic coverage. Eight Kazakh-specific Cyrillic letters (*ә, қ, ң, ғ, ө, ү, ұ, і*) appear in the corpus. The radar chart (Fig. 1) shows that "*қ*" alone accounts for about 40 % of the diacritic tokens, with "*ң*" and "*ғ*" the next most common. Consequently, automatic evaluation cannot succeed by handling only Russian spellings.

Answer-key balance. The answer keys were originally placed so that each position (A-D) had the correct option exactly one quarter of the time, eliminating positional bias at generation time. After human curation, where annotators occasionally rewrote, swapped, or pruned options, the distribution drifted, and Fig. 7 now shows a modest skew across positions. We report this shift to inform reviewers about the residual position bias introduced during manual cleanup.

[1] Available at HF.

4.3 Linguistic profile

Frequent vocabulary. The histogram in Fig. 5 confirms that stems are dominated by function words and wh-terms, whereas answers introduce content words such as *қазақ* 'Kazakh', *дәстүрлі* 'traditional' and named entities. This design forces models to rely on content-specific cues rather than stereotyped question templates.

Category–specific variation. The heat-map of type–token ratios highlights clear lexical contrasts: creative domains such as *Cinema* display the highest diversity within columns, while everyday areas (*Agriculture*, *Traditions*) use a narrower but still non-trivial vocabulary. Such variation allows error analysis that links model failures to specific cultural sublexica.

Summary. Taken together, the figures demonstrate that KazBench–KK offers (i) broad topical coverage, (ii) compact but information rich stems, (iii) balanced answer positions, and (iv) authentic Kazakh orthography. These properties make the dataset a realistic stress test for language models that claim cultural knowledge of Kazakhstan.

5 Results

We selected a diverse panel of 21 checkpoints that (i) span the major open-source families (Llama-3, Gemma-3, Qwen 2.5, Mistral, Nemotron, DeepSeek) and (ii) cover the full spectrum of tuning regimes (base SFT, community SFT-tune, and RL/Instruct). We excluded any model that participated in our data-generation pipeline—those very large, API-only LLMs that seeded the MCQs-because evaluating them on a benchmark they helped create would inflate scores and mask true generalisation. This "no-leak" policy avoids circularity and lets us gauge how well *independent* models, with parameter counts from 8B to 70B, handle culture-specific content. Within that cohort, reinforcement-/instruction-tuned models dominate

On logit-level multiple-choice scoring, reinforcement-/instruction-tuned models dominate: Gemma-3-27B-*it* (0.72), both Llama-3-70B Instruct variants (0̃.71), and Nvidia's Nemotron-Super-49B RL model (0.69) form a clear first tier. Model scale still matters - Nemotron-Nano-8B RL plunges to 0.35 - but domain-focused fine-tunes can partly offset size: the 8B Sherkala chat model (0.69) and KazLLM-70B (0.69) rival much larger base checkpoints. Pure SFT baselines such as Gemma-3-12B-pt (0.62) and Qwen-32B (0.62) trail their RL counterparts by 6–10 points, confirming the benefit of preference optimization even when no text generation is required. Overall, reinforcement alignment combined with sufficient parameters remains the most reliable recipe for KazBench-KK accuracy, though well-targeted community SFTs can yield competitive gains.

At the category level, *Cinema* and *Onomatopoeia* are consistently the hardest sections, dipping below 0.60 for nearly every model, including top-tier Gemma-3-27B-*it* (0.69 and 0.67, respectively) and falling into the mid-0.40s for smaller checkpoints. Conversely, politically grounded knowledge is easy: all first-tier models top 0.79 on *Politics & Social Stratification*, with Gemma-3-27B-*it* at 0.79 and Llama-3-70B Instruct at 0.81. Nvidia's Nemotron-Super-49B shows a distinctive strength in *Musical Instruments* (0.69) and *Architecture* (0.72), whereas the Sherkala 8B chat model punches above its weight in *Humor* (0.71) and *Cuisine* (0.67)-categories where many SFT baselines lag. KazLLM-70B peaks at *Swearing & Offensive Expressions* (0.70), reflecting its culture-specific tuning. The overall spread suggests that cultural trivia tied to media, sound symbolism, and pop-culture films remains challenging, while hierarchical or historically codified knowledge (political titles, social classes, formal rituals) is much easier for models to retrieve.

Model name	Type	Accuracy
google/gemma-3-27b-it	rl	0.7216
meta-llama/Llama-3.3-70B-Instruct	rl	0.7090
meta-llama/Llama-3.1-70B-Instruct	rl	0.7030
nvidia/Llama-3_3-Nemotron-Super-49B-v1	rl	0.6936
inceptionai/Llama-3.1-Sherkala-8B-Chat	sft-tune	0.6909
issai/LLama-3.1-KazLLM-1.0-70B	sft-tune	0.6892
google/gemma-3-12b-it	rl	0.6794
mistralai/Mistral-Small-24B-Instruct-2501	rl	0.6761
Qwen/Qwen2.5-32B-Instruct	rl	0.6334
google/gemma-3-12b-pt	sft	0.6241
Qwen/QwQ-32B	sft	0.6165
deepseek-ai/DeepSeek-R1-Distill-Llama-70B	sft	0.6019
Qwen/Qwen2.5-14B-Instruct	rl	0.6002
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	sft	0.5996
google/gemma-3-4b-pt	sft	0.5854
google/gemma-3-4b-it	rl	0.5828
meta-llama/Llama-3.1-8B-Instruct	rl	0.5750
issai/LLama-3.1-KazLLM-1.0-8B	sft-tune	0.5656
nvidia/Llama-3.1-Nemotron-Nano-8B-v1	rl	0.3542
TilQazyna/llama-kaz-instruct-8B-1	rl	0.2768

Table 5: Overall accuracy of evaluated models. Model types: **rl** = reinforcement-tuned, **sft** = base supervised fine-tune, **sft-tune** = post supervised fine-tune.

Why an Offline-Only Evaluation All checkpoints were executed locally-without any hosted-API calls-for four technical reasons.

(1) Apples-to-apples comparability: restricting the pool to models that ship raw weights prevents API-only systems from benefiting from undisclosed tool use or server-side retrieval, so every score reflects the base language model alone.

(2) Decoding transparency: local inference lets us pin the exact tokenizer build, sampling algorithm, and context window; commercial endpoints may apply proprietary post-processing that we cannot inspect or replicate.

(3) Logit access for analysis: computing per-option log-likelihoods, error heat-maps, or calibration curves requires raw logits-information that most APIs do not expose.

These constraints keep the leaderboard a clean test of model weights, tokenization, and decoding policy-nothing else.

6 Conclusions

This study introduces *KazBench-KK*, a 7,111-item benchmark that assesses how well contemporary language models grasp cultural knowledge encoded in Kazakh. Built through a semi-automatic pipeline that blends expert guidance, web mining, and careful human curation, the dataset covers seventeen domains ranging from clan hierarchy to popular cinema.

The evaluation paints a mixed picture. Large, reinforcement-aligned models, like Gemma-3-27B-it and the Llama-3-70B Instruct pair-handle codified facts such as historical events with confidence, but their accuracy drops on items tied to film references or sound-symbolic words. Smaller community fine-tunes, notably Sherkala-8B and KazLLM-70B, narrow the gap in conversational categories like humour, swearing, and cuisine, showing that targeted data can offset limited parameter count in specific niches.

Practically, the league table offers a guide: Choose a heavyweight model when the task demands institutional knowledge, and reach for a lean, locally tuned model when nuance in everyday language matters more. For researchers, the consistent underperformance on Cinema and Onomatopoeia highlights clear gaps where additional data collection is likely to yield rapid gains.

Finally, the methodology itself is portable. Because each stage of the pipeline-seed selection, keyword expansion, retrieval, and filtering-relies on general tools, other language communities can replicate the process to create their own culturally specific benchmarks.

7 Future Work

Future research could expand KazBench-KK by integrating open-ended questions and dialect-specific knowledge from underrepresented rural regions. Moreover, the semi-automated benchmarking pipeline introduced here can be extended beyond textual data, facilitating culturally grounded benchmarks in multimodal domains such as images, audio, and video. Applying this methodology across diverse modalities would support a more comprehensive understanding and representation of Kazakh culture and other low-resource cultural contexts.

8 Limitations

Our benchmark cannot claim exhaustive coverage of Kazakh culture. Web-derived material is skewed toward urban, Russian-influenced outlets, so the lexicon of rural dialects and oral genres (e.g., regional *aitys*) remains underrepresented. Although the generation pipeline balanced answer keys at creation time, manual curation introduced a mild positional skew (Fig. 7). The questions are single-sentence MCQs; they do not test open-ended generation, discourse planning, or code-switching.

9 Ethics

Data provenance. All text was scraped from publicly accessible websites; we removed pages that contained personal names, contact details, or paywalled material. The released dataset stores only short question stems and answer options, minimising potential copyright concerns.

Annotator welfare. Four native-speaker linguists contributed on a *voluntary* basis; they received no monetary compensation, but gave their informed consent, could skip any item, and were free to withdraw at any time.

Bias and cultural sensitivity. Web sources may reflect gender, regional, or political biases; the benchmark therefore inherits those biases. Some items reference sensitive topics (e.g. clan affiliation, swearing); we flagged such questions with metadata so that downstream users can filter them if desired.

Acknowledgments

We thank Arman Zharmagambetov for his valuable feedback and insightful discussions that significantly contributed to the development of this work.

We also thank the reviewers for their thoughtful comments and suggestions, which helped improve the quality and clarity of the paper.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition lm benchmarking.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024a. Kazakh constitution: Multiple choice benchmark. Available on Hugging Face.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024b. Kazakh unified national testing: Multiple choice benchmark. Available on Hugging Face.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024c. Mmlu on kazakh language: Translated mmlu benchmark. Available on Hugging Face.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds,

Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. Kaznerd: Kazakh named entity recognition dataset.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. Kazparc: Kazakh parallel corpus for machine translation.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. Kazsandra: Kazakh sentiment analysis dataset of reviews and attitudes.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models.

Appendix

A Question Category Distribution

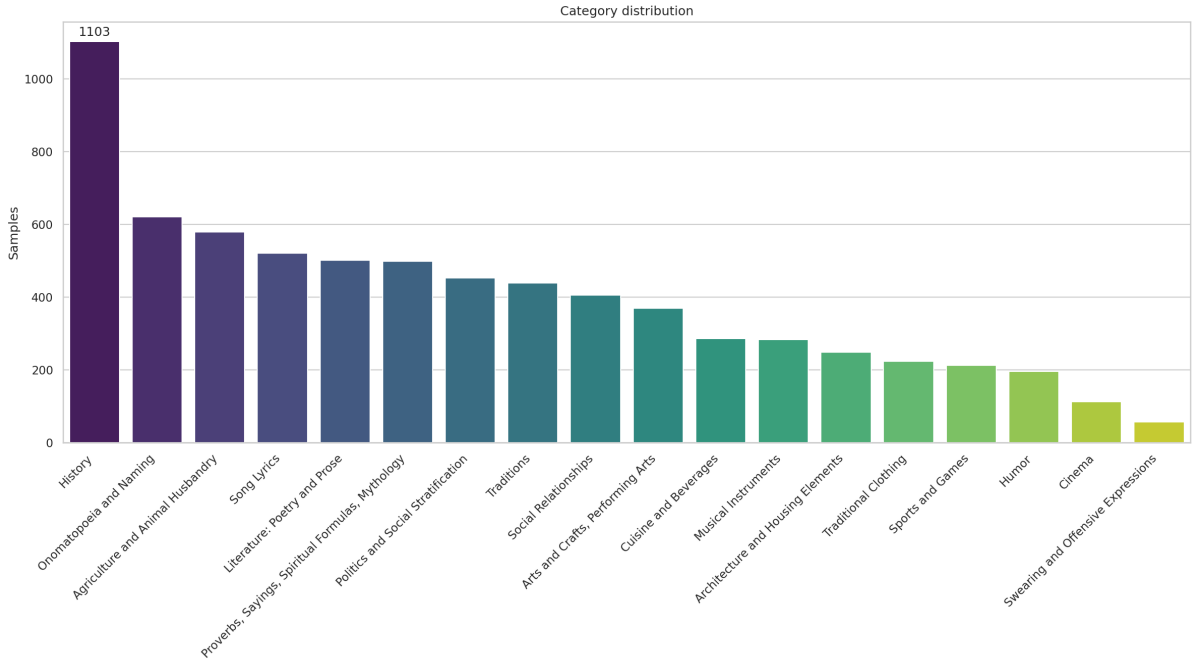

Figure 2: Distribution of questions across major cultural categories in KazBench-KK.

B Sub-Category Distribution

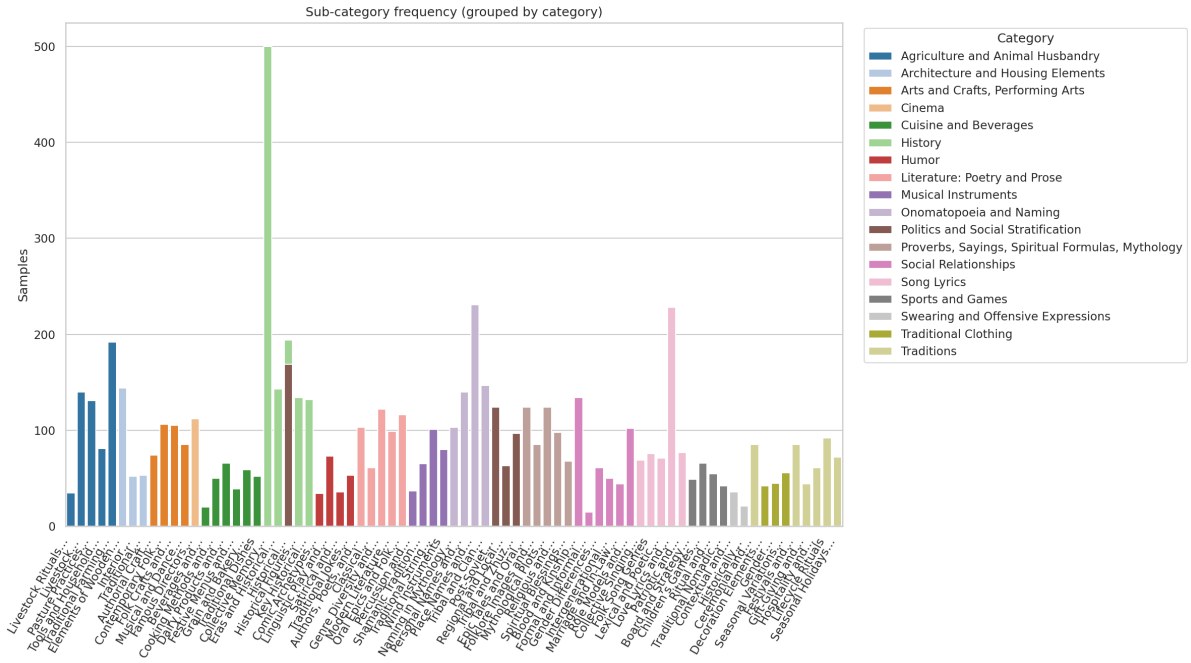

Figure 3: Granular breakdown of question counts per sub-category, demonstrating the breadth of domain-specific coverage.

C Question Length Analysis

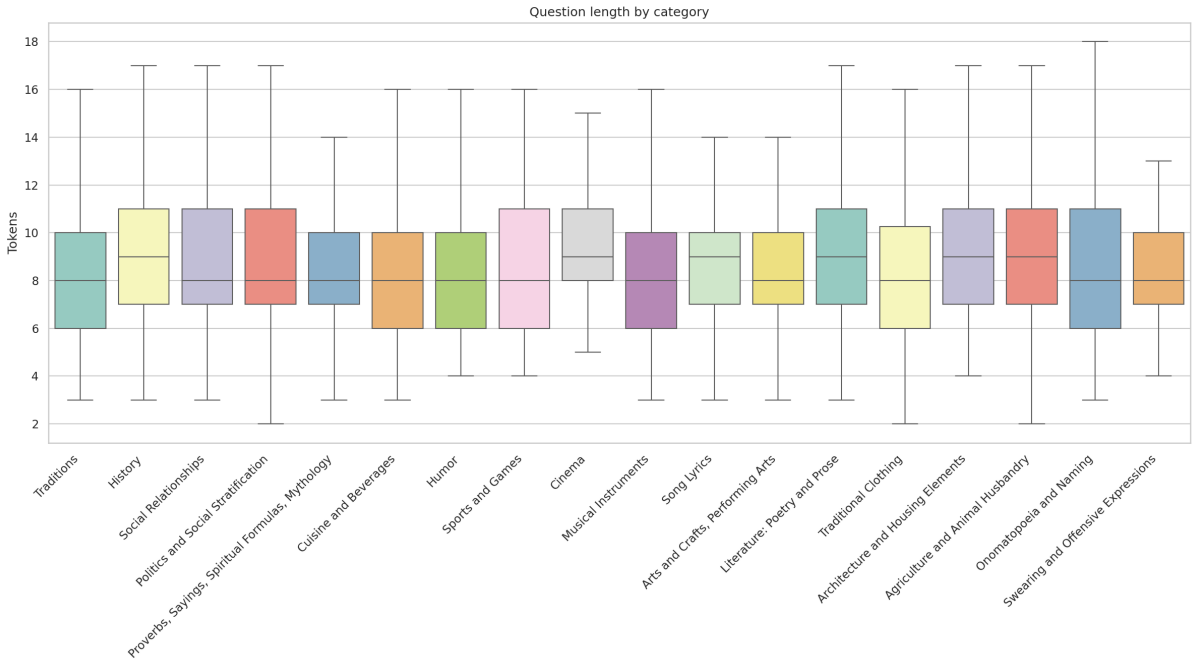

Figure 4: Box plot of question stem lengths (in tokens), showing central tendency and variability across domains.

D Top Token Frequency in Questions

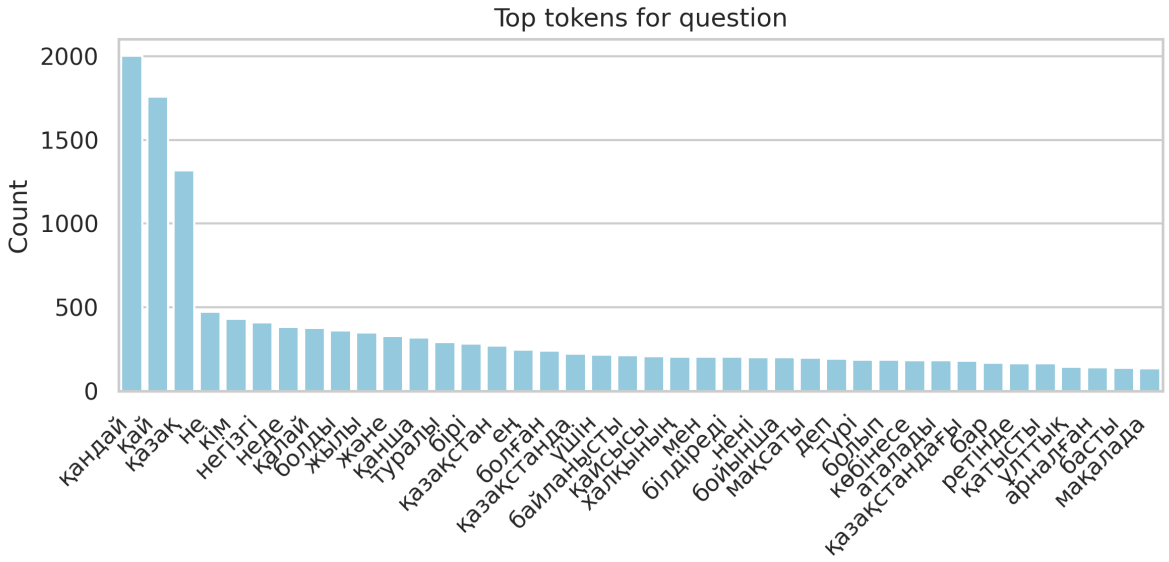

Figure 5: Most frequent tokens in question stems, highlighting common wh-terms and grammatical structures.

E Lexical Diversity by Category

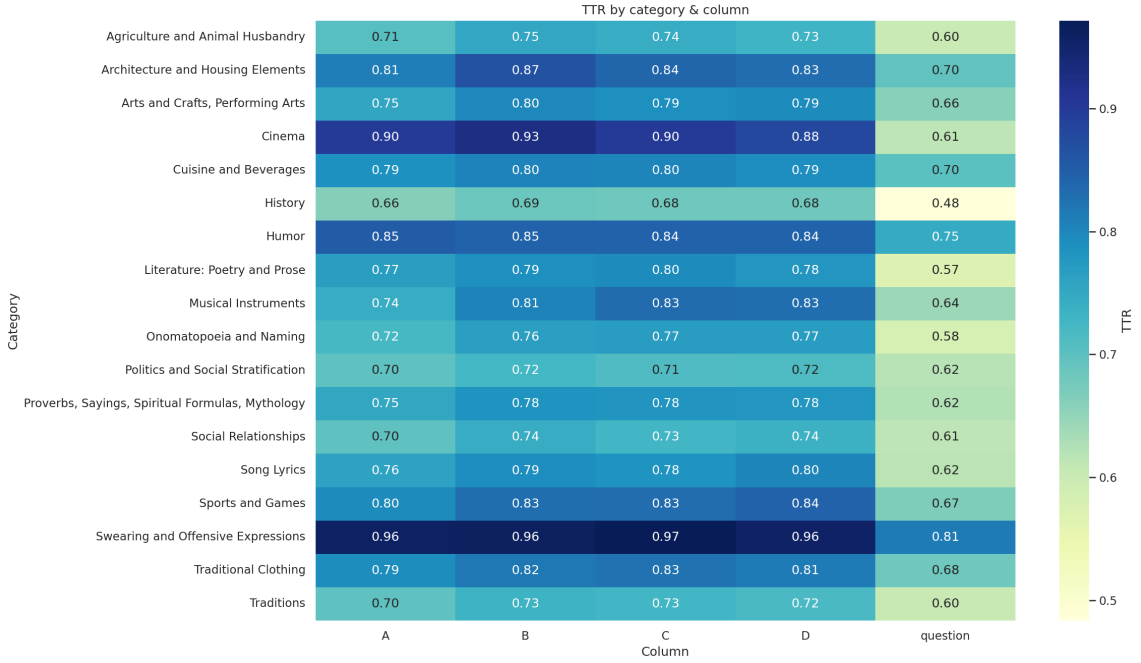

Figure 6: Type-token ratio (TTR) heatmap across categories, illustrating domain-specific variation in lexical richness.

F Answer Key Distribution

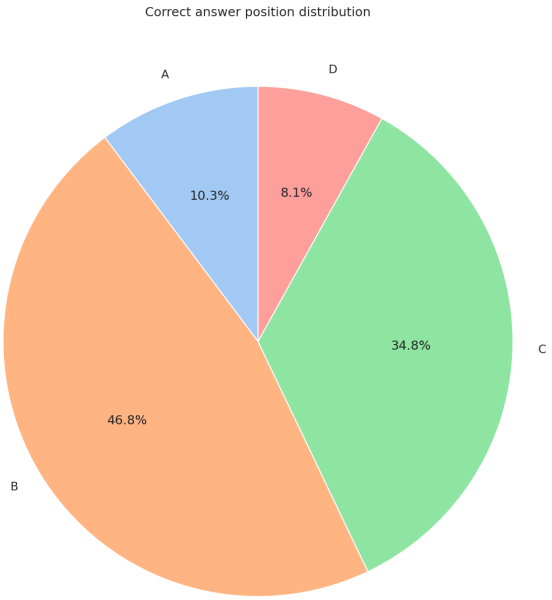

Figure 7: Distribution of correct answer positions (A–D), exposes bias in the dataset after human evaluation and fixes.

G Per-Category Model Accuracy

Model	Arch	Arts	Hist	Cinema	Cuisine	Lit	Swear	Instr	Onom	Polit	Proverb	Agric	Social	Sport	Song	Trad	Cloth	Humor	Avg
google/gemma-3-27b-it	0.759	0.708	0.706	0.688	0.748	0.669	0.649	0.671	0.667	0.788	0.725	0.765	0.746	0.651	0.737	0.779	0.714	0.740	0.722
meta-llama/Llama-3.3-70B-Instruct	0.747	0.714	0.709	0.625	0.696	0.655	0.684	0.675	0.657	0.817	0.703	0.741	0.741	0.675	0.708	0.729	0.723	0.663	0.709
meta-llama/Llama-3.1-70B-Instruct	0.719	0.703	0.697	0.625	0.724	0.653	0.667	0.650	0.652	0.810	0.707	0.737	0.744	0.679	0.685	0.713	0.737	0.673	0.703
nvidia/Llama-3_3-Nemotron-Super-49B-v1	0.723	0.686	0.691	0.598	0.671	0.641	0.649	0.689	0.633	0.792	0.677	0.741	0.741	0.656	0.683	0.715	0.710	0.694	0.694
inceptionai/Llama-3.1-Sherkala-8B-Chat	0.699	0.662	0.692	0.625	0.668	0.681	0.632	0.678	0.641	0.777	0.689	0.727	0.712	0.675	0.685	0.702	0.665	0.714	0.691
issai/LLama-3.1-KazLLM-1.0-70B	0.727	0.668	0.703	0.571	0.675	0.657	0.702	0.636	0.622	0.773	0.693	0.725	0.714	0.665	0.687	0.708	0.696	0.684	0.689
google/gemma-3-12b-it	0.731	0.673	0.669	0.661	0.664	0.647	0.544	0.657	0.630	0.737	0.709	0.694	0.680	0.623	0.693	0.713	0.719	0.679	0.679
mistralai/Mistral-Small-24B-Instruct-2501	0.687	0.681	0.685	0.589	0.671	0.625	0.684	0.661	0.634	0.724	0.659	0.712	0.697	0.618	0.668	0.715	0.692	0.704	0.676
Qwen/Qwen2.5-32B-Instruct	0.699	0.614	0.604	0.598	0.570	0.649	0.509	0.618	0.612	0.717	0.619	0.642	0.638	0.608	0.643	0.658	0.643	0.694	0.633
google/gemma-3-12b-pt	0.671	0.611	0.589	0.518	0.629	0.561	0.649	0.594	0.531	0.695	0.665	0.665	0.675	0.561	0.637	0.692	0.688	0.643	0.624
Qwen/QwQ-32B	0.651	0.605	0.601	0.589	0.573	0.637	0.526	0.590	0.597	0.658	0.615	0.639	0.645	0.599	0.599	0.622	0.589	0.699	0.617
deepseek-ai/DeepSeek-R1-Distill-Llama-70B	0.699	0.576	0.603	0.464	0.587	0.565	0.632	0.565	0.504	0.667	0.625	0.639	0.643	0.561	0.601	0.620	0.634	0.638	0.602
Qwen/Qwen2.5-14B-Instruct	0.631	0.627	0.573	0.536	0.601	0.605	0.456	0.601	0.572	0.658	0.561	0.613	0.638	0.599	0.585	0.649	0.580	0.622	0.600
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	0.635	0.614	0.583	0.589	0.577	0.629	0.491	0.565	0.576	0.634	0.579	0.634	0.589	0.547	0.589	0.640	0.598	0.633	0.600
google/gemma-3-4b-pt	0.643	0.616	0.579	0.509	0.584	0.543	0.509	0.544	0.536	0.631	0.595	0.613	0.601	0.613	0.578	0.608	0.563	0.602	0.585
google/gemma-3-4b-it	0.618	0.578	0.575	0.438	0.580	0.557	0.544	0.640	0.548	0.636	0.581	0.615	0.589	0.552	0.572	0.576	0.643	0.566	0.583
meta-llama/Llama-3.1-8B-Instruct	0.647	0.614	0.573	0.482	0.535	0.545	0.614	0.530	0.507	0.600	0.589	0.606	0.618	0.561	0.570	0.576	0.580	0.622	0.575
issai/LLama-3.1-KazLLM-1.0-8B	0.598	0.568	0.576	0.455	0.549	0.511	0.649	0.516	0.462	0.638	0.569	0.611	0.628	0.524	0.557	0.597	0.585	0.602	0.566
nvidia/Llama-3.1-Nemotron-Nano-8B-v1	0.341	0.351	0.359	0.339	0.374	0.311	0.386	0.392	0.327	0.355	0.365	0.370	0.340	0.406	0.347	0.346	0.402	0.342	0.354
TilQazyna/llama-kaz-instruct-8B-1	0.233	0.235	0.282	0.295	0.318	0.281	0.193	0.325	0.264	0.291	0.327	0.287	0.249	0.288	0.261	0.264	0.237	0.265	0.277

Table 6: Per-category accuracy (and macro average) for each evaluated model. Column abbreviations: **Arch** = Architecture/Housing, **Arts** = Arts/Crafts, **Lit** = Literature, **Swear** = Swearing expressions, **Instr** = Musical instruments, **Onom** = Onomatopoeia, **Polit** = Politics/Social, **Proverb** = Proverbs & Mythology, **Agric** = Agriculture, **Trad** = Traditions, **Cloth** = Traditional clothing.

E Semi-Automated Data Generation Pipeline

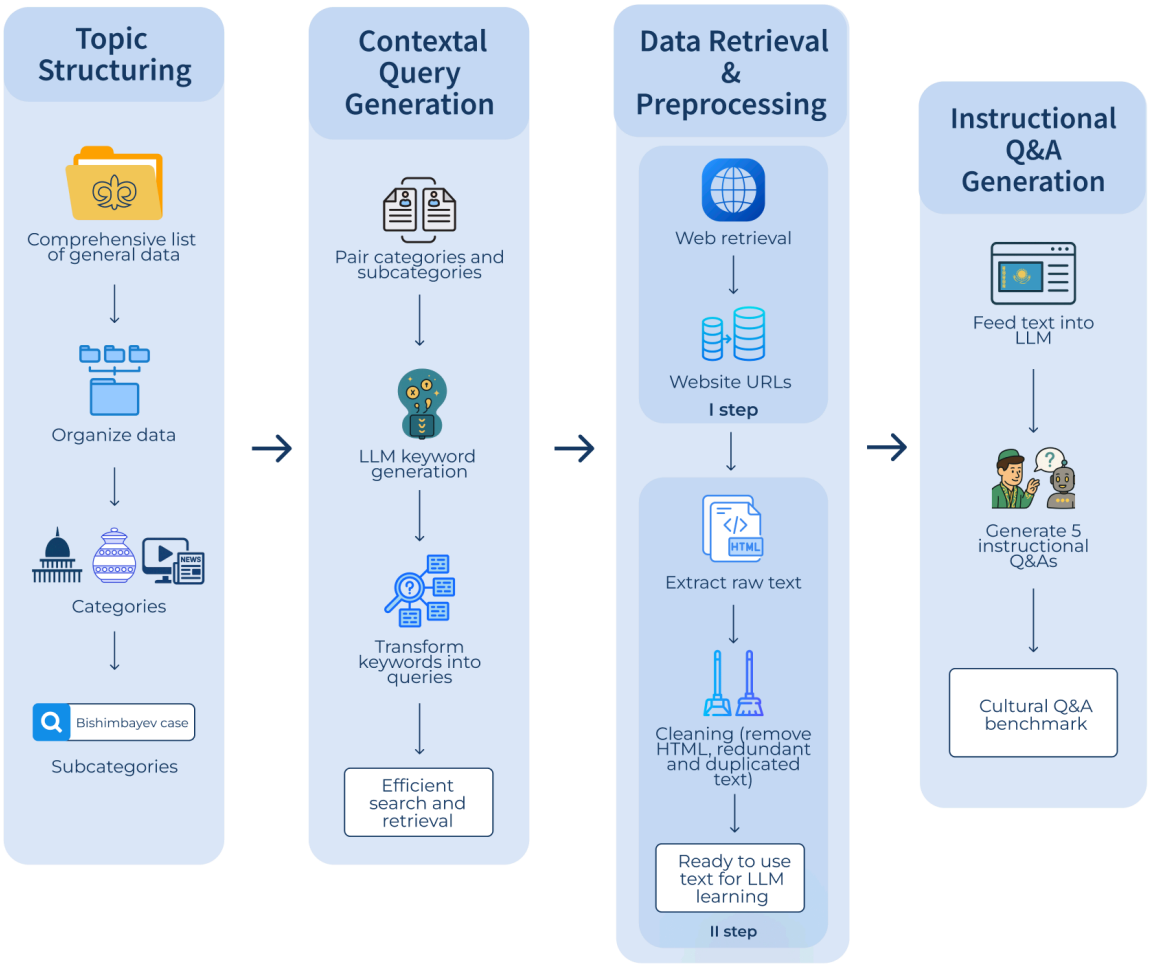

Figure 8: Overview of the semi-automated pipeline used to generate culturally-grounded instructional Q&A benchmark.

Searchable Language Documentation Corpora: DoReCo meets TEITOK

Maarten Janssen
Faculty of Mathematics and Physics
Charles University, Czechia
janssen@ufal.mff.cuni.cz

Frank Seifart
SeDyL CNRS, France
Humboldt-Universität zu Berlin, Germany
frank.seifart@cnrs.fr

Abstract

In this paper, we describe a newly created searchable interface for DoReCo, a database that contains spoken corpora from a worldwide sample of 53, mostly lesser described languages, with audio, transcription, translation, and - for most languages - interlinear morpheme glosses. Until now, DoReCo data were available for download via the DoReCo website and via the Nakala repository in a number of different formats, but not directly accessible online. We created a graphical interface to view, listen to, and search these data online, providing direct and intuitive access for linguists and laypeople, including members of speech communities. The new interface uses the TEITOK corpus infrastructure to provide a number of different visualizations of individual documents in DoReCo and provides a search interface to perform detailed searches on individual languages.

1 Introduction

Over the past 30 years, spoken corpus data have been produced through linguistic fieldwork on hundreds of languages around the world, often in attempts to document languages that are threatened of becoming extinct (Seifart et al., 2018). Typically, these were archived as part of language documentation collections in repositories such as TLA[1] and ELAR[2]. However, within these collections, the corpus data are often not easily identifiable and subject to access restrictions. Recently, the DoReCo database brought together selected high-quality corpus data from such collections, harmonized their annotations, and made them available for download (Seifart et al., 2024).

But even in DoReCo, these data are served only as raw source data, that is, as files from their respective tools. So in order to access them, users have to download the data, install the corresponding tool, and use the data locally using that tool, for instance, ELAN[3]. This means that it is not trivial for casual users to access such language documentation corpora, even though they could be valuable resources for simply getting an impression of the language or for university teaching, as well as linguistic research.

In this paper, we demonstrate how to make spoken corpora stemming from fieldwork-based language documentation directly accessible online, including for online corpus searching, by converting the source data to a corpus search tool with an online interface, building on existing tools and formats. In the example here, we convert DoReCo to TEITOK, a web-based corpus management platform that provides specific tools for spoken data and interlinear glossed data (Janssen, 2016). We first briefly describe DoReCo and TEITOK and then describe how the DoReCo data were converted into a TEITOK corpus. And finally we will demonstrate how the TEITOK online interface of the DoReCo data can be used to quickly and efficiently access the fieldwork data. The use of TEITOK also enables the corpus for use with NLP pipelines, either using the data to train NLP models or to use NLP models to further enrich the data.

2 DoReCo

DoReCo (Language Documentation Reference Corpus) is a collection of spoken corpora on a diverse set of 53 languages from around the world, with a focus on small and endangered languages. It was conceived to make data that were painstakingly collected in fieldwork in often remote areas available for cross-linguistic and cross-cultural research. As such, it addresses the problem of overreliance on what has been termed WEIRD (Western Educated Industrial Rich Democratic) populations and their languages in cognitive science (Henrich et al.,

[1] https://archive.mpi.nl/tla/
[2] https://www.elararchive.org/
[3] https://www.mpi.nl/corpus/html/elan/

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 58–64
August 1, 2025 ©2025 Association for Computational Linguistics

2010; Blasi et al., 2022).

Most of the corpora in DoReCo stem from efforts to document endangered languages in the framework of documentary linguistics (Himmelmann, 1998). From such documentary collections, DoReCo selected data that were suitable for cross-linguistic corpus-based research (Schnell and Schiborr, 2022). The selection criteria included the quality and consistency of annotation, the quality of the accompanying audio-recordings and that these materials could be made available using CC-BY licenses. The majority of DoReCo data are spontaneously produced traditional or personal narratives, in addition to some conversations and stimulus retellings, but not isolated examples. Each corpus had been transcribed, translated and, for 39 languages, also morphologically annotated by experts on the language prior to their inclusion in DoReCo. These experts are also the authors of the individual corpora that are edited and made available through DoReCo.

Within DoReCo, these data have been processed to add time alignment of transcription and audio through a combination of automatic forced alignment and manual corrections (Paschen et al., 2020). As a result, the start and end times for each phone, morph, and word unit are now annotated - a design motivated by research questions on phonetic lengthening (Blum et al., 2024). Other data processing steps in DoReCo included the harmonization, across the 53 corpora, of the tier structure and tier names, the documentation of the phonetic value of symbols used in the transcription, and the creation of csv files for each language, one with one word per line and another with one phone per line. DoReCo was first published in 2022, and the latest major update, containing 53 languages, was published in 2024. All DoReCo data are distributed under CC BY(-NC)(-ND-/SA) licenses.

3 TEITOK

TEITOK is an online platform for creating, managing, visualizing, and searching annotated corpora. All corpus documents in TEITOK are stored in a tokenized TEI/XML format[4]. It has a modular setup with various search and visualization methods. The default search is performed using Corpus WorkBench (CWB) (Evert and Hardie, 2011), which allows rich queries that can combine various token attributes, sequences of tokens, and can take metadata into account. The default document visualization shows linguistic information and is designed to display lemmatization, POS tagging, and dependency data. But there are also visualization modules for facsimile-aligned manuscript-based corpora, for time-aligned audio-based corpora (Janssen, 2021), and for interlinear glossed text corpora.

TEITOK was initially developed for the diachronic corpus PostScriptum (Vaamonde et al., 2014) and the learner corpus COPLE2 (Mendes et al., 2016), and has since been used for a wide variety of corpora including the multilingual Universal Dependencies corpus[5], the parliamentary corpus ParlaMint (Janssen and Kopp, 2024), dialectal corpora such as Madison[6], and corpora on less-resourced languages like CoDiaJE on Judeo-Spanish (Quintana, 2020). A list of publicly accessible TEITOK projects can be found on the TEITOK website[7].

TEITOK actively supports corpus editing and does not typically rely on corpora that have been fully developed outside of the platform. It allows users to run NLP pipelines by default using UDPIPE[8] on their data from the interface, in order to easily enrich a corpus with NLP data such as tagging, lemmatization, and dependency parsing. For fieldwork data, there typically are no NLP pipelines available, but TEITOK also allows training a tagger on the manually annotated data in the corpus, to automatically pre-tag subsequent documents with the recently trained tagger. And it provides an intuitive interface to add and correct annotations, so that errors in the automatic annotations can be corrected. This mechanism has been used, for instance, in the CoDiaJE corpus mentioned above to create a POS tagger from scratch for a language for which no NLP tools were available.

TEITOK is actively maintained and extended with new functionalities, and has an active user base. It is open source and can be easily installed anywhere from the repository[9], or run in a virtual environment from DockerHub[10]. TEITOK has been generally well received both by corpus creators and corpus users. The fact that it makes use of well established formats and tools such as TEI/XML

[4] https://tei-c.org/

[5] https://lindat.mff.cuni.cz/services/teitok/ud214/index.php

[6] http://teitok.clul.ul.pt/madison/

[7] http://www.teitok.org/index.php?action=projects

[8] https://lindat.mff.cuni.cz/services/udpipe/

[9] https://gitlab.com/maartenes/TEITOK/

[10] https://hub.docker.com/r/maartenpt/teitok

and the Corpus WorkBench means that many people will be familiar with various aspects of the interface even if they do not know the tool itself.

4 DoReCo in TEITOK

In order to create a searchable version of DoReCo in TEITOK, all original DoReCo files were converted to the TEITOK file format. The interface follows the same design layout as the DoReCo website, even though it is hosted on a different server, highlighting that it provides a visualization of the existing DoReCo version, not a re-edition.

Spoken corpora, including the DoReCo corpora, often closely transcribe what is said by the speakers, keeping track of pauses, corrections, false starts, etc. Transcription of such phenomena is typically performed in fieldwork-specific tools such as the Field Linguist Toolbox[11] or in speech-driven tools like ELAN. Since such tools use plain text for the transcription, all labels (or codes) for speech phenomena like corrections or false starts are transcribed by using special characters and labels inside the transcription.

The encoding of these phenomena by means of special characters is not ideal for a number of reasons. The first is that the labels tend to vary from corpus to corpus, so it is always necessary to provide a legend along with the corpus to explain the labels. The second is that these manually added labels are often not computer readable if they are not applied 100% consistently. The third is that the labels impede easy searching of the corpus: if we use the symbol / for a pause, then searching for "the man" will not yield results that have a pause in the middle (the / man).

The TEITOK conversion converts all DoReCo labels for disfluencies etc. into TEI/XML markup. XML is a formal language that has to be used systematically and TEI provides a set of standardized, well-described markers. This makes the resulting TEI documents compatible with other spoken data. The meaning of the markers used can be looked up in the TEI documentation for those who are not familiar with them. And they do not interfere with searches because searches are done on sequences of tokens ignoring markers. In the next section, we show how the conversion was done, and then how the converted corpus can be used for visualization and searching.

[11] https://software.sil.org/toolbox/

4.1 Conversion

The conversion from DoReCo to TEITOK was done completely automatically by a custom script that combines the metadata from the DoReCo metadata table with the transcription data from the ELAN (EAF) files. The script reads each line in the metadata table and then for each line creates a TEI/XML file in TEITOK style and saves it under the identifying name of that line. The metadata are placed in their appropriate TEI fields in the header (teiHeader), while the transcription is placed in the body (text). As a corpus search environment, TEITOK does not work with tiers, but rather with running text. For spoken corpora with multiple speakers, this typically implies an "interview style" representation of the text, in which speech turns are presented in chronological order, determined by their start time, also in case of overlapping turns.

The technical implementation of the conversion of the EAF files is as follows. Each annotation unit on the DoReCo REF tier(s), which represents a chunk of speech defined by the corpus creators as sentences, intonation units, or larger units like paragraphs, and which is associated with a translation unit, is turned into an utterance (u). The utterances are ordered chronologically by their start time to generate the interview-style representation of the text. The utterances get adorned with attributes taken from all the tiers that correspond to the utterance: the start and end time from the interval, the speaker identifier (who) from the name of the tier (ref@XX), the identifier (id) from the REF tier, the text from the TX tier, and the translation from the FT tier.

Inside the utterance, it creates tokens (tok) for each annotation within the range of the utterance from the WD tier, with the inner text corresponding to the content of the WD tier and attributes from all dependent tiers. Within each token, it creates morphemes (m) from the MB tier with its respective attributes. When the start and end times are available for tokens and morphemes, they are also added to the respective nodes.

The CWB searches do not work with units smaller than the token, which means that morphs and phones (approximated by units transcribed with X-SAMPA symbols) are not directly searchable. Therefore, the content of the MB and PH tiers are (also) kept as single string on the token, concatenating the content of the various elements. Morphemes are separated by a dot, while the X-SAMPA charac-

ters are separated by a space to increase readability. The identification of morpheme breaks follows the language-expert annotations in DoReCo.

The disfluency labels in the DoReCo transcription are converted into their respective TEI codes, as shown in Table 1. This way, the custom codes used in DoReCo are converted to standardized markers and separated from the text.

An example of a one-word sentence *Эвīлэн* from the file `2007_Ekonda_Udygir_Viktor_FSk3` in the corpus on the Siberian language Evenki (Kazakevich and Klyachko, 2024) is given in Table 2 (with some details left out for clarity).

The header of the TEI file tracks whether or not the audio file for the transcription is available. Generally, DoReCo corpora consist of a core set of annotations which have been time-aligned and for which the audio is available, and a larger set for which that is not the case. For eight DoReCo corpora, however, the audio files are only available at repositories outside DoReCo after registration, so these were not made available in the TEITOK corpus either. When the audio is not available, audio related functions are disabled for that file.

The entire conversion is fully automatic and would work not only for possible future versions or extensions of DoReCo, but also for any ELAN data following the DoReCo set-up.

4.2 Visualizations

There are three main ways to visualize the individual files in the DoReCo-TEITOK: a linguistic view, a speech-oriented view, and an interlinear glossed view. The linguistic view displays the full text of the transcription with the audio file displayed on top. Moving the mouse over one word will display a pop-up that shows all the information available for that word: the word itself, the morphological breakdown with their glosses (where available), the POS tag, and the X-SAMPA transcription.

The speech-oriented visualization displays the waveform of the audio file on top and below that the transcription of the utterances. Clicking on an utterance will play that utterance. Playing the audio will highlight which word in the transcription is currently being pronounced, and the word will also appear as a caption in the waveform image. While the waveform scrolls horizontally, the transcription scrolls vertically. Naturally, the speech-oriented visualization is available only for those transcriptions that have audio files that accessible. And the word-level visualization is only available for the files that were time-aligned at the word level. An example of a waveform visualization from the Evenki DoReCo corpus (Kazakevich and Klyachko, 2024) is given in Figure 1.

The default visualization of DoReCo-TEITOK is set to the interlinear glossed text (IGT) visualization. This is because for transcriptions that have a morphological breakdown, neither the linguistic nor the speech-oriented visualization will display the morphemes. The IGT view displays each utterance in sequence, with first the utterance, then the words of the utterance with below each word the token-level annotations such as POS, gloss, and X-SAMPA. Below that, it displays the morphemes of each word with the morpheme level annotations such as form and gloss, and, finally, the utterance-level annotations such as full text translation and the option to listen to the utterance. An example of an IGT visualization from the Evenki DoReCo corpus is given in Figure 2.

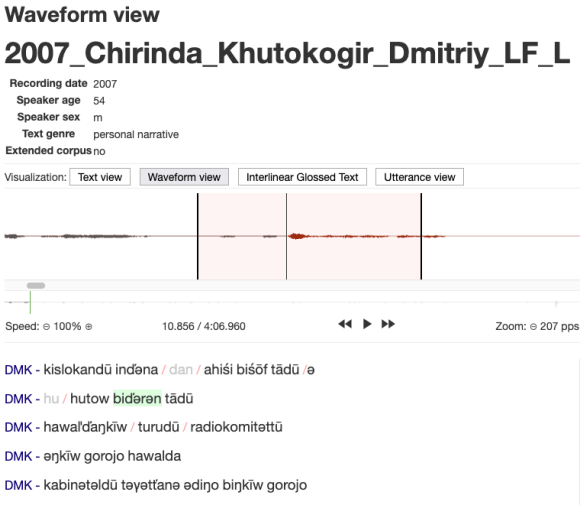

Figure 1: Waveform view example (Evenki corpus (Kazakevich and Klyachko, 2024))

4.3 Searches

From the collection of converted TEI/XML files, an indexed corpus is created in Corpus WorkBench (Evert and Hardie, 2011), making the various kinds of metadata searchable, along with all the attributes on the utterances and the tokens. The corpus can be searched using the Corpus Query Language (CQL). CQL is a well established and powerful query language used by many tools including for instance CQPWe(Hardie, 2012) and SketchEngine(Kilgarriff et al., 2014), and should be familiar to many potential users - but for people not familiar with it, TEITOK provides a user friendly

Filled pause	<<fp>uhm>	<vocal><desc>uhm</desc></vocal>
	<<fp>>	<pause type="filled"/>
Prolongation	<<pr>looonger>	<tok obs="prolongued">longer</tok>
Backchannel	<<bc>mm>	<vocal><desc>mm</desc></vocal>
False start	<<fs>fal->	<del type="falsestart">fal-</del>
Ideophone	<<id>tick>	<tok obs="ideophone">tick</tok>
Onomatopeic	<<on>moo>	<vocal type="onomatopeic"><desc>moo</desc></vocal>
Foreign material	<<fm>Weberei>	<foreign><tok>Weberei</tok></foreign>
Unidentifiable	<<ui>vubi>	<unclear><tok>vubi</tok></unclear>
	<<ui>>	<gap reason="unidentifiable"/>
Singing	<<sg>>	<gap reason="singing"/>
Silent pause	<p:>	<pause type="silent"/>
Word-internal pause	<<wip>>	<pause type="word-internal"/>

Table 1: Disfluency code conversion

```
<u who="VNU" tier="ref" start="317.92" end="318.42" eid="0089_doreco_even1259_2007_Ekonda_Udygir_Viktor_FSk3" text="¯." gloss="He began to play." id="u-86">
    <tok who="VNU" tier="wd" start="317.92" end="318.42" form="wīln" pos="v" phon="@wi:l@n" morph="wɪ.-l.-.-n" id="w-358">
        wīln
        <m who="VNU" tier="mb" start="317.92" end="318.15" form="wī" gloss="" id="m-358-1"/>
        <m who="VNU" tier="mb" start="318.15" end="318.25" form="-l" gloss="INCH" id="m-358-2"/>
        <m who="VNU" tier="mb" start="318.25" end="318.28" form="-" gloss="NFUT" id="m-358-3"/>
        <m who="VNU" tier="mb" start="318.28" end="318.42" form="-n" gloss="3SG" id="m-358-4"/>
    </tok>
</u>
```

Table 2: Example utterance in TEITOK/XML (Evenki corpus (Kazakevich and Klyachko, 2024))

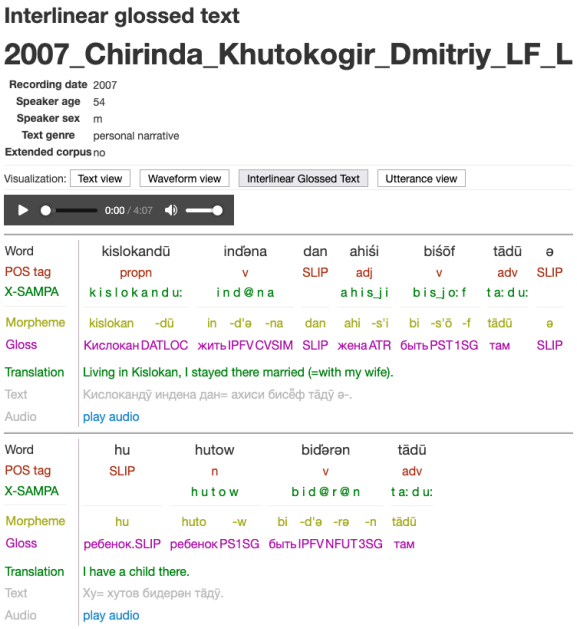

Figure 2: Interlinear glossed text view example (Evenki corpus (Kazakevich and Klyachko, 2024))

GUI to build search queries.

CQL can be used to search for words (or sequences of words) and to restrict that search to specific documents or utterances. These searches can combine any of the attributes present in the corpus: the form and X-SAMPA representation of the word, the part-of-speech (POS) tag (when available) or glosses. They can also be restricted to utterance by speakers of a certain sex or age, and to documents of a specific genre or to the core vs. extended (without time-alignment and audio) corpus sections.

This makes it possible to quickly find examples in the corpus, which facilitates its use, for instance, in teaching in linguistics programs. The results can also be used for statistical data by grouping the results by one of the categories. This makes it possible, for instance, to see the distribution of words over the different POS tags, to see whether certain types of words are more frequently used by women, or in narrative texts. The search results are rendered as utterances, and when a sound file is available, it will have a play button next to the result, making it possible to directly listen to the utterances.

Since all text-based codes in the original DoReCo data have been converted to TEI/XML codes according to Table 1, all text is searchable and disfluencies, gaps, and other markings do not hamper the search, while the information they provide is still available.

5 Conclusion

In this paper, we have shown how we created a searchable, directly accessible version of the DoReCo corpus making use of the built-in capacities of the TEITOK platform. This TEITOK ver-

sion of DoReCo is much easier to use for casual users and allows expressive searches and frequency counts to quickly find examples or quickly extract some general information on the language, for example in teaching settings.

TEITOK visualization and search functions focus on textual information present in DoReCo. It disregards the speech-related aspects of DoReCo, especially the time-alignment of annotation with audio at the phone-, morph- and word-level. For analyses taking these into account, the original DoReCo files in combination with speech-specific tools like ELAN and Praat offer functions that a corpus tool like TEITOK does not.

Currently, the DoReCo corpus in TEITOK only represents the information already provided by DoReCo. But having fieldwork corpora from ELAN made available in TEITOK not only makes it possible for casual users to search the corpus online, but also makes it possible for the corpus creators to enrich their corpus data with further annotations, such as a lemmatization, POS tags, Named Entities, or full dependency treebanks in the framework of Universal Dependencies[12]. The platform has been designed to help provide the necessary manual annotations, and furthermore provides an interface to then use the manually annotated data to train NLP tools like taggers, parsers, and named entity recognition tools.

The current project applied TEITOK to DoReCo data, but the visualization and search functions shown here would work for many other language documentation corpora. By providing fieldwork corpus data on more languages in the same interface and following the same principles, the cross-linguistic coverage for comparative corpus research could be enhanced even further. Therefore, it would be beneficial to the community if more language documentation corpora were made available in the same fashion.

References

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

Frederic Blum, Ludger Paschen, Robert Forkel, Susanne Fuchs, and Frank Seifart. 2024. Consonant lengthening marks the beginning of words across a diverse sample of languages. *Nature Human Behaviour*, 8(11):2127–2138.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Andrew Hardie. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380 – 409.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Maarten Janssen. 2021. A corpus with Wavesurfer and TEI: Speech and video in TEITOK. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 261–268, Berlin, Heidelberg. Springer-Verlag.

Maarten Janssen and Matyáš Kopp. 2024. ParlaMint in TEITOK. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 121–126, Torino, Italia. ELRA and ICCL.

Olga Kazakevich and Elena Klyachko. 2024. Evenki DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

[12] https://universaldependencies.org/

pages 2657–2666, Marseille, France. European Language Resources Association.

Aldina Quintana. 2020. CoDiAJe–the annotated diachronic corpus of judeo-spanish. *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, (9):209–236.

Stefan Schnell and Nils Norman Schiborr. 2022. Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics*, 8:171–191.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation 25 years on. *Language*, 94(4):e324–e345.

Frank Seifart, Ludger Paschen, and Matthew Stave, editors. 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Gael Vaamonde, Ana Luísa Costa, Rita Marquilhas, Clara Pinto, and Fernanda Pratas. 2014. Post Scriptum: archivo digital de escritura cotidiana. *Janus. Humanidades digitales: desafíos, logros y perspectivas de futuro.*, pages 473–482.

A Practical Tool to Help Automate Interlinear Glossing: a Study on Mukrī Kurdish

Hiwa Asadpour*[1], Shu Okabe*[2,3], Alexander Fraser[2,3,4]
[1]Goethe University Frankfurt
[2]School of Computation, Information and Technology, Technische Universität München (TUM)
[3]Munich Center for Machine Learning
[4]Munich Data Science Institute
* Equal contribution
Correspondence: asadpour@lingua.uni-frankfurt.de, shu.okabe@tum.de

Abstract

Interlinear gloss generation aims to predict linguistic annotations (gloss) for a sentence in a language that is usually under ongoing documentation. Such output is a first draft for the linguist to work with and should reduce the manual workload. This article studies a simple glossing pipeline based on a Conditional Random Field and applies it to a small fieldwork corpus in Mukrī Kurdish, a variety of Central Kurdish. We mainly focus on making the tool as accessible as possible for field linguists, so it can run on standard computers without the need for GPUs. Our pipeline predicts common grammatical patterns robustly and, more generally, frequent combinations of morphemes and glosses. Although more advanced neural models do reach better results, our feature-based system still manages to be competitive and to provide interpretability. To foster further collaboration between field linguistics and NLP, we also provide some recommendations regarding documentation endeavours and release our pipeline code alongside.

1 Introduction

Language documentation aims to create and archive corpora alongside resources on a language usually classified as endangered. To do so, linguists carry out fieldwork and then process the collected data. Each annotation (e.g., transcribing the recordings, analysing the transcription) is mostly done manually; it is hence costly in terms of time and requires advanced linguistic knowledge. This is the 'transcription bottleneck' (Brinckmann, 2008), which underlines the gap between the amount of unannotated recordings and the fully annotated sentences. In this article, we focus on one of the central linguistic annotations, interlinear glosses, and aim to predict them automatically, to create a draft for the linguists to post-edit. It has been previously shown that such automation can actually help linguists both in terms of time and annotation quality (Baldridge and Palmer, 2009; Palmer et al., 2009).

1	Source	de	tirsī	kābrāy
2	Segmented	de	tirs=ī	kābrā–ī
3	Gloss	in	fear=EZ	fellow–OBL
4	Translation	out of the fear of the man		

Figure 1: Sentence annotated in the IGT format.

Figure 1 shows an example of an annotated sentence in the Interlinear Glossed Text format (IGT). The source sentence (1) is segmented into morphemes (2), the smallest meaningful units in the language. Each morpheme has a corresponding linguistic annotation, the gloss (3). We observe mainly two categories: grammatical glosses indicate the role of the morpheme (e.g., 'OBL' for oblique), while lexical glosses express its meaning (e.g., '*tirs*' for fear in English). Finally, the sentence is translated (4) in a meta-language used for the documentation (e.g., in English here).

Several languages and corpora have already been studied by the Natural Language Processing (NLP) community for the gloss generation task, for instance, during the SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023). We focus, however, on the usability of an automatic glossing model in a real-life setting of an annotation workflow. This means that we take into account actual technical constraints that hinder the use of the most up-to-date NLP models.

To do so, we base our work on a corpus from one of the authors' fieldwork data (Asadpour, 2021) to enable linguistic analysis of the glossing. The studied language is Mukrī Kurdish, a variety of Central Kurdish, whose morphological complexity can be challenging. As a Kurdish language, it has a rich agglutinative system characterised by *ezafe* (linking) constructions, polypersonal agreement, and a variety of affixed, cliticised, and reduplicated

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 65–75
August 1, 2025 ©2025 Association for Computational Linguistics

morphemes.

We present a simple pipeline using a feature-based model to label each source morpheme in Mukrī Kurdish with a gloss. Our work mostly focused on how to make such an NLP model more accessible for field linguists and closer to their workflow. Our model is indeed achieving performance around a few accuracy points behind state-of-the-art models, while it only requires stable Python dependencies with minimal computational resources (CPU of a standard computer). The pipeline can also output annotations in a format compatible with commonly used linguistic fieldwork tools.

Our contributions are as follows: (i) we release a feature-based minimal system for automatic glossing[1], (ii) we apply it to a manually annotated text from a real fieldwork corpus of one of the authors, and (iii) analyse the linguistic relevance of the predictions and learnt patterns.

Section 2 describes Mukrī Kurdish and the glossed corpus we studied. We explain our CRF pipeline methodology in Section 3. We present its performance and analyse the linguistic patterns learnt by the model in Section 4. We also point out a few recommendations for both field linguists and NLP practitioners in Section 5.

2 Language and fieldwork corpus

2.1 The language: Mukrī Kurdish

Mukrī Kurdish (also spelt Mukrīyānī) is primarily spoken in the northwestern region of Iran, specifically in middle and southern parts of West Azerbaijan and northern parts of Kurdistan provinces. The geographical area traditionally associated with Mukrī Kurdish is centred around the city of Mahābād (historically known as Sāblāx or Sāwjbłāx) and extends to surrounding cities, towns and villages, including Bokān, Pīrānšār, Sardašt, Šino and Naxada (Asadpour, 2021). This region, historically known as Mukrīyān, forms part of the larger Iranian Kurdistan area that borders Iraqi Kurdistan to the west.

Mukrī Kurdish belongs to the Central Kurdish (Sorānī) dialect group within the Indo-European language family. It is closely related to other Central Kurdish varieties spoken in both Iran and Iraq. However, it maintains distinctive features that set it apart from standard Sorānī as spoken in Silēmānīya or Hawlēr (Erbil) in Iraqi Kurdistan (Haig and Matras, 2002; Asadpour, 2021, 2022).

Among Central Kurdish varieties, Mukrī Kurdish has several distinctive characteristics. On the phonological aspect, certain vowel and consonant realisations differentiate it from standard Sorānī varieties, including retention of some archaic phonological features. On the lexical side, its unique vocabulary is influenced by its geographic position between different Kurdish dialect areas and contact with Jewish and Christian Neo-Aramaic, Armenian, and Azerbaijani Turkish communities (Asadpour, 2021).

Moreover, Mukrī Kurdish has a rich morphological structure with prefixes, suffixes, and enclitics. Correct morphological labelling requires an awareness of the surrounding context, such as in the example below:

Source	ne–	bird	–ī	=ewe
Gloss	NEG.PST–	take.PST	–2SG	=ASP
Translation	you did not take			

with a negation prefix *ne–*, a past verb stem *bird*, a person suffix *–ī*, and the aspectual enclitic *=ewe*. Verbal morphology, in particular, requires both left and right contexts for correct segmentation and interpretation. We note here that certain morphological markers are consistent and predictable both in form and position. For instance, verbs begin with mood/aspect prefixes (e.g., negation in the example), end with person suffixes (e.g., *–ī* for 2SG), and aspectual enclitics may also appear in the final position (e.g., *=ewe*).

2.2 Corpus preparation

We use the corpus collected through fieldwork by one of the authors (2004–in progress) in the Mukrī variety of Central Kurdish (Sorānī). The corpus includes narrative, conversational and procedural texts, ensuring diversity in genre and register. Annotation was done manually following the IGT format and Leipzig Glossing Rules (Lehmann, 2004; Bickel et al., 2008). Besides, the segmentation annotation tier marks morpheme boundaries with hyphens, while clitics are separated by equal signs (cf. tier 2 in Figure 1).

We split the corpus into training and test datasets (80:20) for our experiments. We also convert the sentences into the format used for the SIGMORPHON Shared Task (Ginn et al., 2023), with one sentence annotation tier per line. This notably ensures compatibility with tools devised for the Shared Task.

[1]The pipeline is released alongside a demonstration at: https://github.com/shuokabe/crf_glossing.

Table 1 displays the size of the fieldwork corpus of Mukrī Kurdish in terms of number of sentences (N_{sent}), number of words and morphemes for both tokens (N_{token}) and types (N_{type}).

		word		morpheme	
	N_{sent}	N_{token}	N_{type}	N_{token}	N_{type}
train	211	1,233	570	2,126	354
test	52	272	184	500	153

Table 1: Fieldwork corpus statistics for Mukrī Kurdish.

3 Gloss generation system

3.1 Gloss generation pipeline

We tackle the gloss generation task as a morpheme labelling task. We assume that the sentence has been previously segmented into morphemes. Interlinear glosses can hence be viewed as labels assigned to each morpheme.

Source	de	tirs=ī	kābrā–ī
Step I	IND	stem=EZ	stem–OBL
Step II	IND	UNK=EZ	fellow–OBL
True gloss	in	fear=EZ	fellow–OBL

Figure 2: Example output at each step from the model.

Our model can be decomposed into two steps, as presented in Figure 2. First, grammatical labels are predicted for each morpheme (step I), with lexical morphemes initially labelled as 'stem' placeholders. Then, these placeholder labels are replaced with actual lexical glosses using a simple dictionary built from frequent associations in the training data (step II).[2] When available, actual bilingual dictionaries or known morpheme-to-gloss mappings can be integrated to augment the lexical coverage in this step. For unknown lexical morphemes, the second step outputs the 'UNK' tag. Figure 3 summarises the pipeline.

As previously considered by (McMillan-Major, 2020; Barriga Martínez et al., 2021), our system is based on a Conditional Random Field (CRF) (Lafferty et al., 2001), which relies on local properties (or features) to predict a label. We use the default parameters in our experiments.

We use generic features to keep it adaptable to other languages, such as the current morpheme, its

[2]The dictionary contains one-to-one associations only, i.e., one source lemma can only have one possible lexical label.

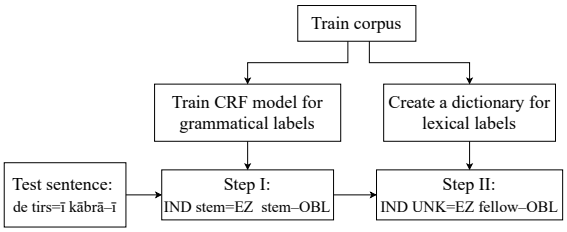

Figure 3: Glossing pipeline flowchart

immediate predecessors and successors, morpheme length, and boundary markers indicating whether the morpheme is separated by a hyphen (–) or an equal sign (=). An example list of features is presented in Appendix A.

3.2 Between simplicity and complexity

Technical requirements The main strength of our system is its simplicity, making it possible to run efficiently on CPUs rather than requiring GPUs. For instance, most participating submissions to the SIGMORPHON Shared Task (Ginn et al., 2023) used neural systems based on transformers (e.g., ByT5 (Xue et al., 2022)) or needed PyTorch to run (e.g., (Girrbach, 2023)'s winning system). In contrast, our approach pushes towards usability in real language documentation settings, where access to GPUs may be limited. This also means the model can run on common laptops within minutes, making it suitable for further integration into annotation workflows.

On the technical side, our CRF uses the sklearn-crfsuite library (Okazaki, 2007) in Python[3], which is widely used and stable. Besides this toolkit, our pipeline does not need any external packages.

Quality of the predictions However, this simplicity comes at a price. Compared to more advanced neural models, our pipeline shows lower overall performance, as shown in Section 4.1. It seems more adapted when the corpus is rather small, notably at the beginning of the annotation phase.

Furthermore, due to the pipeline approach, errors at step I impact the second step. In Figure 2, we see that the first morpheme is wrongly predicted with a grammatical tag (IND), although it should have been a 'stem' label for lexical glosses. Besides, even though our experiments show that lexical glosses can be relatively easily labelled with dictionaries in many cases due to the annotation

[3]https://sklearn-crfsuite.readthedocs.io/.

regularity in documentation corpora, this reliance means that unknown morphemes cannot be handled at all, as for the second morpheme ('tirs') which was never seen in the training corpus.

Interpretability and flexibility Another characteristic of our pipeline is that it allows for better interpretability, as developed in Section 4.4, given its feature-based nature. This can be helpful in understanding the patterns in the predictions and behaviour of the model, ensuring better transparency for the linguists compared to more black-box neural models. This follows previous analyses as in (Barriga Martínez et al., 2021; Okabe and Yvon, 2023a,b).

Besides, our current pipeline remains generic and only requires an annotated training dataset. When language-specific phenomena are known, the CRF can integrate them as additional features; when more annotations are made, the dictionary can be easily expanded for new words, and the CRF will be more robust.

In a nutshell, we chose to focus on a system with reduced technical complexity, trading performance for better accessibility, because we have real-life settings in mind. We recall that the purpose of automatic glossing is to *reduce* the proportion of manual workload by providing a first draft to start with for the linguist.

3.3 Workflow integration

More broadly than the glossing task, we strove to reduce the gap in the standard annotation workflow. For smoother integration, we created scripts to convert the predicted sentence annotations towards formats widely used in linguistic tools such as FieldWorks Language Explorer (FLEx) (Rogers, 2010), Toolbox[4], and ELAN (Wittenburg et al., 2006). This is to further reduce the friction of using yet another tool.

Below is how our feature-based pipeline can be put into practice in an existing framework for language documentation. Once the time-aligned audio recording is transcribed, with a consistent orthography, the sentences are segmented into words, but also into *morphemes*. The next step is to annotate a small batch of sentences with glosses; usually, the natural order of sentences is followed (e.g., each sentence of a recorded story), ensuring lexical consistency. Then comes the automatic glossing tool. Starting with as many training (i.e., fully annotated) sentences as possible, the model is applied to the rest of the corpus. The idea here is, naturally, to continue the annotation of sentences (possibly from the draft) and to compare the glosses. If specific linguistic phenomena are wrongly predicted systematically, dedicated features can be integrated into the CRF, or more sentences could be given. The latter solution also applies to lexical glosses since our approach depends on the coverage of the dictionary. Finally, the predictions are converted back to the format of the chosen annotation tool.

We note here that our approach does not solve the 'NLP gap' problem yet, as stated in (Gessler, 2022), since it runs separately and not concurrently from existing linguistic tools. It is, however, a step towards an actual integration in annotation software, where we reduced the technical constraints pertaining to the latest glossing models.

4 Experimental results

4.1 Comparison with the SIGMORPHON shared task on interlinear glossing

First, we compare our model with the most recent automatic glossing models to assess its quality in general. The SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023) offered two tracks: the closed one only contained the source sentence with no segmentation information, while the open one notably had the morphological segmentation of the source sentence. The latter setting is closer to ours, where we have actual morphological boundaries of the source sentence.

The seven languages that were studied are diverse both geographically and linguistically (six language families). The released corpora are also of varying size, reflecting different stages of documentation (from 31 training sentences up to several thousand). We focus on six languages to test our model against the other submissions; we do not cover Arapaho, which had the largest corpus by far (40k sentences; 4 times the size of the second largest corpus).

We compare our model (`CRF+dict`) with a simple baseline, `dict`, which assigns the most frequent label seen in the training dataset. This replicates a dictionary-based functionality which is implemented in certain annotation tools (e.g., ELAN (Wittenburg et al., 2006) or FLEx (Rogers, 2010)).

Moreover, we present the results of the baseline model from the Shared Task (BASE_ST), based on a transformer architecture (Ginn, 2023), and the

[4] https://software.sil.org/toolbox/.

best performance reached during the Shared Task (mostly from (Girrbach, 2023); BEST_ST), usually a neural model. Additionally, we report the scores obtained by the state-of-the-art GlossLM model (Ginn et al., 2024), when it was fine-tuned on the corresponding language datasets (GlossLM$_{FT}$).

We evaluate the models according to the two main metrics used in the Shared Task: the accuracies computed at the word or morpheme levels.

	ddo	git	lez	ntu	nyb	usp
dict	65.3	28.1	81.2	81.5	64.4	72.8
CRF+dict	82.2	29.2	85.0	87.6	74.8	75.6
BASE_ST	75.7	16.4	34.5	41.1	84.3	76.6
BEST_ST	85.8	31.5	**85.4**	**89.3**	**88.0**	78.5
GlossLM$_{FT}$	**89.3**	**34.9**	71.3	81.5	87.7	**84.5**
Δ_{BEST}	-7.1	-5.7	-0.4	-1.7	-13.2	-8.9
dict	79.1	51.2	85.8	87.1	72.9	79.5
CRF+dict	89.2	51.8	**88.6**	91.7	82.0	82.0
BASE_ST	85.3	25.3	51.8	49.0	88.7	82.5
BEST_ST	92.0	**52.4**	87.6	**92.8**	**91.4**	84.5
GlossLM$_{FT}$	**92.8**	28.9	74.7	86.0	90.7	**86.4**
Δ_{BEST}	-3.6	-0.6	+1.0	-1.1	-9.4	-4.4

Table 2: Accuracy at the word (top rows) and morpheme (bottom) levels on the test set of the Shared Task. Best scores are in **bold**. Δ_{BEST} indicates the difference between our system and the best performance otherwise.

Table 2 presents the scores for both evaluation levels on the six corpora. First, we see that the glossing task can already be well achieved by a dictionary-based approach, as seen in the high accuracy reached by the dict baseline (except for Gitksan, git). This is due to the regularity in the annotations found in the dataset, especially for lexical morphemes and glosses. These units tend to be consistently annotated in the same way, which is one key assumption for our system.

Still, grammatical morphemes contain more variability with different glosses that could be attributed to the same unit; this is, hence, better captured by CRFs. We see a noticeable improvement of more than 16 points for Tsez (ddo), which is a morphologically rich language.

This approach is also consistently better than the Shared Task baseline (except for Uspanteko, usp). Moreover, despite its simplicity, it remains competitive with the best systems submitted to the shared task. Our model is only a few points behind the best models of the Shared Task, except in Nyangbo (nyb). Indeed, with an average accuracy of 72.4 for words and 80.9 for morphemes, the model would have been ranked fourth among eleven submissions. Compared to the current state-of-the-art GlossLM model, our system performs worse on their in-domain languages (i.e., which have more training data and, hence, were used for pre-training) but leads to notably better prediction on the languages with fewer data, such as Lezgi (lez) or Natugu (ntu).

More generally, we note that our CRF-based approach works better when the training data is smaller (git, lez, and ntu have below 800 training sentences), while more complex models naturally perform better with more sentences (Ginn et al., 2024). Hence, our system could help the early stages of documentation while alleviating some technical constraints.

4.2 Results on Mukrī Kurdish

Table 3 presents the accuracy scores on our Mukrī Kurdish corpus, computed using the same methodology as in the previous section. Given its size, we are in earlier documentation stages, where our system works relatively better (cf. Section 4.1).

	word	morpheme
dict	38.2	53.3
CRF+dict	50.7	64.1

Table 3: Accuracy on Mukrī Kurdish (top: word level, bottom: morpheme level).

We see for Mukrī Kurdish that the glossing performance with our model is still imperfect, actually in between the quality observed for Gitksan and Lezgi in Table 2. However, we notice a significant improvement over a pure dictionary-based approach, which is often used in linguistic annotation workflows.

Moreover, we additionally compare the usual precision, recall, and F-score separately for grammatical and lexical glosses. We notice that, as expected, the use of a CRF model improves the quality of grammatical label prediction (F-score of 48.3 to 66.3) due to their ambiguity. Our two-step pipeline also benefits the lexical tags thanks to a better separation of grammatical and lexical morphemes before replacement.

Among the 500 morphemes in the test set, 53 of them were tagged as UNK. This means that either the lexical morpheme was not seen in the training (in most cases), or the morpheme was wrongfully labelled with 'stem' instead of a grammatical tag.

Process Type	Error Rate (%)
Simple Affixation	9.60
Compounding	18.90
Cliticisation	14.20
Reduplication	37.10
Circumfixation	28.40
Infixation	33.80

Table 4: Error Rates by Morphological Process

Error rates (1 − accuracy) varied significantly across different morphological processes, as shown in Table 4. The lowest error rate was observed for simple affixation (9.60%), suggesting that the model effectively captures regular concatenative morphology even when trained with fairly few examples. For non-concatenative phenomena, however, performance deteriorated sharply, with reduplication showing the highest error rate at 37.10%, followed by infixation (33.80%) and circumfixation (28.40%). These results confirm that sequential models have particular difficulty with morphological operations involving copying, template relations, or internal alternation structures. They are harder to predict and thus require a more complex approach than a simple CRF modelling.

These findings are consistent with theoretical discussions in morphological typology (McCarthy, 1981; McCarthy and Prince, 1999), which distinguish between concatenative and non-concatenative morphology. Reduplication, in particular, involves correspondence constraints between base and copy elements, which are difficult to capture with the current surface-level statistical model alone.

4.3 Qualitative analysis

We discuss the linguistic peculiarities of Mukrī Kurdish and how they are reflected in the test data and predictions. Table 5 displays how ambiguous a given grammatical gloss is. We see that some highly systematic morphemes, such as *ne–*, *–eke*, and *=ewe*, appear consistently and should be easier to learn. In contrast, forms like *ī* have multiple roles (*ezafe*, 3SG, possessive, oblique), making them harder to disambiguate, and hence to predict.

As such, for canonical constructions, the pipeline showed strong performance, correctly identifying core and consistent morphemes such as *ezafe* markers, possessive suffixes, and common definite articles. For example, in the phrase *ser=ī yexdānē* (‘the door of the wardrobe’), the system accurately recognised ‘=ī’ as an *ezafe* or genitive marker linking the possessed noun (*ser*, ‘door [lit. head]’) to its possessor (*yexdānē*, ‘wardrobe’). This is consistent with the typical agglutinative structure found in many Iranian languages, where grammatical relations are expressed by postposed affixes (MacKenzie, 1961; Öpengin and Haig, 2014; Asadpour, 2022).

Error analysis We mainly noticed it struggles with under-represented (or absent) phenomena in the training corpus and ambiguous morphemes. For instance, discourse particles and switch-reference markers were poorly captured, especially in spoken narrative texts where such pragmatic features are prominent. The sentence in Figure 4 is a representative example.

S	[...] degeł	lē–de–de–ā	w
P	[...] with	at–IND–IND–3SG	PTCP
G	[...] with	PVB–IND–give.PRS–3SG	and

Figure 4: Example of wrong analysis. S: segmented source sentence, P: prediction from our system, G: gold glossed sentence.

In this case, the particle ‘*lēdedā*’ (‘lē–de–de–ā’) was wrongly analysed, and the conjunction ‘*w*’ was treated as a clitic rather than a full discourse element. As the gold standard shows, ‘*lēdedā*’ functions as a verb root combined with aspect markers around, while ‘*w*’ functions as a coordinating conjunction. This illustrates one of the limitations of a simple CRF-based model: longer dependencies are not well-captured. Since our features mainly look at the immediate neighbours of a morpheme, it still struggles with polymorphemic words, as in here.

A frequent error we saw concerned the treatment of agreement markers. For instance, the morpheme ‘ī’ was often assigned OBL1 instead of 3SG, which is likely due to overlapping surface forms.

Application to language documentation The results on the test data suggest that, despite the morphological complexity, many patterns in Mukrī Kurdish are consistent enough to be handled by automatic systems. Common and systematic affixes (especially for verbs and nouns) are good candidates for automatic glossing. However, ambiguous and pragmatic elements are likely to require manual review and correction. A tool that pre-annotates glosses based on these regularities can, however,

Label	Example	Description	Position	Consistency
IND (de–)	de–ke, de–łē, de–č–m	Indicative prefix	Verb-initial	High
IRR (bi–)	bi–hēn, bi–nūs, bi–ke	Irrealis prefix	Verb-initial	High
NEG (ne–)	ne–bird, ne–kew, ne–mā	Negation prefix	Verb-initial	High
PVB	heł–de–gir, lē–de–de, ber–de	Preverbal particles	Before verb	Medium
ASP (=ewe)	bird–ī=ewe, ke=ewe, dāte=ewe	Aspectual enclitic	Word-final	High
DEF (–eke)	kitēb–eke, čikołe–eke	Definite marker	Noun-final	High
PL (–ān)	kitēb–ān, žin–ān	Plural marker	After noun	High
OBL (–ī)	bird–ī, č–ī, goř–ī	Oblique case	Noun-final	Medium (ambig.)
EZ (=ī)	čend=ī, birā=ī	*Ezafe*	After noun	Medium (ambig.)
PRSNT	āhā, hā, hā	Presentative	Independent	Low
DISC	āhā, wiłāhī	Discourse markers	Variable	Low

Table 5: Consistency of 10 frequent grammatical labels in the test dataset.

notably reduce the burden on linguists by allowing them to correct rather than annotate from scratch.

This is in line with one of the author's feedback as a fieldworker. Using the CRF-trained model significantly reduced the time spent on routine glossing by pre-labelling frequent grammatical patterns and high-frequency morphemes with reasonable accuracy. This allowed him to focus more on irregular forms, novel constructions, and higher-level linguistic analysis.

4.4 Interpretation of the model

Since our system relies on a CRF, we can interpret the features and patterns that were learnt by the model. For instance, the left part of Table 6 displays the 10 most weighted local properties.

Feature		Transition	
source feature	gloss	gloss_1	$\rightarrow \text{gloss}_2$
morph: m	1SG	EZ	→ REFL
morph: ew	DEM	IND	→ –
morph: emin	1SG	INDF.PRO	→ INDF.PRO
morph: eto	2SG	PVB	→ –
morph: de	IND	=	→ 3SG
morph: t	2SG	VOC	→ RDP
morph: bi	IRR	IMP	→ DISC
morph: nā	NEG	–	→ OBL
morph: ēk	INDF	3SG	→ NEG3
morph: n	3PL	OBL1	→ POST1

Table 6: Left: top 10 features; right: top 10 label transitions learnt by the CRF.

We notice that key morphological patterns in Mukrī Kurdish were correctly identified. For instance, both the independent pronoun 'emin' and its bound form 'm' are associated with the first-person singular gloss. Other frequent and crucial grammatical morphemes are also learnt, such as the negation marker 'nā' or the indefinite suffix 'ēk'. Most of them are consistent annotations with little ambiguity and occur often. These associations are closely aligned with typological descriptions of Kurdish, where agglutination dominates, and each morpheme encodes a single grammatical meaning.

This supports usage-based theories of morphological acquisition (Bybee, 2010), which posit that speakers rely heavily on co-occurrence patterns to disambiguate morphological function. Our results also suggest that statistical models approximate native speakers' intuitions about morpheme function.

Similarly, the model learns label transitions; the most highly weighted ones are in the right part of Table 6. Some of the transitions highlight crucial morphosyntactic patterns. First, the transition from 'EZ' to 'REFL' captures a common construction in Mukrī Kurdish where reflexive pronouns often follow an *ezafe* marker. The model has also correctly identified some pronominal clitics appearing after a clitic boundary, as shown by the strong association between the clitic marker '=' and '3SG' (third-person singular). The transition from 'IND' (indicative) to '–' (morpheme boundary) reflects the morphological structure of Mukrī verbs, where the indicative prefix is typically followed by other verbal morphology (as in Figure 4). These patterns demonstrate that the model has actually captured central morphosyntactic regularities in Mukrī Kurdish, such as clitic placement or verbal morphology.

However, the model occasionally violated these constraints when exposed to less frequent patterns. This suggests that surface statistics, while informa-

tive, may not be sufficient to fully capture more complex morphosyntactic principles.

In short, while our pipeline performs reasonably well on regular morphological patterns represented in the training data, it struggles with rare constructions, phonologically conditioned allomorphy, and morphologically complex phenomena that require more global structural awareness. This is because the CRF relies on local statistical cues, which cannot handle rare, unseen or structurally divergent constructions. While the model does not explicitly learn abstract grammatical rules, it manages to infer recurrent associations between morphemes and their glosses based on distributional patterns present in the training data, which makes it effective for canonical morphology.

5 Recommendations for stakeholders

This article is the result of a collaboration between field linguistics and NLP; as such, we found a few recommendations for all parties involved or supporting language documentation, in line with (Flavelle and Lachler, 2023).

For field linguists, maintaining consistent segmentation and annotation conventions is essential for both humans and NLP models. On this point, following widely used conventions such as the Leipzig Glossing Rules (Lehmann, 2004; Bickel et al., 2008) can also help cross-lingual models, which might have seen the same grammatical glosses in other languages. In this regard, starting with a small but high-quality dataset is enough to start the first automatic gloss pipeline (e.g., the Gitksan corpus in the SIGMORPHON Shared Task had 31 sentences, and we have slightly more than 200 sentences).

For members of the language community, simplified interfaces and localised training materials can enable active participation in validation and annotation. Workshops to build consensus on terminology and validate results help to ensure cultural appropriateness and community ownership of digital resources.

For NLP researchers, the challenge is to improve the robustness of the model and to deal with more complex morphological phenomena while keeping in mind a real-life deployment of the glossing tool. Making the tools more user-friendly is also appreciated; specialised error analysis tools and visualisations would help to diagnose wrong predictions easily. Finally, a better evaluation protocol should be used to account for the error gravity; in the end, we aim at a system that helps rather than confuses the annotators.

6 Related Work

Interlinear gloss generation, in collaboration between linguistics and NLP for language documentation, has initially been explored with feature-based taggers. (Baldridge and Palmer, 2009) and (Palmer et al., 2009) both discuss the relevance and efficiency of active learning in such a context. They notably found that the benefit of better sampling techniques depends on the expertise of the annotators. (Samardžić et al., 2015) also applied a two-step pipeline with a tagger for grammatical glosses and a lexicon for the lexical glosses. Their experiments were, however, based on a much larger corpus for a better-documented language.

Moeller and Hulden (2018) show that CRFs are a reliable approach to predict *grammatical* glosses compared to a neural model for a corpus with 3,000 annotated words. Using the same methodology, Barriga Martínez et al. (2021) also find that CRFs outperform RNNs and biLSTMs on their corpus. Then, McMillan-Major (2020) proposes a pipeline combining two CRFs, one to predict from the source sentence and another one from the translation, an underexploited resource so far. All these methods are closely related to our methodology because CRFs are reliable in capturing local dependencies, especially in low-resource settings. However, due to the number of potential labels, lexical glosses cannot be predicted with CRFs alone.

This is one reason behind the consideration of neural models for glossing. Zhao et al. (2020) extend the methodology of (McMillan-Major, 2020) by considering both the source and translated sentences as inputs to a multi-source neural model (based on a transformer architecture; Vaswani et al., 2017).

Finally, a major milestone on the topic is the SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023). Among the two possible tracks, the open one provided the morpheme-level segmentation of the source sentence. In this category, which is an easier task due to the additional information, the best performing model was (Girrbach, 2023), which trained a hard attention model. Two other submissions were also neural and based on transformers (Cross et al., 2023; He et al., 2023). Okabe and Yvon (2023b) have also compared their

feature-based systems against a simple CRF-based baseline; however, the former was not as accessible and convenient as our system, while the latter model was not released. The state-of-the-art for the task is currently achieved by the GlossLM model (Ginn et al., 2024), which also relies on the transformer architecture.

7 Conclusion

We have deployed an automatic glossing pipeline on a fieldwork corpus in Mukrī Kurdish, a Central Kurdish variety, to assess not only how it performs but also how usable such NLP tools are in practice. We have seen that our CRF-based system improved the prediction quality compared to the currently implemented full dictionary-based approach, which further reduces manual workload. It notably managed to learn the most frequent patterns while struggling with rarer phenomena and annotation, as expected. This is, however, not a major issue since any model output remains an annotation draft: they need to be corrected and controlled eventually. In our case, the system lowered the manual annotation effort noticeably, with a fairly robust reliability for repetitive annotations.

Even though our feature-based pipeline may not match the quality of state-of-the-art neural approaches (lagging by 3 points in accuracy on average in Table 2), it offers a more interpretable and adaptable alternative that is well-suited to early-stage documentation projects, such as for Mukrī Kurdish. We believe these characteristics outweigh the benefits of marginal gains obtained with more advanced models.

Finally, we are releasing the glossing pipeline under an open-source license to foster its use by both field linguists and NLP practitioners. We strove to provide a simple tool that can work with the usual infrastructure at hand in language documentation.

We stress again that this work, at the intersection of computational and documentary linguistics, aimed to bridge the gap between the vastly different technical environments of both fields. We also tried to lower the technical barrier by providing scripts to convert the annotations towards popular formats used in language documentation.

Our future work includes integrating the model into an actual annotation software so that it can be used even more easily. Moreover, we will also explore how performance can be improved by adapting known linguistic rules in the feature set, as some linguists already use rule-based processing to some extent.

Limitations

From the NLP perspective, the proposed model is not particularly novel, as similar models relying on CRFs were considered as a baseline for experiments. It does not reach a state-of-the-art performance either, given its simplicity. The model is, however, released not only to provide a fairly competitive yet simple baseline for future works in NLP but also to foster its use among field linguists. We believe, indeed, that the current pipeline can be integrated into actual annotation workflows, possibly after further simplifying user interaction with the model. Hence, our system choice is the result of a compromise between prediction quality and technical complexity.

From the linguistic side, some non-negligible errors remain in Mukrī Kurdish, which shows that the model cannot handle complex morphological patterns yet. For this article, we tried to release a model which could also be applied to other languages directly, i.e., without language-specific features. Thanks to the flexibility allowed by the features, the system can be better tailored to any language which will be studied.

Acknowledgments

We thank the anonymous reviewers for their comments. We are deeply grateful to all who have contributed their time and knowledge during Asadpour's fieldwork in the Mukrīyān region, which began in 2003 and has continued over the years. This research would not have been possible without their trust and generosity. Parts of the work related to the preparation and writing of this paper have received funding from the European Research Council (ERC) under grant agreement No. 101113091 – Data4ML, an ERC Proof of Concept Grant, supporting the contributions of Shu Okabe and Alexander Fraser. Asadpour's fieldwork and participation were conducted independently of this funding.

References

Hiwa Asadpour. 2021. Cross-dialectal diversity in Mukrī Kurdish I: Phonological and phonetic variation. *Journal of Linguistic Geography*, 9(1):1–12.

Hiwa Asadpour. 2022. *Typologizing word order variation in Northwestern Iran*. Ph.D. thesis, Goethe University Frankfurt, Frankfurt, Germany.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Balthazar Bickel, Bernard. Comrie, and Martin Haspelmath. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Leipzig: Max Planck Institute for Evolutionary Anthropology, Department of Linguistics. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Caren Brinckmann. 2008. Transcription bottleneck of speech corpus exploitation.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–229, Toronto, Canada. Association for Computational Linguistics.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Ginn. 2023. Sigmorphon 2023 shared task of interlinear glossing: Baseline model. *Preprint*, arXiv:2303.14234.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 151–165, Toronto, Canada. Association for Computational Linguistics.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. *STUF - Language Typology and Universals*, 55(1):3–14.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Christian Lehmann. 2004. Interlinear morphemic glossing. In *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung.*, volume 17 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 1834–1857. Berlin & New York: W. de Gruyter.

David Neil MacKenzie. 1961. *Kurdish Dialect Studies*. Oxford University Press, London.

John J. McCarthy. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12(3):373–418.

John J. McCarthy and Alan S. Prince. 1999. *Faithfulness and identity in Prosodic Morphology*, page 218–309. Cambridge University Press.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*,

pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shu Okabe and François Yvon. 2023a. LISN @ SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 202–208, Toronto, Canada. Association for Computational Linguistics.

Shu Okabe and François Yvon. 2023b. Towards multilingual interlinear morphological glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Ergin Öpengin and Geoffrey Haig. 2014. Regional variation in kurmanji: A preliminary classification of dialects. *Kurdish Studies*, 2(2):143–176.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. Association for Computational Linguistics.

Chris Rogers. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. In *Language Documentation & Conservation 4*, pages 78–84.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A CRF features

Table 7 presents the features for the following sentence ('*out of the fear of the man*') at the fourth position[5] (first '**ī**'):

1	2	3	**4**	5	6	7
de	tirs	=	**ī**	kābrā	–	ī.

In general, we check the same local properties (source entity itself and its length) for the current (0), previous (-1), and next (+1) positions. Depending on the presence of a morpheme boundary, we also check the 'actual' previous morpheme (-2) to account for morpheme dependencies inside polymorphemic words.

position	feature	example for 'ī'
0	morpheme	ī
0	length	1
0	morpheme boundary?	False
-1	morpheme	=
-1	length	1
-1	morpheme boundary?	True
-2	morpheme	tirs
-2	length	4
+1	morpheme	kābrā
+1	length	5

Table 7: List of the computed features for a given entity. Position indicates the relative position compared to the entity (0: current position, -1: previous position, and +1: next position).

[5]We count source entities: both actual morphemes and morpheme boundaries.

Field to Model: Pairing Community Data Collection with Scalable NLP through the LiFE Suite

R Karthick Narayanan[1], Siddharth Singh[1,2], Saurabh Singh[2], Aryan Mathur[2], Ritesh Kumar[1,2], Shyam Ratan[2], Bornini Lahiri[1,3], Benu Pareek [1,2], Neerav Mathur[2], Amalesh Gope[1,4], Meiraba Takhellambam[1,5], Yogesh Dawer[2],

[1]Council for Diversity and Innovation, [2]Unreal Tece LLP,
[3]Indian Institute of Technology-Kharagpur, [4]Tezpur University, [5]Manipur University
Correspondence: riteshkrjnu@gmail.com

Abstract

We present LiFE Suite as a "Field-to-Model" pipeline, designed to bridge community-centred data collection with scalable language model development. This paper describes the various tools integrated into the LiFE Suite that make this unified pipeline possible. Atekho, a mobile-first data collection platform, is designed to empower communities to assert their rights over their data. MATra-Lab, a web-based data processing and annotation tool, supports the management of field data and the creation of NLP-ready datasets with support from existing state-of-the-art NLP models. LiFE Model Studio, built on top of HuggingFace AutoTrain, offers a no-code solution for building scalable language models using the field data. This end-to-end integration ensures that every dataset collected in the field retains its linguistic, cultural, and metadata context, all the way through to deployable AI models and archive-ready datasets.

1 Introduction

Mobilising language documentation resources to produce language technologies for low-resource and Indigenous languages faces two significant challenges:

1. the lack of accessible, community-friendly data collection tools, and
2. fragmented workflows that separate field linguistics from computational modelling.

Despite advancements in crowdsourcing, linguistic data collection remains predominantly expert-driven. The tools available to Indigenous language speakers are often either complex proprietary systems, such as Karya[1], or basic audio recorder apps that lack essential features such as prompt integration, multilingual support, and metadata capture, all of which are critical for systematic language documentation. Similarly, the tools used by field linguists and computational linguists rarely support direct interoperability. Field linguistics tools seldom leverage the benefits of automation and machine learning that NLP technologies can offer, while NLP tools often struggle to process the rich, multi-layered annotations typical of language documentation corpora. We introduce LiFE Suite, an integrated pipeline that enables communities and researchers to collect, process, and model language data without requiring programming skills or specialised infrastructure. We describe how the suite supports multimodal, multilingual, and metadata-rich workflows that empower both field linguists and NLP practitioners to build language technologies from real-world field linguistic data.

2 Review of Existing Tools

A variety of tools have been developed to support field linguists, community language workers, and NLP practitioners in data collection, management, annotation, and lexicon creation. However, these tools tend to be fragmented, often serving either field linguistics or NLP, but rarely both. Below, we review commonly used tools in these domains and highlight the gaps that motivate the design of LiFE Suite.

2.1 Field Linguistics Tools

Tools primarily used by field linguists or community members for speech and multimodal data collection, management, and lexicon creation include:

1. **Toolbox (formerly Shoebox)**[2]: One of the earliest linguistic tools developed by SIL International, designed for text data entry and dictionary creation (Robinson et al., 2007).

[1] https://www.karya.in

[2] https://software.sil.org/shoebox, https://software.sil.org/toolbox

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 76–84
August 1, 2025 ©2025 Association for Computational Linguistics

2. **FieldWorks Language Explorer (FLEx)**[3]: A widely used SIL tool for managing linguistic and cultural data, including lexicon development and interlinear glossing. LiFE Suite is designed to interoperate with FLEx, supporting the import of LIFT XML data produced by FLEx (Butler and Volkinburg, 2007).

3. **LexiquePro**[4]: Software for creating and formatting lexicon databases, mainly focused on dictionary publication and sharing, with limited editing capabilities (Guérin and Lacrampe, 2007).

4. **WeSay**[5]: Designed to help non-linguists and native speakers build dictionaries of their own languages. It is based on SIL's Semantic Domain and Rapid Word Elicitation methods, promoting community-led lexicon development (Perlin, 2012).

5. **Woefzela**[6]: A smartphone-based tool for offline data collection, supporting multiple sessions and metadata capture. It has been successfully deployed in South Africa for quality-controlled data collection (Vries et al., 2014).

6. **SayMore**[7]: A tool for organising multimedia recordings and their metadata. It also supports basic transcription and translation workflows (Moeller, 2014).

7. **Living Dictionaries**[8] help communities build and manage their own word collections. People can add words, meanings, sounds, pictures, and videos. They can search, filter, and organize entries by topics. The tool works offline and allows data to be shared or imported using common file formats like CSV, PDF, and JSON (Daigneault and Anderson, 2023).

8. **Aikuma and LIG-Aikuma**[9] are mobile apps designed to support speech data collection for under-resourced and endangered languages. Originally developed as Aikuma and later extended as LIG-Aikuma, these apps offer features such as audio recording, respeaking for clarity, oral translation, and elicitation using prompts. LIG-Aikuma also supports metadata capture, geolocation tagging, and data export compatible with ELAN. While LIG-Aikuma remains available, its development has slowed since 2018. A more recent adaptation, Williaikuma, offers updated features for sentence-level elicitation and Praat integration, demonstrating continued interest in mobile tools for linguistic fieldwork (Bird et al., 2014; Gauthier et al., 2016).

While these tools have advanced the practice of field linguistics, they suffer from several limitations that hinder their broader adoption and integration into computational workflows. Most of these tools are standalone desktop or mobile applications, often lacking compatibility with Linux operating systems and restricting usage to Windows or Mac environments. Users are typically required to switch between multiple specialised tools for different tasks, such as ELAN for video transcription, Audacity or Praat for audio segmentation, and FLEx for lexicon management, each with its own steep learning curve. Additionally, data produced by these tools is often stored in non-standard or tool-specific formats, making interoperability with NLP systems difficult without additional processing or programming expertise. Finally, data sharing in reusable, open formats remains cumbersome, limiting long-term accessibility and cross-tool usability.

2.2 NLP Annotation Tools

In contrast, NLP practitioners use a different set of tools for data annotation and management, including:

1. Label Studio[10]: An open-source, web-based data labelling platform supporting audio, text, image, video, and time-series annotation. It allows export to multiple ML-ready formats, making it popular for preparing training data (Tkachenko et al., 2020-2022).

2. Shoonya[11]: An open-source platform focused on enhancing digital content for India's underrepresented languages, supporting large-scale annotation for machine translation and other language technologies.

[3] https://software.sil.org/fieldworks
[4] https://software.sil.org/lexiquepro
[5] https://software.sil.org/wesay/
[6] https://sites.google.com/site/woefzela/
[7] https://software.sil.org/saymore, https://github.com/sillsdev/saymore
[8] https://livingdictionaries.app
[9] https://lig-aikuma.imag.fr
[10] https://labelstud.io/; https://github.com/heartexlabs/label-studio
[11] https://ai4bharat.iitm.ac.in/shoonya, https://github.com/AI4Bharat/Shoonya

3. BRAT[12], doccano[13], and INCEpTION[14]: Popular open-source tools for text annotation at the token, span, and document levels. These tools offer features for text classification, sequence labelling, and sequence-to-sequence applications (Stenetorp et al., 2012; Nakayama et al., 2018; Klie et al., 2018).

However, these NLP tools generally do not support field data collection or integrate with linguistic data management workflows. They are designed for annotation and model preparation, often assuming pre-processed, clean data rather than raw, community-collected, multimodal datasets.

2.3 The Need for an Integrated Pipeline

As highlighted, field linguistics tools and NLP tools often operate in isolation, each addressing specific stages of the data lifecycle but failing to offer a cohesive, integrated experience. LiFE Suite seeks to bridge this gap by providing a unified, no-code pipeline that supports the full workflow—from community-led data collection with structured metadata, to linguistic data management and annotation, and ultimately to NLP model training and deployment. Designed to serve both field linguists and NLP practitioners (Figure 1), LiFE Suite reduces the technical barriers that currently separate these communities, enabling them to collaboratively develop language technologies for low-resource and Indigenous languages.

3 LiFE

LiFE Suite [15] is an open-source, AI-powered platform developed by UnReaL-TecE [16], a venture led by linguists to enable seamless language data collection, management, processing, and analysis. All the components of the suite are developed as web apps with HTML, CSS and JavaScript at the frontend, Python and FastAPI (for serving different APIs) at the backend, MongoDB as the backend database and IndexedDB as the frontend database. The apps are served using Flask.

The suite integrates state-of-the-art technologies, including Large Language Models (LLMs), Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), Machine Translation (MT), and advanced text processing models, providing an efficient and scalable solution for linguistic data workflows. LiFE Suite brings together three key components to support this pipeline. Atekho, a mobile-first progressive web application (PWA) designed for community-centred, offline and online multimodal data collection with integrated metadata scaffolding; MATra-Lab, a web-based platform for organizing, segmenting, transcribing, translating, and annotating linguistic datasets; and LiFE Model Studio, a no-code model-building environment built on top of HuggingFace AutoTrain, enabling the training and deployment of speech and multimodal models. Together, these components form an end-to-end system that bridges community-led data collection with scalable NLP model development and long-term archival, making linguistic technologies more accessible and sustainable for low-resource and Indigenous languages.

3.1 Atekho

Named after the Great Andamanese word for “language", Atekho (Figure 2) is a mobile-first, progressive web application for data collection, designed to empower communities and researchers to co-create living linguistic and cultural archives. By supporting multimodal data capture, including audio, video, image, and text, Atekho enables the documentation of linguistic, cultural, environmental, and oral traditions in both spontaneous and staged settings.

[12] http://brat.nlplab.org, https://github.com/nlplab/brat

[13] https://doccano.herokuapp.com, https://github.com/doccano/doccano

[14] https://inception-project.github.io, https://github.com/inception-project/inception

[15] https://github.com/unrealtecellp/life

[16] UnReaL-TecE is an organisation that is founded to develop and maintain this platform. Unlike a large number of other platforms, which could not be maintained because of various practical reasons, we expect this organisation to take care of long-term maintenance of the app and ensure that it remains available in the future.

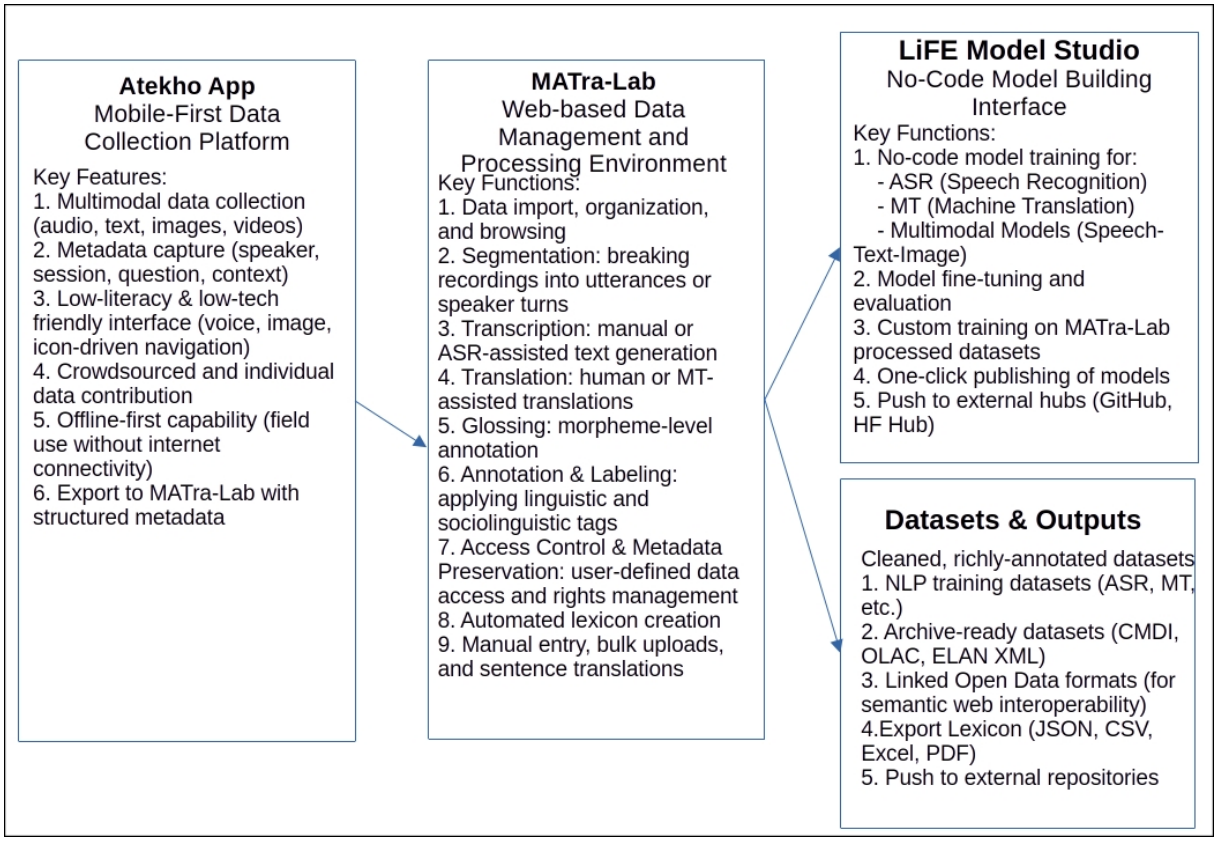

Figure 1: LiFE Pipeline

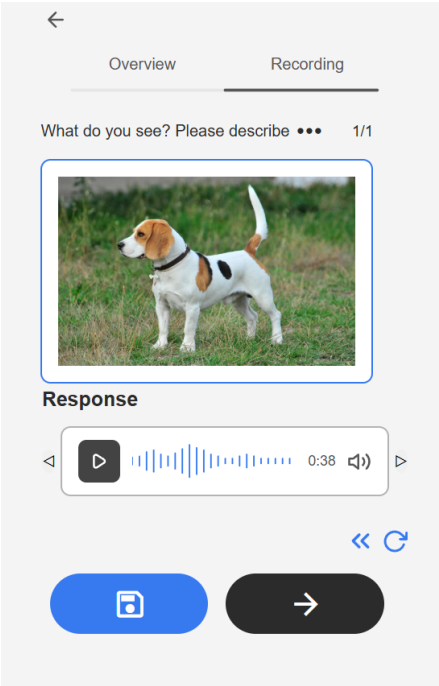

Figure 2: Atekho Interface

Atekho is built with accessibility and inclusion at its core. Its voice and icon-based user interface is designed for users with limited literacy or technology experience, making it particularly suited for community-centered projects in remote or low-resource contexts. The application operates in offline-first mode, ensuring functionality in rural areas with limited or no internet connectivity. In addition to capturing content, Atekho supports collaborative metadata scaffolding, allowing users to tag recordings with speaker details, contextual information, and community-generated annotations. Its workflows are customizable, enabling projects to define their own data structures and inventory formats. This flexibility makes Atekho adaptable to a wide range of documentation initiatives, including sociolinguistic surveys, oral literature preservation, and environmental knowledge documentation. As part of the LiFE Suite, Atekho seamlessly integrates with MATra-Lab, allowing collected data and metadata to flow directly into more advanced processing pipelines. Once synchronized with MATra-Lab, recordings can be segmented, transcribed, translated, glossed, and annotated using an AI-in-the-loop mechanism. This interoperability positions Atekho not just as a data collection tool, but as the starting point of an end-to-end "Field to Model" pipeline, bridging community-driven documentation with scalable NLP model

development.

By placing ownership and control in the hands of the communities whose heritage it seeks to preserve, Atekho supports the creation of living archives that are ethically grounded, accessible, and sustainable.

3.2 MATra-Lab: Web-Based Linguistic Data Management and Processing

MATra-Lab (Figure 3) is a web-based platform designed to support the management, processing, and annotation of multilingual and multimodal datasets, with a particular focus on the linguistic diversity of Indian languages. It provides researchers across subfields, such as field linguistics, sociolinguistics, and computational linguistics, with an integrated environment for processing audio, video, text, and image data. By combining multimodal data processing, AI-powered tools, collaborative management, and an intuitive interface, MATra-Lab offers an end-to-end environment to produce scalable, reproducible, and NLP-ready linguistic resources.

Figure 3: Transcription in Matra Lab

3.2.1 Data Ingestion and Management

In addition to providing automated ingestion of data collected using Atekho, MATra-Lab supports the direct upload of field-collected datasets along with rich metadata, including participant information and item-level metadata. This metadata enables users to sort, filter, and organise data for efficient navigation and management of large and heterogeneous collections (Figure 4).

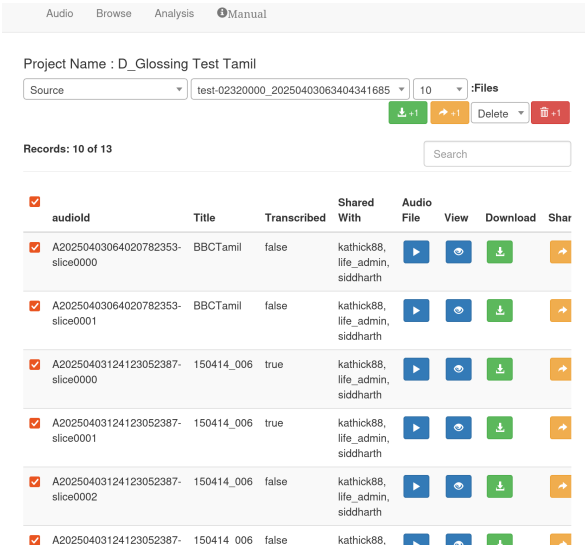

Figure 4: Browse and Filter in Matra Lab

3.2.2 Supported Workflows and Tasks

MATra-Lab supports multilayered annotation and processing workflows that linguists and computational linguists may use. It allows users to apply document-level (for all kinds of multimodal documents) and span-level (for audio, video and text documents) labels for tasks such as morphological analysis, syntactic tagging, discourse annotation, and semantic labeling. It currently allows for the following kinds of tasks -

1. Time-aligned audio and video transcription in multiple scripts at any level from individual phones to complete discourse. It also allows for speaker diarisation, mapping audio files or parts of audio files to prompts and anonymising parts of the audio.

2. Translation of audio, video, text and images.

3. Annotation of audio and video chunks with custom labels.

4. Annotation of text at both document and span level with custom labels.

5. Interlinear glossing of audio, video and text documents.

6. OCR and labelling of images.

3.2.3 Building Lexicons

In addition to providing data processing support, MATra-Lab also includes a Lexicon Module for building and managing multilingual glossaries and dictionaries. It supports:

1. Automated extraction of lexical items from existing annotated data,
2. Manual entry of lexicon items,
3. Bulk uploads from external sources, and
4. Sentence-aligned translations to support context-rich dictionary development.

Users can browse, edit, and export lexicons in multiple formats, including JSON, RDF, CSV, Excel, and PDF, facilitating integration with other linguistic tools or dissemination to wider audiences. The module also supports collaborative dictionary development with fine-grained access control, allowing teams to manage user permissions for viewing, editing, and exporting lexicon data.

This added functionality extends the utility of MATra-Lab beyond corpus management, allowing dictionary making, terminology development, and community-led lexical documentation within the same unified workflow.

3.2.4 AI-in-the-loop

A core feature of MATra-Lab is its graphical user interface (GUI), which allows users to apply pretrained models for a range of natural language processing (NLP) tasks without requiring coding expertise. The tasks where AI models currently provide support are the following:

1. Transcription (both in IPA and native scripts for supported languages), speaker diarisation, translation, and glossing for audio data using Automatic Speech Recognition (ASR), Machine Translation (MT) and other relevant models,
2. Digitisation of scanned documents and images using Optical Character Recognition (OCR) tools,
3. Advanced text processing using Large Language Models (LLMs) and multimodal models.

All of these automation facilities are made available by integrating models and APIs from different sources.

1. **HuggingFace Hub**[17]: integration with publicly available models on HuggingFace Hub is available out-of-the-box.
2. **Bhashini API**: Bhashini APIs are provided by the Ministry of Electronics and Information Technology, Government of India. These provide access to state-of-the-art open-source models supporting different kinds of tasks in Indian languages viz ASR, speaker diarisation, transliteration, language identification, etc.
3. **LiFE Model Studio**: The models that are trained by the users using the LiFE Model Studio can be used in MATra Lab for automating the tasks.
4. **Stanza**[18]: Stanza models are integrated to provide automatic interlinear glossing and morphosyntactic information including part-of-speech categories, morphological features and dependency relations for the supported languages (Qi et al., 2020).
5. **Language agnostic models**: Some models for language-agnostic tasks such as voice activity detection (viz Silero VAD (Team, 2024) and PyAnnote (Bredin, 2023; Plaquet and Bredin, 2023)), speaker diarisation, (viz. PyAnnote) and universal phoemiser (viz. Allosaurus (Li et al., 2020)) are also integrated into the app.
6. **In-house Models**: Limited support for tasks such as interlinear glossing, sentiment analysis, aggression level, etc in some languages are provided through our in-house rule-based and machine learning-based models.

3.2.5 Data Export and Sharing

The platform allows users to download their dataset in multiple structured and semi-structured formats viz. JSON, CSV, XLSX, Markdown, TextGrid, CHAT, etc, for further processing and use with other apps and libraries.

The platform also offers collaborative workspace functionality, enabling file-level sharing with fine-grained access control. Users can define permissions for:

1. Full access (edit and download),
2. Restricted access (online viewing and annotation without download), or
3. Partial access (limited operational permissions).

[17] https://huggingface.co/docs/hub

[18] https://stanfordnlp.github.io/stanza

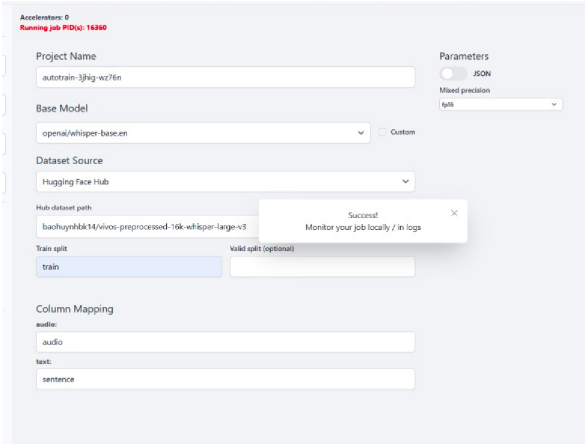

Figure 5: LiFE Model Studio

This collaborative architecture promotes team-based data management while ensuring data security and control.

3.3 LiFE Model Studio

LiFE Model Studio (Figure 5) is a no-code model training interface built on HuggingFace AutoTrain. It allows users to train new models or fine-tune existing ones for a variety of NLP tasks, using the datasets they have created and processed in MATra-Lab. While model building is optional, users who choose to extend their workflow within the LiFE Suite can derive datasets directly from their MATra-Lab projects and use them in LiFE Model Studio to build task-specific models, such as speech recognition, machine translation, or multimodal processing models.

Trained models are immediately available within the platform for further use in data processing workflows, creating a feedback loop that improves annotation, transcription, and translation over time. Additionally, these models can be exported to external repositories, such as GitHub or HuggingFace Hub, for wider public access and reuse. We have added the following functionalities to the existing AutoTrain interface -

1. **MATra Lab Integration**: We have added support for directly importing specific kinds of data from MATra Lab dataset into the interface and use that for fine-tuning the required models.

2. **Audio-based Tasks**: We have added support for audio-based tasks such as Automatic Speech Recognition, speaker diarisation and voice activity detection.

3. **Additional Tasks**: We are in the process of integrating support for additional libraries (such as scikit-learn) and tasks to enable no-code training for additional tasks (such as interlinear glossing).

LiFE Model Studio thus completes the Field-to-Model pipeline, providing users with a scalable, no-code solution for bringing community-collected data all the way to deployable language technologies.

4 Case Studies: Demonstrating the Field-to-Model Pipeline

To demonstrate the effectiveness of the Life Field-to-Model pipeline, we present two ongoing case studies that apply the pipeline in real-world, community-centred language technology projects. These case studies illustrate how the pipeline enables end-to-end data collection, management, processing, and model development in two distinct linguistic contexts.

4.1 SpeeD-TB

The first case study, Speed-TB (Kumar et al., 2023), focuses on six Tibeto-Burman languages spoken in Northeast India—Bodo, Meetei, Chokri, Kokborok, Nyishi, and Toto—and is funded by the Bhashini initiative of the Government of India. Using the Life Suite, data is collected through structured questionnaires, stimulus-based narration, role-play, and spontaneous speech, with community members actively participating in data contribution and validation. The collected data is processed in MATra-Lab and used in LiFE Model Studio to build baseline speech recognition models. The project explores fine-tuning/training models such as conformer-multilingual-asr by AI4Bharat[19], Whisper[20](Radford et al., 2022), wav2vec 2.0(Baevski et al., 2020), and NVIDIA NeMo[21](NVIDIA, 2025).

4.2 Irula Language

The second case study focuses on Irula, a Dravidian language spoken in Tamil Nadu, India. In collaboration with the Keystone Foundation, a community-based organisation, the project builds on existing

[19] https://dibd-bhashini.gitbook.io/bhashini-apis/available-models-for-usage

[20] Whisper: https://github.com/openai/whisper; Paper: https://arxiv.org/abs/2212.04356

[21] NVIDIA NeMo: https://developer.nvidia.com/nvidia-nemo

resources from a community radio station to collect and process Irula speech data. The team experiments with fine-tuning the conformer-multilingual-dravidian model by AI4Bharat and other multilingual models to develop a dedicated Irula ASR system.

5 Conclusion

In this paper, we have presented a new workflow for building language technologies for underresourced languages using primary data collected from the field. This workflow is enabled through the LiFE Suite, an open-source AI-powered platform. We give details of the suite and how it enables the operationalisation of the Field-to-Model workflow. We also present two case studies where we use the workflow and the suite for building language technologies.

In both the case studies, the community remains at the center—not only as data contributors but as co-creators and validators of the resulting language technologies. These case studies serve as proof-of-concept implementations, demonstrating that the Field-to-Model pipeline is viable, scalable, and capable of supporting community-driven speech technology development for underrepresented languages.

While these case studies demonstrate the practical value and scalability of the Field-to-Model pipeline, our experience also highlights several challenges and constraints that must be addressed to make the workflow more inclusive and widely adoptable.

Limitations

While the LiFE Suite offers a comprehensive, no-code pipeline for community-centered language documentation and NLP model development, several limitations persist. Firstly, although Atekho is designed for offline use, both MATra-Lab and LiFE Model Studio require stable internet connectivity and access to web-based interfaces, which may pose challenges in remote or resource-constrained environments. Secondly, MATra-Lab's reliance on pre-trained models from platforms like HuggingFace means that its performance is contingent on the availability and quality of existing models, which may not adequately represent all low-resource or Indigenous languages. Regarding LiFE Model Studio, while it provides an accessible interface for model fine-tuning, it currently does not support training models from scratch; users can only fine-tune existing pre-trained models . Additionally, despite its no-code design, users may still require a foundational understanding of NLP concepts to effectively navigate model selection and fine-tuning processes. Lastly, the computational demands of model training and inference necessitate access to GPUs, which may not be readily available to all users, potentially limiting the suite's accessibility and scalability.

Acknowledgments

We would like to thank Mission Bhashini, Ministry of Electronics and Information Technology (MEITY), Govt of India, for supporting the SpeeD-TB project and the development of the LiFE suite. We would also like to express our heartfelt thanks to all the community members of Irula, Toto, Chokri, Kok Borok, Nyishi, Bodo and Meitei, who contributed immensely in the two case studies and by providing valuable feedback on the LiFE suite.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Lynnika Butler and Heather Volkinburg. 2007. Review of fieldworks language explorer (flex). *Language Documentation and Conservation*, 1.

Anna Luisa Daigneault and Gregory D. S. Anderson. 2023. Living dictionaries: A platform for indigenous and under-resourced languages. *Dictionaries: Journal of the Dictionary Society of North America*, 44(02):57–74.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. Ligaikuma: A mobile app to collect parallel speech for under-resourced language studies. pages 381–382.

Valérie Guérin and Sébastien Lacrampe. 2007. Lexique pro. *Language Documentation and Conservation*, 1(2):293 – 300.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Ritesh Kumar, Meiraba Takhellambam, Bornini Lahiri, Amalesh Gope, Shyam Ratan, Neerav Mathur, and Siddharth Singh. 2023. Collecting speech data for endangered and under-resourced indian languages. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL 2023)*, pages 31–38.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Sarah Ruth Moeller. 2014. Saymore, a tool for language documentation productivity. 08:66–74.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

NVIDIA. 2025. Nvidia nemo: Open-source toolkit for conversational ai. `https://developer.nvidia.com/nvidia-nemo`. Accessed: 2025-05-11.

Ross Perlin. 2012. Wesay, a tool for collaborating on dictionaries with non-linguists. *Language Documentation & Conservation*, 6:181 – 186.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with toolbox and the natural language toolkit. *Language Documentation and Conservation*, 1.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Nic Vries, Marelie Davel, Jaco Badenhorst, Willem Basson, Etienne Barnard, and Alta de Waal. 2014. A smartphone-based asr data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.

Low-resource Buryat-Russian neural machine translation

Dari Baturova[1], Sarana Abidueva[2], Ivan Bondarenko[1], Dmitrii Lichko[3]

[1]Novosibirsk State University, Russia
[2]Saint Petersburg State University, Russia
[3]NUST MISiS, Russia

baturova.dari@gmail.com, abidueva.sarana02@gmail.com
i.bondarenko@g.nsu.ru, lichko2002@mail.ru

Abstract

This paper presents a study on the development of a neural machine translation (NMT) system for the Russian-Buryat language pair, focusing on addressing the challenges of low-resource translation. We also present a parallel corpus, constructed by processing existing texts and organizing the translation process, supplemented by data augmentation techniques to enhance model training.

We managed to achieve BLEU score of 20 and 35 for translation to Buryat and Russian respectively. Native speakers have evaluated the translations as acceptable.

Future directions include expanding and cleaning the dataset, improving model training techniques, and exploring dialectal variations within the Buryat language.

1 Introduction

The Buryat language is the national language of the Buryat people and is spoken in Russia, Mongolia, and China. It belongs to the Mongolic language group. However, due to its geographic distribution, the Buryat language has evolved differently in each country, influenced by the dominant languages and cultural contexts of the respective regions. In this article, we focus specifically on the variety of Buryat spoken in Russia, which is identified by the ISO code bxr.

Although Buryat is an official language of the Republic of Buryatia in the Russian Federation, the overwhelming majority of intellectual activity there is carried out in Russian and the number of young people speaking Buryat is rapidly declining. UNESCO included the Buryat language in the "Atlas of the world's languages in danger" (UNESCO, 2010).

As part and consequence of this problem, Buryat is underrepresented in computational linguistics, has limited available corpora and linguistic resources. That creates the main challenge of conducting a machine translation system for Buryat.

The Buryat language has undergone several transitions in its writing system throughout its history. Since 1939, it has been written using the Cyrillic alphabet. Before then, from 1930 to 1939, it utilized a Latin-based alphabet. Going further back, since the 18th century, the traditional Mongolian script served as the writing system for Buryat. This adds another complexity to the construction of machine translation, as some of the available literature was in Latin.

Furthermore, the preservation and revitalization of endangered languages through modern technologies have become critical goals in both linguistic research and cultural heritage preservation. In this context, creating robust machine translation systems not only aids communication but also contributes to the documentation and promotion of underrepresented languages.

2 Related Work

Low-resource machine translation has been an active area of research in recent years, driven by the need to support underrepresented languages.

Several large-scale initiatives have extended machine translation capabilities to hundreds of languages, including low-resource ones. Notable examples include the work of Bapna et al. (2022), the No Language Left Behind (NLLB) project (NLLB Team et al., 2022), and the efforts described in Fan et al. (2021). These projects demonstrate the potential of multilingual models to address challenges in low-resource settings.

Other studies focus on developing machine translation systems for individual low-resource languages by fine-tuning multilingual models. Examples include work on Erzya (Dale, 2022), Ngambay (Sakayo et al., 2023), Zarma (Keita et al., 2024), Karachay-Balkar (Berberov et al., 2024), and Aro-

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 85–93
August 1, 2025 ©2025 Association for Computational Linguistics

manian (Jerpelea et al., 2025). These efforts often rely on community-driven datasets and highlight the importance of adapting models to specific linguistic and cultural contexts.

In June 2024, Google Translate[1] added support for the Buryat language. According to company representatives, this was made possible by leveraging their language model PaLM 2. However, the dataset used for training has not been released as open source.

Research on Buryat in the context of natural language processing includes work by Konovalov and Tumunbayarova (2018), who explored word vector representations for Buryat using models such as pointwise mutual information (Church and Hanks, 1990), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013), trained on data from the Buryat Wikipedia[2]. More recently, Shliazhko et al. (2024) introduced a multilingual variant of the GPT-3 large language model, trained on 61 languages, including Buryat, using corpora such as Wikipedia and the C4 dataset (Raffel et al., 2020).

3 Parallel Corpus Construction

Our parallel corpus was constructed through three main approaches: manual translation of Russian texts by hired translators, collaboration with local organizations in the Republic of Buryatia that have bilingual textual resources and web-based data collection.

The dataset is available as an open source[3].

3.1 Manual translation of texts

In order to create a quality Russian–Buryat parallel corpus from zero, we crafted a semi-automated system for the selection and preparation of the Russian source texts. The Taiga Corpus news segment[4] contains articles from various online media like Lenta.ru, Interfax, Komsomolskaya Pravda, N+1, Fontanka.ru and Arzamas. For our project, we utilized the text corpus that was news genre to ensure maximal variety, along with clearness of meaning, and a stable correspondence between the source and the target sentences.

To improve the ease of translation, we programmed fragments consisting of several sentences (up to five) into coherent passages instead of treating every sentence as isolated. Long paragraphs were split into smaller chunks, while ensuring they were semantically cohesive.

Text that did not suit requirements was discarded and the remaining content was processed using text embedding model `aiforever/sbert_large_mt_nlu_ru`[5] for the vectorized representations of the sentences. We then K-means clustered the data for initial selections to be more diverse and representative based on semantic similarity. The final Russian dataset contained 95,300 text fragments.

The fragments were sent to three professional translators who translated the text into Buryat. Currently, the corpus consists of 11,392 Russian–Buryat sentences that have been translated manually, with work still progress.

3.2 Collaborations with local organizations

In order to extend the corpus, we cooperated with some regional institutions that deal with Buryat language materials. These were:

1. The State Translation Service of the Republic of Buryatia, which contributed bilingual presidential decrees, government resolutions, and other legal acts of subordinate level decisions. They were at first in DOCX format and were converted by the means of some automation into a format of a parallel table – a step-by-step process.

2. The Buryat Research Center of the Siberian Branch of the Russian Academy of Sciences (BRC SB RAS) which enabled access to five parallel literary texts, but these texts had frequent mismatches at the level of sentence alignment due to translation losses or free rendering of the text. Reasonable estimates claim that only the texts which were structurally most homogeneous were chosen to be included in the corpus, other texts were set aside for possible later processing.

3.3 Web-based Data Collection

A significant portion of the Buryat-Russian parallel data was collected from the web.

[1] https://translate.google.com

[2] https://bxr.wikipedia.org/wiki/

[3] https://huggingface.co/datasets/buryat-translation/buryat_russian_parallel_corpus

[4] https://tatianashavrina.github.io/taiga_site/

[5] https://huggingface.co/ai-forever/sbert_large_mt_nlu_ru

ааараг 1. 1) редкий, необычный; **ааараг ушар** редкий случай; 2) *перен.* аморальный; негодный, никчёмный; **ааараг урагшагүй ябадалтай хүнүүд** никуда не годные люди (*о бездельниках, тунеядцах и т.д.*); **2.** редко; **алдуу хэхэнь ааараг хүн даа** человек-то он такой, что редко ошибается.

Figure 1: Example of the dictionary article

Several dictionaries are available for the Buryat language. For our purposes, we selected the Buryat-Russian dictionary by Shagdarov and Cheremisov (Shagdarov and Cheremisov, 2010) due to its extensive scope and detailed coverage, which surpasses that of other dictionaries. This dictionary contains 30,000 words, provides grammatical information and usage examples, characteristic of Buryat culture. We encountered several challenges during the data extraction process. First, the dictionary was available only as a PDF scan, which resulted in suboptimal optical character recognition (OCR) quality. We experimented with both ABBYY FineReader OCR and Tesseract OCR, but neither significantly improved the accuracy of the text extraction. Second, the extensive and complex nature of the information made it difficult to extract parallel data. The dictionary entries included multiple meanings separated by commas, semicolons, Arabic or Roman numerals, letters, and additional details enclosed in brackets. Buryat words were presented in bold, grammatical information in italics, and Russian translations in regular font. Example of the dictionary article presented in Figure 1. To parse this structured information, we relied on regular expressions. Third, some pages suffered from unrecognized fonts, which required alternative approaches. For these cases, we employed the large language model Claude Sonnet 3.5[6]. However, using a multimodal large language model to process the entire book was not feasible due to the high associated costs.

Religious literature is another common source of parallel data. For Buryat, the only available resource was the Bible[7]. We aligned the text using regular expressions based on the enumerated verses. However, the translations were not always precise. Certain content in the Buryat Bible was omitted, and in some instances, multiple verses were combined into a single sentence. These issues were also addressed using regular expressions to ensure proper alignment and extraction.

We identified several bilingual books, which were translations between Buryat and Russian. A key challenge was aligning sentence pairs from these texts, as differences in structure and translation styles introduced inconsistencies. To address this, we fine-tuned the LaBSE encoder (Feng et al., 2022) on a previously collected dataset using the methodology described in (Dale, 2022), allowing us to effectively extract parallel sentences.

We also explored the use of Wikipedia as a potential source of parallel data. However, the corresponding articles in Buryat and Russian were found to be significantly different. This discrepancy likely stems from the fact that much of the Buryat Wikipedia content was translated from Russian prior to 2015, while the Russian articles have undergone substantial changes since then. As a result, we were unable to extract high-quality sentence pairs from Wikipedia and it was excluded from the final dataset.

A buryat monolingual corpus[8] was created by collecting texts from books sourced from websites[9] [10] [11] and Internet news articles in Buryat[12]. The monocorpus is used to enhance tokenizer of translation models. To further expand the parallel corpus, a subset of the news articles was translated into Russian using the large language model Claude 3.5 Sonnet (20240620). At the time of creation, this model provided best translation quality, enabling us to significantly enrich the dataset and improve the overall performance of the translation system.

Finally, to improve the quality of the dataset, we filtered out poorly aligned sentences using a heuristic based on sentence length and cleaned up Russian borrowings by applying a heuristic based on Levenshtein distance, both methods following the approach outlined in Dale (2022). After this cleaning process, we obtained a final dataset of 33 thousand words and 94 thousand sentences. The detailed breakdown of amounts by source is provided in Table 1.

[6] https://www.anthropic.com/claude

[7] https://ibt.org.ru/buryatskiy/vsya-bibliya/elektronnaya-kniga

[8] https://huggingface.co/datasets/buryat-translation/buryat_monocorpus

[9] https://old.buryatika.ru/

[10] https://soyol.ru/culture/books/

[11] https://nomoihan.com/books/

[12] https://burunen.ru/bur/

Source	Amount
Dictionary phrases	45,169
Dictionary words	33,449
Book alignments by BRC SB RAS	12,893
News translated with Claude	11,380
The Bible	8,591
Organized manual translations	11,392
Book alignments by model	4,415
Tatoeba	808

Table 1: Data sources, total of 94 thousand sentences and 33 thousand words

4 General concept of neural network

4.1 Creation of a Russian-Mongolian Parallel Corpus

High-quality neural machine translation requires large amounts of parallel data for training. However, the Buryat language is severely underrepresented in digital resources and is considered a low resource language. In such cases, transfer learning techniques or model adaptation based on related languages are often employed to improve translation performance.

The closest high-resource cognate language to Buryat is Modern Mongolian. It is more widely represented in digital space and is commonly included in multilingual models. Pretraining on Mongolian can thus serve as a valuable step for enhancing Russian–Buryat translation.

In order to test this theory, we tried to find Russian Mongolian parallel corpora that was publicly accessible. The only relevant dataset was found in the OPUS collection, which is one of the largest repositories of multilingual corpora. The corpus contains 387,310 sentence pairs. However, many of these translations were found to be of insufficient quality or poorly aligned, making the dataset unsuitable for direct use.

This led us to the conclusion that generating Russian-Mongolian parallel data by translating Russian texts into Mongolian using pretrained multilingual models was a better option. To determine the most accurate model for this task, we evaluated several candidates that support both Russian and Mongolian:

1. `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022)
2. `facebook/nllb-200-1.3B` (NLLB Team et al., 2022)
3. `google/madlad400-3b-mt` (Kudugunta et al., 2023)

The steps outlined below were taken to determine the best model to use for creating a Russian-Mongolian parallel corpus:

1. Each candidate machine translation model was used to translate a shared set of Russian sentences into Mongolian.
2. The generated translations were compared to the corresponding Mongolian references in the OPUS corpus using the ChrF++ metric.
3. The model with the highest average ChrF++ score was selected as the most accurate for Russian–Mongolian translation (see Table 2). The same Russian source corpus used for the Russian–Buryat data — the Taiga corpus — served as the basis for the synthetic Russian–Mongolian dataset. In this case, text clustering was not applied, as a large volume of data was preferred for pretraining purposes.

The model `facebook/nllb-200-1.3B` was found to perform best and was used to translate a total of 90,548 Russian sentences from the Taiga corpus.

4.2 Model Selection for Russian–Buryat Machine Translation

Given the low-resource nature of the Buryat language, selecting an appropriate neural architecture is critical for achieving reasonable translation quality. When selecting the model architecture, we opted for encoder-decoder type, as the cross-attention mechanism enables the model to better capture dependencies within the input and incorporate contextual information during decoding — a crucial aspect in machine translation. Experimental results presented in Raffel et al. (2023) and Fu et al. (2023) have shown that encoder-decoder models consistently outperform decoder-only architectures in translation tasks.

The first model selected for training on the Russian–Buryat parallel corpus was Google's `mt5-large`. This model was chosen due to its strong performance on machine translation tasks and broad multilingual support, including related languages such as Mongolian, making it a suitable candidate for low-resource scenarios.

The second model, `nllb-200-distilled-600M` by Meta (formerly Facebook), was specifically designed for multilingual machine translation with a focus on

Model	Average ChrF++ Score
facebook/nllb-200-distilled-600M	26.4
facebook/nllb-200-1.3B	**27.8**
google/madlad400-3b-mt	10.8

Table 2: Comparison of machine translation models using the ChrF++ metric

low-resource languages. Its compact architecture and high translation efficiency, as demonstrated in the model comparison presented in Section 4.1, make it particularly well suited to the task at hand.

4.3 Final Training Procedure

Before initiating the main training process, it was necessary to update the tokenizer vocabulary by incorporating new tokens specific to the Buryat language, which is not included in the original models. For well-represented languages in the training data, it is typical for each word to correspond to 2–3 tokens on average. However, Buryat words are segmented into a significantly larger number of tokens, indicating insufficient vocabulary coverage (Figure 2).

To address this, we utilized a Buryat monolingual corpus to extend the tokenizer. We used a dedicated dataset, described in Section 3.3, and supplemented it with Buryat sentences extracted from the training data. A new SentencePiece tokenizer was trained on this combined corpus.

The missing tokens identified in the newly trained tokenizer were then added to the original vocabulary of the NLLB tokenizer. Corresponding embedding vectors were initialized and appended to the model's embedding layer, ensuring that the model could represent and learn these new units during training.

The roles of language tags are crucial to the NLLB tokenizer. These special tokens are added to the beginning of source and target sentences to explicitly indicate the language. For Russian–Buryat translation, we added the tag `bxr_Cyrl` to both the tokenizer and the model configuration.

Following this preparation, we proceeded with training the neural machine translation models. Training was performed in both directions (Russian–Buryat and Buryat–Russian), with the direction chosen randomly for each batch. Details of the training corpus, hyperparameter settings, and results are provided in Section 5.

5 Experiments

We now turn to the experimental setup. Multiple versions of the `mt5-large` and `nllb-200-distilled-600M` models were trained. Each version was trained on an incrementally larger dataset, as the Russian–Buryat parallel corpus was continuously updated with newly translated sentence pairs.

For both models, the following hyperparameters were used:

- Batch size: 16
- Maximum sequence length: 512
- Number of training steps: 60,000

To evaluate translation quality, we used the BLEU and ChrF++ metrics, which are widely adopted in machine translation research.

The `mt5-large` model was pre-trained on the Russian–Mongolian parallel corpus described in Section 4.1.

In contrast, no additional pretraining was applied to the NLLB model, as it demonstrated strong performance during the Russian–Mongolian model comparison and achieved results comparable to the `facebook/nllb-200-1.3B` model used to generate the synthetic corpus.

The training results of all model versions are presented in Table 3.

The initial version of the model, referred to as `Fine-tuned NLLB-v0`, was trained before the manual translation process had begun. As a result, this version did not include any of the high-quality human-translated data. This limitation affected the overall translation quality, but the model served as a useful baseline for evaluating the impact of incorporating manually translated content in later versions.

Starting from the first version, manually translated data was incorporated into the training set. Additionally, we refined the regular expressions used for mining data from the dictionary and introduced back-translated data generated by Claude. As expected, translation quality improved with

bxr	bxr words	bxr tokens
Зарим хүнүүдтэ һүлдэ сэсэн мэргэн үгэ хайрлада…	[Зарим, хүнүүдтэ, һүлдэ, сэсэн, мэргэн, үгэ, х…	[_Зарим, _хүн, үүд, тэ, _h, үл, дэ, _с, э, сэн…
һүзэглэгшэдэй бэе бэедээ дуратай байхые урмашу…	[һүзэглэгшэдэй, бэе, бэедээ, дуратай, байхые, …	[_h, үз, эг, лэг, ш, эд, эй, _бэ, е, _бэ, ед, …
бидэ гурбан эрэшүүлые зорюута буурал хүгшөөдэ…	[бидэ, гурбан, эрэшүүлые, зорюута, буурал, хүг…	[_бид, э, _гур, бан, _эр, эш, үү, лые, _зор, ю…

Figure 2: Example of Buryat token segmentation

	ru-bxr		bxr-ru	
model	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
Fine-tuned NLLB-v0	6.56	22.14	1.31	10.00
Fine-tuned NLLB-v1	18.74	46.17	32.20	53.37
Fine-tuned mT5-v1	12.49	39.39	14.47	37.28
Fine-tuned NLLB-v2 (last)	**20.61**	**48.68**	**35.43**	**56.21**
Google Translate	8.93	37.61	29.58	52.35
Claude 3.5 Sonnet 20240620	8.00	34.80	25.12	52.03

Table 3: Evaluation of our and Google Translate model on test-set

each iteration. At the first stage, based on the observed performance, we decided to continue using only the NLLB-based model for further development. Once additional manually translated data became available, we trained the second version of the NLLB model, which, at the time of writing, represents the latest iteration. This version achieved the best results for Russian–Buryat translation.

To assess the performance of our model `Fine-tuned NLLB-v2`, we compared against publicly available systems: Google Translate and Claude 3.5 Sonnet. As shown in Table 3, our model outperforms both baselines in both directions (Russian–Buryat and Buryat–Russian), achieving ablhigher scores in both BLEU and ChrF++.

Translation performance varies across text types and directions (Table 4). The NLLB-v2 model achieved higher scores on manual translations, likely because it is most familiar with this domain. In the case of Bible texts, Google Translate performs best in the Buryat-to-Russian direction—possibly due to similar phrasing in its training corpus—while NLLB is stronger in the opposite direction. Phrasebook examples result in the lowest scores overall, which could be explained by their short length, limited context, and the frequent presence of set expressions, all of which make them difficult to translate reliably. In literary and legal texts, NLLB-v2 and Claude show similar performance in the Buryat-to-Russian direction, though reasons remain unclear. It is possible that Claude was trained on similar data.

It is important to note, however, that both the training and test sets used in our experiments were derived from the same pool of source texts, although split and processed independently. While this setup allows or stable evaluation, it may introduce a slight bias in favor of our model due to potential domain similarity. Still, the consistent advantage in scores suggests that our model performs better for Russian–Buryat translation than Google Translate and Claude, particularly in the Russian-to-Buryat direction.

6 Online translator

To make our machine translation model accessible to the public, we released it online[13]. Figure 3 demonstrates the graphic user interface of the translator. To make the model suitable for usage in web, we made quantization of the model with ctranslate (Klein et al., 2020).

[13] `https://www.burtranslate.ru/`

		ru-bxr		bxr-ru	
Source Type	**Model**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
Manual translations	Fine-tuned NLLB-v2	**21.88**	**52.15**	**38.10**	**60.20**
	Google Translate	8.56	39.18	23.69	52.31
	Claude 3.5 Sonnet 20240620	9.43	38.40	33.42	59.43
The Bible	Fine-tuned NLLB-v2	**20.49**	**47.56**	40.07	57.57
	Google Translate	10.00	36.83	**54.36**	**68.61**
	Claude 3.5 Sonnet 20240620	3.67	29.10	11.72	37.29
Phrasebooks	Fine-tuned NLLB-v2	**6.25**	**28.69**	**11.20**	30.89
	Google Translate	5.94	25.74	8.33	28.75
	Claude 3.5 Sonnet 20240620	4.43	25.82	8.43	**31.64**
Literature and regulations	Fine-tuned NLLB-v2	**16.83**	**39.86**	18.01	40.48
	Google Translate	9.18	36.12	13.20	37.45
	Claude 3.5 Sonnet 20240620	9.40	33.93	**24.94**	**49.64**

Table 4: Evaluation of our model and Google Translate on test-set by source types.

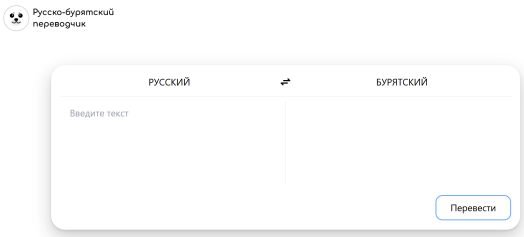

Figure 3: Graphic user interface of online translator

7 Human evaluation

We asked six native speakers of Buryat to participate in the evaluation of translations. Given that the average age of the participants was 57.8 years, we opted for a simplified rating scale consisting of two criteria: accuracy and fluency, both rated on a 5-point scale (with 5 indicating a perfect translation):

- Accuracy: Assesses how faithfully the translation preserves the meaning of the original sentence. Accuracy: Assesses how faithfully the translation preserves the meaning of the original sentence.
- Fluency: Evaluates the grammatical correctness and naturalness of the translation in the target language.

Each participant assessed 15 sentences in each translation direction (bxr-ru and ru-bxr), resulting in a total of 90 unique sentences evaluated manually. To ensure reliability, each sentence was reviewed by two different raters. The evaluated texts were randomly selected from the test corpus. The average scores are summarized in Table 5.

The manual evaluation suggests that the translations produced by the model are generally acceptable, particularly in terms of accuracy. However, lower fluency scores — especially in pessimistic cases — indicate that the output sometimes lacks grammatical correctness or natural phrasing. This highlights the need for further improvement.

8 Conclusion

In this work, we introduce a Buryat-Russian machine translation model, along with a parallel corpus of 127K sentence pairs and a monolingual Buryat corpus of 214K sentences. All resources are publicly released to support further research in low-resource language technologies.

Our model shows slightly better performance compared to Google Translate's Buryat-Russian system on our test dataset. Native speakers have evaluated the translations as acceptable for practical use.

We hope that this work will contribute to the development of computational linguistics for the Buryat language and provide a foundation for future research. By making these resources available, we aim to support efforts toward the preservation and promotion of Buryat in the digital domain.

Limitations

Machine translation systems for Buryat have great potential to support language learning and increase the availability of content in Buryat. However,

Table 5: Average manual evaluation scores for bxr-ru and ru-bxr

Metric	Group 1	Group 2	Group 3	Total Averages
ru-bxr				
Average Accuracy	4.13	4.06	3.19	3.79
Average Fluency	4.00	3.97	2.43	3.47
Pessimistic Accuracy	3.80	3.40	2.60	3.27
Pessimistic Fluency	3.67	3.27	1.73	2.89
bxr-ru				
Average Accuracy	3.34	3.63	3.06	3.34
Average Fluency	3.33	3.57	2.53	3.14
Pessimistic Accuracy	2.73	2.87	2.47	2.69
Pessimistic Fluency	2.73	2.87	1.40	2.33

these systems are not without significant limitations that need to be addressed.

A major concern is the accuracy of translations. Machine translation often makes mistakes, such as generating non-existent words, providing incorrect definitions, or producing grammatical errors. These inaccuracies can lead to misunderstandings and may even influence the language negatively if users unknowingly adopt incorrect forms. Additionally, the current model is still under development and cannot yet be fully trusted. Users are advised to double-check translations, especially in critical contexts, as over-reliance on automated systems can result in errors being propagated.

Another challenge is the lack of representation of Buryat dialects. Most models are trained on the literary standard of the language, leaving out the rich diversity of regional variations. This focus on a single dialect makes it harder for speakers of other dialects to benefit from the system and limits learners' exposure to the full range of linguistic expression within the Buryat language.

Cultural and contextual nuances also present difficulties. Machine translation struggles with idiomatic expressions, metaphors, and culturally specific references, which can lead to mistranslations or loss of meaning. For a language like Buryat, which carries deep cultural significance, this limitation can hinder effective communication.

Finally, the scarcity of high-quality training data further restricts the system's capabilities. Limited and imbalanced datasets can introduce biases and reduce performance, particularly in informal or specialized contexts. Addressing these challenges will require expanded and more diverse datasets, as well as ongoing refinement of the model.

While machine translation systems offer valuable support for Buryat, careful attention must be paid to these limitations to ensure their responsible and effective use.

Acknowledgments

The work is supported by the Mathematical Center in Akademgorodok under the agreement № 075-15-2025-349 with the Ministry of Science and Higher Education of the Russian Federation.

We would like to thank the translators who contributed to this project, as well as The State Translation Service of the Republic of Buryatia and The Buryat Research Center of the Siberian Branch of the Russian Academy of Sciences for providing parallel data. We are also grateful to David Dale for his great work on open-source low-resource translation.

Finally, we thank our relatives for their help in evaluating the models and for their support during the study.

References

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi N. Baljekar, Xavier García, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, and 5 others. 2022. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983.

Ali B. Berberov, Bogdan S. Teunaev, and Liana B. Berberova. 2024. The first neural machine translation system for the karachay-balkar language. In *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, pages 1720–1723.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

David Dale. 2022. The first neural machine translation system for the Erzya language. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 45–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *Preprint*, arXiv:2304.04052.

Alexandru-Iulius Jerpelea, Alina Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for Aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228, Abu Dhabi, UAE. Association for Computational Linguistics.

Mamadou Keita, Elysabhete Ibrahim, Habibatou Alfari, and Christopher Homan. 2024. Feriji: A French-Zarma parallel corpus, glossary & translator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–9, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

VP Konovalov and ZB Tumunbayarova. 2018. Learning word embeddings for low resource languages: the case of buryat. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 331–341.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Toadoum Sari Sakayo, Angela Fan, and Lema Logamou Seknewna. 2023. Ngambay-French neural machine translation (sba-fr). In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 39–47, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Lubsan Shagdarov and Konstantin Cheremisov. 2010. *Buryat-Russian dictionary*, volume 1-2. Republic typography, Ulan-Ude.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

UNESCO. 2010. *Atlas of the world's languages in danger*. UNESCO.

Author Index