Voices Unheard: NLP Resources and Models for Yorùbá Regional Dialects

Orevaoghene Ahia[1,5] **Anuoluwapo Aremu**[5,6] **Diana Abagyan**[1] **Hila Gonen**[1]
David Ifeoluwa Adelani[3,4,6] **Daud Abolade**[6] **Noah A. Smith**[1,2] **Yulia Tsvetkov**[1]
[1]University of Washington [2]Allen Institute for AI [3]Mila - Quebec AI Institute
[4]McGill University [5]Lelapa AI [6]Masakhane NLP
oahia@cs.washington.edu

Abstract

Yorùbá—an African language with roughly 47 million speakers—encompasses a continuum with several dialects. Recent efforts to develop NLP technologies for African languages have focused on their standard dialects, resulting in disparities for dialects and varieties for which there are little to no resources or tools. We take steps towards bridging this gap by introducing a new high-quality parallel text and speech corpus YORÙLECT across three domains and four regional Yorùbá dialects. To develop this corpus, we engaged native speakers, travelling to communities where these dialects are spoken, to collect text and speech data. Using our newly created corpus, we conducted extensive experiments on (text) machine translation, automatic speech recognition, and speech-to-text translation. Our results reveal substantial performance disparities between standard Yorùbá and the other dialects across all tasks. However, we also show that with dialect-adaptive finetuning, we are able to narrow this gap. We believe our dataset and experimental analysis will contribute greatly to developing NLP tools for Yorùbá and its dialects, and potentially for other African languages, by improving our understanding of existing challenges and offering a high-quality dataset for further development. We release YORÙLECT dataset and models publicly under an open license [1].

1 Introduction

While great strides have been made in developing NLP resources for low-resource languages, the majority of these efforts have been directed towards the "standard" dialect of these languages, largely neglecting the long tail of non-standard dialects spoken by millions (Faisal et al., 2024; Alam et al., 2024). Dialects of a language exhibit nuanced yet distinguishable differences in lexicon, pronunciation, spelling, and syntax, mirroring regional, societal, and cultural differences (Chambers and Trudgill, 1998). Usually, a "standard" dialect is the dialect with the highest population of speakers, and sometimes the only dialect with a standard orthography (Milroy and Milroy, 2012).

African languages are linguistically diverse (Adebara and Abdul-Mageed, 2022; Siminyu and Freshia, 2020), yet severely under-resourced. Most of these languages have numerous varieties, (usually regional), some of which are mostly-spoken and lack a standard orthography (Batibo, 2005; Heine and Nurse, 2000). Developing language technologies has been incredibly challenging for African languages (Nekoto et al., 2020; Muhammad et al., 2023; Ogundepo et al., 2023; Adelani et al., 2023; Dione et al., 2023; Adelani et al., 2024, 2021b), partly due to the scarcity of extensive language resources required for developing systems that are robust to the variations in linguistic features (Adebara and Abdul-Mageed, 2022; Siminyu and Freshia, 2020).

To address this problem, in this work we focus on curating dialectal resources for Yorùbá, a low-resource language with 47 million native speakers around the world. Yorùbá language is native to Southwestern Nigeria, Republic of Benin, and Republic of Togo. Yorùbá encompasses a dialect continuum including several distinct regional dialects (Rowlands, 1967). Due to Yorùbá's low-resource status, the majority of published NLP work have been done on the Standard Yorùbá dialect (Oguremi et al., 2024; Aremu et al., 2023; Ahia et al., 2021; Dione et al., 2023; Shode et al., 2023; Ogundepo et al., 2023; Akinade et al., 2023; Adelani et al., 2023; Muhammad et al., 2023; Adelani et al., 2021a; Adebara et al., 2022, 2021; Lee et al., 2023).

We introduce the first-ever corpus of high quality, contemporary Yorùbá speech and text data parallel across four Yorùbá dialects; Standard Yorùbá, Ifẹ̀/ i f ɛ /, Ìlàjẹ/ i l a dʒ ɛ / and Ìjẹ̀bú/ i dʒ ɛ b u / in three domains (religious, news, and Ted

[1]Code and data available at https://github.com/orevaahia/yorulect

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4392–4409
November 12-16, 2024 ©2024 Association for Computational Linguistics

English	Standard	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Domain
All the efforts to talk to ASUU chairman failed because he said he has nothing to say	Gbogbo ìgbiyànjú láti bá alága ASUU sọ̀rọ̀ lò jásí pàbó nitori ó ni òun kò ni ohunkóhun láti sọ .	Gbogbo ìgbiyànjú láti bá alága ASUU sọ̀rọ̀ re jasi afo to ri ó sọ fo òún ni ohun kóhun láti sọ .	Gbogbo ègbiyànjú látẹ bá alága ASUU sọ̀rọ̀ lọ̀ jásí pàbó torí ó ghíi òun nẹ́ ihunkíhun ún sọ .	Dede ìgbiyànjú áti bá alága ASUU fọ̀ rèé já ní pàbo torí ó fọ̀rọ́ pé ó ghún né irú kirun gho fé fò .	News
They called unto God in the upper room for the release of the holy spirit .	Wón ké pe ọ̀lórun ni yàrá orí òkè fún itújáde ẹ̀mí mímó .	Wón ké pè ọ̀lórun ni yàrá orí òkè fún itú jáde ẹ̀mí mímó .	Igán ké pe ọ̀lóun nẹ́ yàrá orí òkè ún ẹ̀tújáde ẹ̀mi mẹ́mọ́	Ghón kélè kpè ọ̀lórun ni yàrá orígho òkè ghún itújáde ẹ̀mí mímó .	Religion
We all look for characteristics that has to do with self-centeredness, and they are similar to this.	Gbogbo wa la máa ń wá àwọn ànimọ́ tó ni i ṣe pẹ̀lú iwa imọtara nikan, ìrísí wọn si jọ èyi.	Dede wa re n wa iwa ànimọ́ rè nii ṣe pẹ̀lú iwa imọlara nikan, irisi wọ si jọ ìwé	Gbogbo ria la máa ghá inọn ànẹ́mọ́ kọ́ nẹ́ẹ́ i se pẹ̀lú èghà èmọtara ọni nikàn, ẹ̀risi rian sèè jọ yèé .	Dede gha rèé mi fẹ́ àghan ànimá yii nẹ́ i se kpẹ̀lu ighà imòtara ọnẹ nùkàn, ìrísí ghàn si jọ èyi	Ted Talks

Table 1: Examples of parallel translations across all dialects and domains in YORÙLECT. Words that are unique across all dialects are highlighted in *red*.

talks). This newly curated benchmark, developed with native speakers, can be used in (text-to-text) machine translation (MT), automatic speech recognition (ASR), speech-to-text translation (S2TT), and speech-to-speech translation (STST) tasks. We discuss in detail the data curation process, criteria for data selection, and the steps we took to ensure data quality and integrity (§3). We first conduct extensive experiments evaluating the zero-shot performance of recent state-of-the-art models for MT, ASR, and S2TT (§4, §5). Our results and analysis indicate that current models are not robust enough to handle existing variation in Yorùbá dialects. Given these poor results, we proceed to adapt (fine-tune) existing models on our training data across all tasks to boost overall performance. With 802 training instances in each dialect, this approach leads to an average increase of 14 and 5 BLEU points for both MT and S2TT respectively, as well as a 20-point decrease in word-error-rate for ASR. Our work aims to motivate the community to build technology for languages alongside their dialects, especially for low-resource dialects of low-resource languages, as this will promote linguistic diversity, and ensure that technological advancements benefit *all* language communities.

2 Yorùbá and its Regional Dialects

The Yorùbá language is spoken natively by roughly 47 million people in Nigeria[2] and in the neighboring countries of the Republic of Benin and Togo and also Côte d'Ivoire, Sierra Leone, Cuba, and Brazil. In Nigeria, Yorùbá speakers are mainly concentrated in the Southwest region, spanning states like Oyo, Ogun, Osun, Ondo, Ekiti, and Lagos, and North Central states like Kogi, and Kwara.

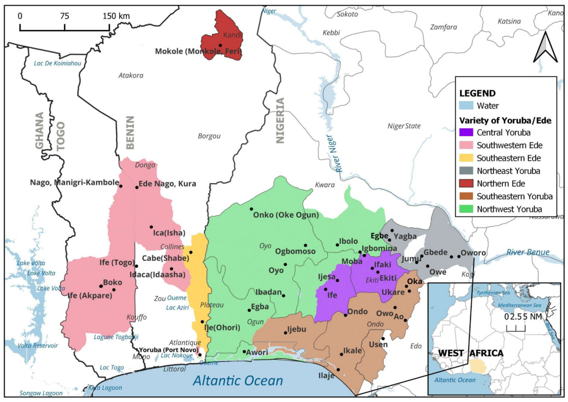

Figure 1: Geographical distribution of Yorùbá dialects in West Africa. Map from (Ozburn, 2023).

The extensive Yorùbá-speaking population and their dispersion across various regions have led to the emergence of geography-specific linguistic variations (Ballard, 1971). The number of existing Yorùbá dialects is estimated between twelve to twenty-six (Ojo, 1977; Adetugbo, 1982; Oyelaran, 1971; Oyelaran and Watson, 1991) and the differences present in these dialects are evident in pronunciation, grammatical structure, and vocabulary (Adetugbo, 1982; Przezdziecki, 2005; Olumuyiwa, 2009; Arokoyo et al., 2019; Olánrewájú, 2022). Also categorized as a Volta-Niger language within the Yoruboid subgroup of the Niger-Congo family, Yorùbá is a tonal language with three basic tones: low, middle, and high (Courtenay, 1969; Oyetade, 1988), as well as two or three contour tones.[3] Previous research (Adeniyi, 2021) has in-

[2] https://en.wikipedia.org/wiki/Yoruba_language

[3] A contour tone is a combination of two more basic tones such as a falling tone made up of a high tone and a low tone, or a rising tone consisting of a low tone followed by a high tone.

dicated that the phonetic nuances of contour tones are a major distinguishing feature among Yorùbá dialects.

Yorùbá dialectal forms in Nigeria can be classified into five regional groupings: Northwest Yorùbá (NWY), Northeastern Yorùbá (NEY), Central Yorùbá (CY), Southwest Yorùbá (SWY), and Southeast Yorùbá (SEY). Phonological, lexical, and grammatical differences distinguish these groupings, given the diverse levels of mutual intelligibility among the "regional" dialects within each category (Arokoyo et al., 2019; Olumuyiwa, 2016; Abiodun et al.). In this work, our focus lies on Ifẹ̀, a dialect in the Central Yoruba classification, Ìjẹ̀bú, and Ìlàjẹ dialects, which belong to the Southeast Yoruba classification. We display the geographical distribution of Yorùbá dialects in West Africa in Figure 1.

Comparative dialectal analysis Standard Yorùbá, Ifẹ̀, Ìjẹ̀bú and Ìlàjẹ dialects exhibit both similarities and differences in their orthographic representations, morphology, and semantics. For instance, standard Yorùbá dialect has fused velar fricative /ɣ/ and labialised voiced velar /g^{W}/ into /w/ (Adetugbo, 1982) and our curated data revealed a similar pattern for Ìjẹ̀bú. In contrast, Ifẹ̀ uses /ɣ/ in certain occurrences while Ìlàjẹ has heavily retained the /g^{W}/ and /ɣ/ in its representations. As a result, at the word level, "àwọn" (3p pl.) is represented similarly in standard dialect and Ìjẹ̀bú but as "ighọn" in Ifẹ̀ and "àghan" in Ìlàjẹ. Besides the contrastive consonant nature, the oral and nasal vowels are also both contrastive in Ifẹ̀ and Ìlàjẹ dialects respectively. Further analyses of YORÙLECT reveal that the low nasalised vowel /ã/ mostly follows "gh" in Ìlàjẹ while the back lower-mid nasalised vowel /ɔ̃/ accompanies "gh" in Ifẹ̀ dialect. One remarkable semantic variation is that standard Yorùbá dialect uses "sọ" and "wi pe" as *say/talk*, however for Ìlàjẹ and Ìjẹ̀bú the morpheme mostly used is "fọ" while Ifẹ̀ uses "ghii", all of which have the same semantics.

3 YORÙLECT Corpus

We curated parallel text and recorded high quality speech data across Ifẹ̀, Ìjẹ̀bú, Ìlàjẹ, and Standard Yorùbá dialects. Our data curation process involves three main steps: (i) text curation and dialect localization; (ii) speech recording; and (iii) text and audio alignment.

3.1 Text Curation and Dialect Localization

We collected textual Standard Yorùbá data from the following sources: (i) Bible study manuals;[4] (ii) the Yorùbá portion of MTTT, a collection of multitarget bitexts based on TED Talks (Duh, 2018); and (iii) Yorùbá news articles within the MAFT corpus (Alabi et al., 2022). Given resource limitations and the demanding nature of this task, we gathered 352 sentences from the Bible study manuals, 247 sentences from TED Talks, and 907 sentences from news articles, amounting to a total of 1,506 sentences. Next, we proceeded to localising the compiled Standard Yorùbá text into the three respective dialects: Ifẹ̀, Ìjẹ̀bú, and Ìlàjẹ by recruiting trained linguists and translators who are literate and also native speakers of the respective dialects. We hired two translators or linguists per dialect and gave each a different domains to localise. The localisation process took about six to eight weeks and this included the localisation, quality assessment and incorporation of corrections. We provided monetary compensation for the localisation of the text.

3.2 Speech Recording

Speaker selection is crucial when creating an ASR corpus; a speaker should be fluent, literate, trained, and familiar with voice recording (Ogayo et al., 2022; van Niekerk et al., 2017). Due to time constraints and speaker availability, we were only able to record speech in standard Yorùbá, Ifẹ̀, and Ìlàjẹ dialects, leaving Ìjẹ̀bú for a later version of the dataset. We retained the linguists and translators who localised the standard Yorùbá text into Ifẹ̀ and Ìlàjẹ dialects. We then recruited two additional native speakers per dialect that are literate in rendering the localised text into audio. All dialectal voice talents received monetary compensation. We first conducted an interview, then asked the new recruits to record random samples of the text and send the recordings for assessment. The audio and corresponding text are vetted, after which we selected native speakers with high reading competence, good voice texture, and reading pace. This brought the total number of voice talents per dialect to four. To ensure that each voice talent within a dialect recorded text across all domains, we divided text in each domain (religion, Ted, news) into four parts. Each person recorded roughly 375 sentences from each domain resulting in a total of 3 hours of

[4] https://faithrebuilder.org/conference-bible-study-manuals

	BLEU ↑				AfriCOMET ↑			
	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard
M2M100	0.00	0.49	0.25	0.49	0.26	0.27	0.26	0.30
NLLB-600M	7.26	7.52	5.78	16.51	0.52	0.50	0.49	0.65
GMNMT	**18.24**	**17.16**	**12.66**	**43.46**	**0.59**	**0.57**	**0.56**	**0.74**
Menyo	2.76	2.66	1.57	7.49	0.44	0.40	0.40	0.52
MT0	5.81	6.68	4.61	17.22	0.52	0.50	0.47	0.65
Aya	7.18	7.71	4.91	16.46	0.49	0.50	0.45	0.63

Table 2: Zero-shot MT evaluation across all models. Google Translate outperforms all other systems and shows greater robustness to dialectal variation. However, a significant performance gap remains compared to the Standard Yoruba dialect.

speech per dialect.

Recording is conducted using the speech recorder application designed by the YorubaVoice project (Ogunremi et al., 2024). The text files were uploaded per domain for each speaker on the YorubaVoice Recorder app. We used an M1 Pro 2021 chip MacBook with an audio-technica AT2020USB-X microphone set-up in an anechoic and sound-isolated voice recording booth for the recording process. Each text is recorded at 48 kHz and the audio files are provided in 16 bit linear PCM RIFF format. The app generates metadata that includes a unique speaker ID, audio ID with corresponding text, and the audio file. Finally, all the recordings were subjected to a quality control process by the data coordinator. We manually verified that the correct text was aligned with the appropriate audio file and re-aligned them when necessary. We also discovered one empty audio file in a particular dialect and proceeded to delete it, along with its corresponding text-audio pairs in all other dialects.

Final data statistics In total, the text portion of YORÙLECT consists of 1506 parallel sentences per dialect and 6024 sentences overall, while the speech portion consists roughly 3 hours of audio each in standard Yorùbá, Ifẹ̀ and Ìlàjẹ, resulting in 9 hours of speech in total. We split the text and audio pairs in each dialect into 804 training samples, 200 validation samples and 502 test samples.

4 Zero-shot Experiments

We start by evaluating the zero-shot performance of current state-of-the-art models on the test portion of YORÙLECT. Based on the results from this initial evaluation, we then adapt the top-performing zero-shot models by finetuning on the training portion of YORÙLECT and report results in §5.1. MT

Dialect	length (hours)	Avg. length (seconds)	Avg. tokens
Standard	2.93	6.99	15.81
Ìlàjẹ	3.30	7.89	15.84
Ifẹ̀	3.03	7.23	15.53
Ìjẹ̀bú	-	-	15.25

Table 3: Statistics of YORÙLECT. The number of train, validation and test samples is consistently (804/200/502) for each dialect.

experiments are conducted on all dialects, while ASR and S2TT experiments are conducted on all expect Ìjẹ̀bú.

4.1 Machine Translation

We evaluate two classes of translation systems: MT-specific models and LMs. Here, the MT-specific models use an encoder-decoder architecture and are trained on large amounts of parallel data in multiple languages, whereas the LMs are decoder-only models trained to maximize likelihood (i.e., next-token prediction) on text in multiple languages. All models we evaluate have standard Yorùbá text in their training data. We only evaluate translation from the standard language or dialect into English since these experiments are zero-shot and we cannot expect the models to generate text in one of the dialects. This essentially enables us to measure the robustness of all of these models to variation in the Yorùbá language.

MT-Specific Models We evaluate M2M-100 (Fan et al., 2020), NLLB (Costa-jussà et al., 2022), and MENYO-20k (Adelani et al., 2021a). M2M-100 and NLLB are multilingual MT models trained on data spanning 100 and 202 languages respectively. MENYO-20k is a Yorùbá-to-English-specific model fine-tuned on top of the multilingual pretrained mT5 model (Xue et al., 2021). MENYO-20k's model is trained with the

	ASR (WER) ↓			S2TT (BLEU) ↑		
	Ifẹ̀	Ìlàjẹ	Standard	Ifẹ̀	Ìlàjẹ	Standard
MMS	**85.38**	**83.79**	**72.50**	-	-	-
SeamlessM4T	96.14	101.99	80.14	5.52	3.30	13.16
Whisper	104.50	127.21	130.96	0.17	0.21	0.23

Table 4: Zero-shot performance on automatic speech recognition and speech translation.

MENYO-20k dataset, a curated multi-domain standard Yorùbá dataset with proper orthography.

Language Models We evaluate two multilingual LMs, Aya (Üstün et al., 2024) and MT-0 (Muennighoff et al., 2023), trained on 101 and 46 languages, respectively (standard Yorùbá included). We prompt the LM to generate translations in a zero-shot setting with the prefix "Translate to English: " added to each sentence and greedily decode the continuation. We do not provide in-context examples in order to create a comparable setting to the evaluation of MT-specific models.

Finally, we include Google Translate (GMNMT)[5] due to its widespread commercial use. We request the NMT model through the API, and cannot control any other aspects of its usage.

Results We measure translation quality using AfriCOMET (Wang et al., 2023) and BLEU (Papineni et al., 2002). Firstly, we report zero-shot performance across all models in Table 2. Although performance is relatively low across the board, among MT-specific models, NLLB performs best across all dialects, outperforming M2M100 and MENYO-20k. Comparing performance on LMs, Aya performs better than MT0 on all dialects except standard Yorùbá. Google Translate outperforms all systems across all dialects. Overall, we see a huge performance gap between standard Yorùbá and the rest of the dialects. This observation is not surprising and is very consistent across all systems. The results in Table 2 also show that Ìlàjẹ has the worst-performing BLEU score across all models. We hypothesize that this is because Ìlàjẹ is largely spoken in Ọndọ́ state, which is geographically distant from Ọyọ́ state where standard Yorùbá originated from.

4.2 Automatic Speech Recognition

We evaluate three models: Whisper (Radford et al., 2022), SeamlessM4T (Communication et al., 2023), and MMS (Pratap et al., 2024). All models include standard Yorùbá in their pretraining data. Whisper is an end-to-end ASR model, implemented as an encoder-decoder transformer, trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The authors argue that it is robust to accents and variations in speech. It was optimized to perform the tasks of transcribing audio into its original language and translating the audio into English text. SeamlessM4T is a multilingual and multimodal model that also translates and transcribes across speech and text. It is trained on 470,000 hours of mined speech and text-aligned data and supports ASR, S2TT, speech-to-speech translation, text-to-text translation and text-to-speech translation, although our focus here is ASR and S2TT. MMS is an ASR-only model finetuned on top of wav2vec 2.0 (Baevski et al., 2020) models across 1,107 languages. In addition to dense finetuning, they also finetune language-specific adapter modules (Houlsby et al., 2019) for each language in their pretraining data.

Results We report word error rate (WER) with the models MMS, SeamlessM4T, and Whisper in Table 4 (left). Performance is generally poor across all models, with MMS performing the best. We hypothesize that MMS performs best due to its training with parameter-efficient finetuning using language-specific adapters. We see an average performance gap of 12 points between standard Yorùbá and the other dialects on MMS and SeamlessM4T. With Whisper, the case is different: while the WER is generally very high, we see that only Ifẹ̀ is substantially better across all dialects. Upon manually reviewing the transcriptions from all models, we noticed that Whisper did not include diacritics in its generated transcriptions. Yorùbá is a tonal language, and diacritics play a crucial role in disambiguating word meanings. We believe that this, coupled with the generation of overly segmented transcriptions contributes to Whisper's exceptionally high word-error rate exceeding 100.

[5]https://translate.google.com/. API last accessed on June 7, 2024.

4.3 Speech Translation

We only evaluate Whisper (Radford et al., 2022) and SeamlessM4T (Communication et al., 2023). Just like MT, we only evaluate translation from the standard language or dialect into English as we cannot expect the models to generate text in any of the dialects without explictly finetuning it do so.

Results In Table 4 (right), we present the zero-shot speech-to-text translation (S2TT) results of SeamlessM4T and Whisper models, the only open-source models we are aware of that include coverage for Standard Yorùbá. Among all the tasks we evaluated, S2TT appears to be the most challenging. Performance is absolutely low for both models with Whisper performing particularly poorly. Across dialects, with SeamlessM4T, Standard Yoruba performs better yet again with an average of 9 points performance gap compared to Ìlàjẹ and Ifẹ̀.

5 Finetuning Experiments

5.1 Machine Translation

Next, we finetune NLLB-600M (Team et al., 2022) on the training portion of our dataset in both directions, English→Dialect and Dialect→English. We experiment with training all dialects jointly under the Yorùbá language code, and training the dialects separately by adding new language codes for each dialect and initializing them with the Yorùbá embedding. In an attempt to further boost performance, we augment our training data with 10k instances from MENYO-20k (Adelani et al., 2021a).[6]

Results In Figure 2 we analyze the translation quality following NLLB finetuning from Dialect→English, comparing it with both the translation quality prior to finetuning and with Google Translate, which serves as the top-performing zero-shot system (Table 2). Our results demonstrate that with only 802 training instances per dialects we outperform Google Translate on the non-standard dialects. While the performance of Google Translate remains notably superior for the standard dialect, we anticipate that scaling up the data could potentially bridge this gap.

We present results for fine-tuning from English→Dialect in Table 12 in the Appendix. Our observation is that performance is generally

[6]MENYO-20k was included in NLLB's pretraining data, however we try to include it in another step of language-specific finetuning.

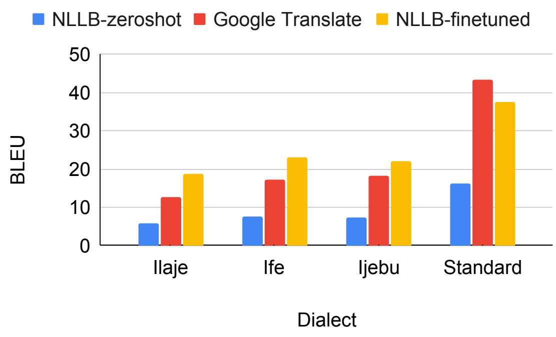

Figure 2: MT results (↑). We compare BLEU across Google Translate, NLLB prior to finetuning, and NLLB after finetuning.

worse than fine-tuning in Dialect→English direction. This is consistent with previous findings that translating into English could be easier than translating from it (Belinkov et al., 2017).

5.2 Automatic Speech Recognition

We finetune MMS (Pratap et al., 2024) and XLSR-Wav2Vec2 (Baevski et al., 2020). For the MMS model, we only finetune the Yorùbá adapter layer, while the other weights of the model are kept frozen.

Results We compare performance after finetuning XLSR and MMS with two different model sizes each: 300M and 1.3B parameters. MMS is a more suitable choice for finetuning because of its parameter efficiency, since we only have to tune the Yorùbá adapter layers. However, we choose to compare it with XLSR as well, as previous studies have reported significant performance improvements by finetuning XLSR (Ogunremi et al., 2024). In Figure 3, we first see that for XLSR, fine tuning a model with less capacity (300M parameters) yields better performance across all dialects compared to fine tuning a model with about $4\times$ more parameters. However, with MMS, we see that finetuning the 1.3B model yields a lower WER compared to finetuning the 300M model. Here, the performance gap is not as drastic as with XLSR.

On average, there is a performance improvement of approximately 20% after finetuning. As expected, across all models, the performance on the Standard Yorùbá dialect remains considerably better than that of Ìlàjẹ and Ifẹ̀. We expect that increasing the size of the finetuning data could help close this gap and could be addressed in future work.

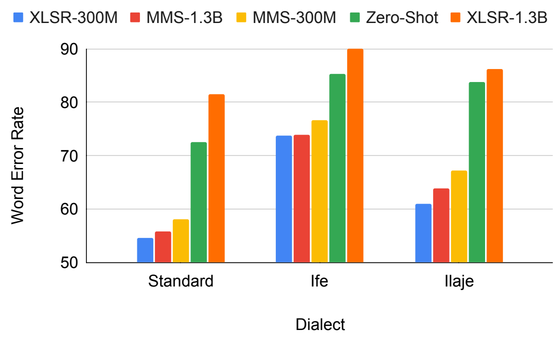

Figure 3: ASR results. (↓) We compare WER between zero-shot and jointly fine-tuning on all dialects on XLSR and MMS models.

5.3 Speech-to-Text Translation

SeamlessM4T (Communication et al., 2023) is the only model we finetune for speech-to-text-translation, since it its the best performing model from zero-shot experiments (see Table 4 and the only other S2TT model (to the best of our knowledge) with Yoruba in its training data asides. We finetune in the (Dialect→English) direction.

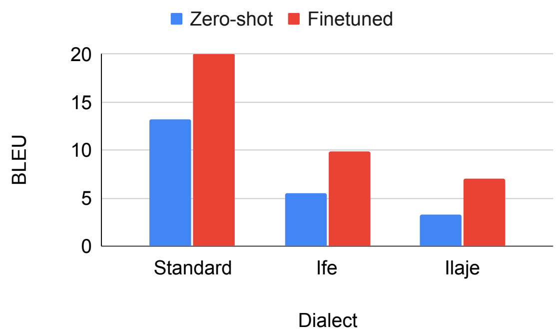

Figure 4: S2TT results (↑). We compare BLEU prior to finetuning and after finetuning SeamlessM4T.

Results The results in Figure 4 show that while we can reasonably boost performance on Standard Yorùbá after finetuning, it still remains a very hard task for the other dialects with just finetuning. We hypothesize that this occurs for two reasons, firstly the amount of Yorùbá S2TT data in SeamlessM4T is smaller than the data available to train ASR (Communication et al., 2023). Secondly, while there is notable lexical variation across Yorùbá dialects, the differences are even more pronounced in spoken language. This significant variation in pronunciation and intonation, coupled with the fact that S2TT data for Yorùbá is scarcer than ASR data makes the task of adaptation particularly challenging.

6 Human Evaluation

We complement automatic evaluation metrics with a human evaluation study to assess the quality of translations and transcriptions from the best models after fine-tuning for MT and ASR. Previous research has shown that word error rate (WER) is not nuanced, as it treats all errors in ASR text—insertions, deletions, and substitutions—the same, without considering their impact on readability (Itoh et al., 2015).[7] For ASR, one native speaker per dialect rated the quality of 30 randomly sampled transcriptions from the test set produced by our best ASR models after finetuning. After listening to the source speech they assess fluency (how natural and grammatically correct the transcription sounds in their dialect) and adequacy (how accurately the transcription conveys the meaning of the source speech) using a Likert scale of (1–5), the higher the better. In Table 5 we show that human raters consider the transcriptions of standard and Ifẹ̀ to be moderately adequate and fluent on average, compared to Ìlàjẹ. These findings align with our observations from automatic metrics.

	Adequacy ↑	Fluency ↑
Standard	3.37	3.03
Ìlàjẹ	2.73	2.62
Ifẹ̀	3.40	2.90

Table 5: Average human ratings of adequacy and fluency of transcriptions from the best ASR models after finetuning.

For MT, we ask human raters to compare the quality of translations from Google Translate with translations after finetuning NLLB, still focusing on fluency and adequacy still using a Likert scale (1–5). We provide the exact phrasings of instruction in the §A.4. Our results, displayed in Table 6, show that Google Translate is rated to be more fluent and accurate on Standard Yorùbá and Ìlàjẹ. However, our finetuned NLLB-600M model is rated to be more more fluent and accurate on Ifẹ̀ and Ìjẹ̀bú. The results on standard Yorùbá, Ifẹ̀ and Ìjẹ̀bú are very consistent with automatic evaluation results in Figure 2. This is not the case with Ìlàjẹ, as our ratings are lower compared to Google Trans-

[7] https://machinelearning.apple.com/research/humanizing-wer

late, which contrasts with our automatic evaluation in Figure 2.

	Adequacy ↑		Fluency ↑	
	GMNMT	NLLB	GMNMT	NLLB
Standard	4.47	4.13	4.73	4.60
Ìlàjẹ	2.73	2.63	2.10	1.83
Ifẹ̀	2.90	3.67	2.73	3.57
Ìjẹ̀bú	3.37	3.96	3.60	4.20

Table 6: Average human ratings of adequacy and fluency of test set translations comparing Google Translate with the best models after fine-tuning NLLB-600M

7 Analysis and Discussion

Does edit distance explain performance gaps? In this analysis we aim to understand how dialectal similarity influences model adaptation during fine-tuning. Ideally, we expect dialects with higher similarity to Standard Yorùbá to perform better. Edit distance (Levenshtein, 1966) is a simple method commonly used in dialectometry to infer pronunciation differences between language dialects (Nerbonne et al., 2020, 1996; Heeringa, 2004). In our work, we use edit distance as a proxy for similarity between Standard Yorùbá and the other dialects in our corpus, expecting that dialects with a higher degree of similarity (lower edit distance) will perform better. We compute the average edit distance per dialect, $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d(s_i, t_i)$, where N is the number of sentences in the test set of the dialect, s is the sentence in Standard Yorùbá, t is the sentence in the corresponding dialect, and $d(s_i, t_i)$ is the edit distance between s_i and t_i at the character-level.

We present the results of this analysis in MT in Table 7. As expected, Ifẹ̀ has the smallest edit distance from Standard Yorùbá and respectively also the best performance after finetuning. However we surprisingly see that while Ìjẹ̀bú has a higher edit distance than Ìlàjẹ, the model performance is higher for Ìjẹ̀bú. We conclude that edit distance has a weak correlation with our MT metrics.

Dialect	Avg. ED	BLEU	AfriCOMET
Ifẹ̀	24.66	22.97	0.59
Ìlàjẹ	38.07	18.64	0.55
Ìjẹ̀bú	41.46	21.98	0.60

Table 7: Average edit distance and MT-Metrics comparison for MT across dialects after finetuning NLLB.

For ASR, we compute edit distance on phonetic transcriptions using the PanPhon library developed by (Mortensen et al., 2016). The phonetic edit distance between standard Yorùbá to Ìlàjẹ and Ifẹ̀ is 34.99 and 44.4, respectively. Here again, we also see no correlations between edit distance and performance on dialect adaptation.

Joint vs. dialect-specific finetuning. Dialects often exhibit rather subtle variations in text and speech. In data-constrained scenarios like ours, it is reasonable to expect that jointly finetuning on all dialects would result in better performance compared to fine-tuning on each dialect individually. In our earlier finetuning experiments detailed in §5, we explored joint training. Now, we try to compare performance between joint training and individual training on MT and ASR tasks. We generally see that on both tasks, joint training is beneficial. In MT, Table 11 in the Appendix shows a huge drop in performance across all dialects when we finetune on each dialect individually. This suggests that by jointly finetuning, the model leverages shared features across dialects for mutual benefit. Although, it is also possible that we observe better results due to 3X increase in data size. However, in ASR, as shown in Table 8, the drop in performance with individual finetuning is not as pronounced as with MT. We believe that in this case, the subtle variations in speech are sometimes significant, making it more challenging to greatly benefit from joint training. We however acknowledge that the data size of each individual dialect is one-fourth of the whole training set, so data paucity might also be influencing these results.

8 Related Work

Previous works that have developing technologies and resources for machine translation (Ahia et al., 2021; Adebara et al., 2022, 2021; Lee et al., 2023; Akinade et al., 2023; Adelani et al., 2021a), automatic speech recognition (Ogunremi et al., 2024; Communication et al., 2023; Baevski et al., 2020) and speech translation (Communication et al., 2023; Oneata and Kamper, 2024) for Yorùbá have largely focused on the standard Yorùbá dialect. This is because, just like other African languages, standard Yorùbá is also very low-resourced, and all efforts have been directed there. Several works have shown that models often exhibit performance disparities between standard languages and their dialectal counterparts (Diab, 2016; Nigmatulina et al., 2020; Kantharuban et al., 2023; Ziems et al., 2023; Faisal et al., 2024; Ahmadi et al., 2024; Joshi et al., 2024; Blaschke et al., 2023; Aji et al., 2022; Abdul-Mageed et al., 2023). Arabic language has roughly 30 regional dialects. Whilst majority of work has being done on Modern Standard Arabic (MSA), Arabic still has the widest coverage of tasks and datasets across several of its dialects (Faisal et al., 2024; Diab and Habash, 2012; Bouamor et al., 2018; Kchaou et al., 2020). Within African languages, some works that aim to build dialect-aware models have conducted their studies on Igbo (Emezue et al., 2024), Luhya (Siminyu et al., 2021; Chimoto and Bassett, 2022), Bemba (Sikasote and Anastasopoulos, 2022) and Kiswahili (Siminyu et al., 2022).

9 Conclusion

We introduce YORÙLECT—the first high quality parallel text and speech corpus for four Yorùbá dialects sourced primarily from native speakers, to enable ASR, MT and S2TT tasks for widely-spoken varieties of Yorùbá. We have provided a detailed documentation of data curation process from standard text creation, to dialect localization and speech recording in communities where these dialects are spoken. Extensive experiments reveal that current models are not robust to dialectal variation, and improve significantly after our dialect-adaptive finetuning. Overall, our data collection methodology, new resources and improved models take a step towards enhancing the quality and equity of NLP technologies for Yorùbá dialects and potentially other African languages.

Ethical Considerations

Our datasets and models will be publicly released under an open license to foster research and continue to promote the development of NLP tools for African languages. Transcriptions, recordings and translations are carried out by paid native speakers who provided consent to use their voice to train our models. We acknowledge that the limited size of the corpus might not represent perfectly communities and speakers of the dialects. Further, dialectal generations, particularly when erroneous, could be perceived as biased or even microaggressions by some native speakers, as well as dialect-specific errors from the models (Wenzel and Kaufman, 2024). While our work provides resources that aim to reduce dialectal biases and unfairness in multilingual NLP systems, future work should focus on careful human evaluation of how these resources are incorporated in end-user tools.

Limitations

A limitation of our work is the robustness of the metrics we use for evaluation. While all of these metrics are standard for all of the tasks, we acknowledge that model-based metrics like AfriCOMET (Wang et al., 2024) could be biased towards standard dialects that their models have been trained on. Exploring model-based metrics that facilitate robust evaluations on dialectal tasks remains a challenge for future work (Faisal et al., 2024).

Additionally, the text portion of our dataset is translated from the standard dialect into English and the non-standard dialects. We acknowledge that this could introduce translation artifacts known as translationese (Volansky et al., 2015) that are not present in the source dialect. However, we believe that the benefits of our dataset outweighs the potential risks of these artifacts.

Acknowledgements

We would like to thank the UW NLP community for valuable discussions of this work. We are grateful to Farhan Samir and Sachin Kumar for discussions on experiments and analysis. We also thank Oduwole Folake and Olomolatan Anuoluwapo for their help with evaluating translations and transcriptions from our models. David Adelani is supported by Canada CIFAR AI Chair program. We gratefully acknowledge support from the National Science Foundation under CAREER Grant No. IIS2142739, and NSF grants No. IIS2125201 and IIS2203097. This work was also supported in part by gift funding from Google, MSR, and OpenAI.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Michael A Abiodun, Samuel O Akintoye, and Jelili Adewale Adeoye. A diachronic study of the loss of/w/in some lexical items in central yoruba dialect group.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yoruba-english mt in low-resource, morphologically-unmarked settings.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2022. Linguistically-motivated Yorùbá-English machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5066–5075, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024. IrokoBench: A new benchmark for African languages in the age of large language models.

Kolawole Adeniyi. 2021. A tonal identification of yoruba dialects. *Dialectologia: revista electrònica*, (28):1–31.

Abiodun Adetugbo. 1982. Towards a yoruba dialectology. *Yoruba language and literature*, pages 207–224.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sina Ahmadi, Daban Q. Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for central kurdish varieties.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. 2023. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Anuoluwapo Aremu, Jesujoba O. Alabi, and David Ifeoluwa Adelani. 2023. Yorc: Yoruba reading comprehension dataset.

Bolanle Elizabeth Arokoyo, Olamide Oluwaseun Lagunju, et al. 2019. A lexicostatistics comparison of standard yoruba, akure and ikare akoko dialects. *Journal of Universal Language*, 20(2):1–27.

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.

John A Ballard. 1971. Historical inferences from the linguistic geography of the nigerian middle belt. *Africa*, 41(4):294–305.

H. Batibo. 2005. *Language Decline and Death in Africa: Causes, Consequences, and Challenges*. Multilingual matters. Multilingual Matters.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

J. K. Chambers and Peter Trudgill. 1998. *Dialectology*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Everlyn Chimoto and Bruce Bassett. 2022. Very low resource sentence alignment: Luhya and Swahili. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual multimodal machine translation.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Karen Ruth Courtenay. 1969. A generative phonology of yorùbá.

Mona Diab. 2016. Processing dialectal Arabic: Exploiting variability and similarity to overcome challenges and discover opportunities. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, page 42, Osaka, Japan. The COLING 2016 Organizing Committee.

Mona Diab and Nizar Habash. 2012. Arabic dialect processing tutorial. In *Tutorial Abstracts at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Kevin Duh. 2018. The multitarget ted talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

Chris Chinenye Emezue, Ifeoma Okoh, Chinedu Mbonu, Chiamaka Chukwuneke, Daisy Lal, Ignatius Ezeani, Paul Rayson, Ijemma Onwuzulike, Chukwuma Okeke, Gerald Nweya, Bright Ogbonna, Chukwuebuka Oraegbunam, Esther Chidinma Awo-Ndubuisi, Akudo Amarachukwu Osuagwu, and Obioha Nmezi. 2024. The igboapi dataset: Empowering igbo language technologies through multi-dialectal enrichment.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages.

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. 2020. Beyond english-centric multilingual machine translation.

Wilbert Heeringa. 2004. Measuring dialect pronunciation differences using levenshtein distance.

B. Heine and D. Nurse. 2000. *African Languages: An Introduction*. African Languages: An Introduction. Cambridge University Press.

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*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Nobuyasu Itoh, Gakuto Kurata, Ryuki Tachibana, and Masafumi Nishimura. 2015. A metric for evaluating speech recognizer output based on human-perception model. In *Interspeech*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *ArXiv*, abs/2401.05632.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich-Belguith. 2020. Parallel resources for Tunisian Arabic dialect translation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 200–206, Barcelona, Spain (Online). Association for Computational Linguistics.

Jaechan Lee, Alisa Liu, Orevaoghene Ahia, Hila Gonen, and Noah Smith. 2023. That was the last straw, we need more: Are translation systems sensitive to disambiguating context? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4555–4569, Singapore. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

J. Milroy and L. Milroy. 2012. *Authority in Language: Investigating Standard English*. Routledge Linguistics Classics. Taylor & Francis.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. 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. ACL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. *Findings of EMNLP*.

John Nerbonne, Wilbert Heeringa, Jelena Proki´c, and Martijn Wieling. 2020. 5 dialectology for computational linguists.

John Nerbonne, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between dutch dialects.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building african voices. In *23rd Annual Conference of the International Speech Communication Association (InterSpeech 2022)*, Incheon, Korea.

Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.

Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2024. Ìròyìnspeech: A multi-purpose yorùbá speech corpus.

Valentine Ojo. 1977. English-yoruba language contact in nigeria. *(No Title)*.

Emmanuel mniyi Olánrewájú. 2022. A contrastive analysis of interrogatives in standard yorùbá and central yorùbá dialects. *Hayatian Journal of Linguistics and Literature*, 6(1):24–46.

Temitope Olumuyiwa. 2009. The high tone syllable in central yoruba dialects. *Nordic Journal of African Studies*, 18(2):9–9.

Temitope Olumuyiwa. 2016. Vowel assimilation in èkìtì dialects of yorùbá language. *Lingue e Linguaggi*, 17:143–154.

Dan Oneata and Herman Kamper. 2024. Translating speech with just images.

Olasope O Oyelaran and RL Watson. 1991. Africanisms in American culture.

Olasope Oyediji Oyelaran. 1971. *Yoruba phonology*. Stanford University.

Benjamin A Oyetade. 1988. *Issues in the analysis of Yoruba tone*. University of London, School of Oriental and African Studies (United Kingdom).

Avery Ozburn. 2023. Language profiles project. Accessed: 2024-06-13.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Marek A Przezdziecki. 2005. *Vowel harmony and coarticulation in three dialects of Yoruba: phonetics determining phonology*. Cornell University.

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

E. C. Rowlands. 1967. Ayo bamgbose: A grammar of Yoruba. (west African language monograph series, 5.) xii, 175 pp. cambridge: University press in association with the west african languages survey and the institute of african studies, ibadan, 1966. 35s. *Bulletin of the School of Oriental and African Studies*, 30(3):736–737.

Iyanuoluwa Shode, David Ifeoluwa Adelani, JIng Peng, and Anna Feldman. 2023. NollySenti: Leveraging transfer learning and machine translation for Nigerian movie sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998, Toronto, Canada. Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Kathleen Siminyu and Sackey Freshia. 2020. AI4D - African language dataset challenge. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 68–77, Seattle, USA. Association for Computational Linguistics.

Kathleen Siminyu, Xinjian Li, Antonios Anastasopoulos, David Mortensen, Michael R. Marlo, and Graham Neubig. 2021. Phoneme recognition through fine tuning of phonetic representations: a case study on luhya language varieties. In *Proceedings of Interspeech 2021*.

Kathleen Siminyu, Kibibi Mohamed Amran, Abdulrahman Ndegwa Karatu, Mnata Resani, Mwimbi Makobo Junior, Rebecca Ryakitimbo, and Britone Mwasaru. 2022. Corpus development of kiswahili speech recognition test and evaluation sets, preemptively mitigating demographic bias through collaboration with linguists. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 13–19, Dublin, Ireland. Association for Computational Linguistics.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. Rapid development of tts corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digit. Scholarsh. Humanit.*, 30:98–118.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. 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 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Marek Masiak, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2023. Afrimte and africomet: Empowering COMET to embrace under-resourced African languages. *arXiv preprint arXiv:2311.09828*.

Kimi Wenzel and Geoff Kaufman. 2024. Designing for harm reduction: Communication repair for multicultural users' voice interactions. In *Proceedings of the*

CHI Conference on Human Factors in Computing Systems, pages 1–17.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Appendix

A.1 Finetuning setup

For MT, we fine-tuned in both directions with a learning-rate of 2e-5 and batch size of 16. We trained for four epochs, and kept the model with the best eval loss. We used a weight decay of 0.01, warmup ratio 0.1, and a cosine annealing scheduler for learning rate. While for ASR finetuning, we fine-tuned with a learning-rate of 1e-3 and batch size of 8 for 20 epochs, as the validation WER continued to drop after preliminary runs with 10 epochs. For S2TT, we fine-tuned for 10 epochs with an optimal learning rate of 3e-4. All training was done on two NVIDIA A40 GPUs.

A.2 Results from Joint vs Individual MT fine-tuning

We present tables comparing jointly fine-tuning to individual fine-tuning on MT across the two training directions in Table 12 and Table 11.

A.3 Results from Joint vs Individual ASR

We present a table comparing jointly fine-tuning to individual fine-tuning on ASR across all models and dialects in Table 8 below.

Model	Standard	Ife	Ilaje
Zero-Shot	72.50	85.38	83.79
MMS-300m-Individual	74.67	93.20	78.24
MMS-1.3bn-Individual	**55.43**	**72.00**	61.80
XLSR-300m-Individual	56.26	81.23	64.22
XLSR-1.3bn-Individual	67.65	78.70	76.36
MMS-300m-Joint	58.11	76.58	67.17
MMS-1.3bn-Joint	55.73	73.95	63.94
XLSR-300m-Joint	54.55	73.72	**61.03**
XLSR-1.3bn-Joint	81.57	90.04	86.30

Table 8: ASR Performance of across all models after fine-tuning individually and jointly

A.4 Human evaluation

We provide exact instructions given to human evlautaors for our ASR and MT tasks in Table 5 and Table 6

Automatic Speech Recognition	You are tasked to evaluate the performance of an Automatic Speech Recognition (ASR) system on your native Yoruba dialect. This task involves assessing the accuracy and quality of transcriptions produced by this system when transcribing audio from a folder that will be provided to you. Your evaluations will help us understand how well these systems handle linguistic variations. Each filename has a corresponding audio file with the same name in the audio folder. Listen to the audio first, then look at the transcription from the model. Next, evaluate the quality of the transcription compared to the audio you listened to and provide a score in the Excel sheet. Please focus on the following key criteria while evaluating the transcriptions: **Fluency** Evaluate how natural and grammatically correct the transcription sounds in your dialect. 1 **Incomprehensible**: The transcription is completely unintelligible and nonsensical. The text is difficult to understand. 2 **Poor grammar and disfluent**: The transcription contains significant errors in grammar, syntax, and vocabulary that affect the clarity and naturalness of the text. 3 **Grammatically correct, potentially unnatural**: The transcription is grammatically correct but may have some errors in spelling, word choice, or syntax. 4 **Fluent and natural**: The transcription contains no grammatical errors, and the text is somewhat easy to read and understand. 5 **Perfectly fluent and natural**: The transcription is completely natural, grammatically flawless, reading as if written by a native speaker. **Adequacy** Assess how accurately the transcription conveys the meaning of the source speech. 1 **Nonsense/No meaning preserved**: All information is lost between the transcription and the source. 2 **Very poor meaning preservation**: The transcription preserves little meaning from the source. 3 **Moderate meaning preservation**: The transcription retains some meaning but still misses important details. 4 **Good meaning preservation**: The transcription retains most of the meaning of the source. 5 **Perfect meaning preservation**: The meaning of the transcription is completely consistent with the source.

Table 9: MT Human evaluation guidelines

Machine Translation	You are tasked to evaluate the performance of two Machine Translation systems on your native Yoruba dialect. This task involves assessing the accuracy and quality of translations produced by these systems, when translating from your dialect into English. Your evaluations will help us understand how well these systems handle linguistic variations. Please focus on the following key criteria while evaluating the transcriptions: **Fluency** Evaluate how natural and grammatically correct the translation sounds in the target language. 1 **Incomprehensible**: The translation is completely unintelligible and nonsensical. The text is difficult to understand. 2 **Poor grammar and disfluent**: The translation contains significant errors in grammar, syntax, and vocabulary that affect the clarity and naturalness of the text. 3 **Mostly grammatically correct, potentially unnatural**: The translation has few grammatical errors and also has some errors in spellings, word choice, or syntax. The language may not be natural. 4 **Grammatically correct and natural**: The translation contains few grammatical errors, the vocabulary is precise, and the text is easy to read and understand. 5 **Perfectly fluent and natural**: The translation is completely fluent, sounds natural and is grammatically correct. **Adequacy** Assess how accurately the transcription conveys the meaning of the source speech. 1 **Nonsense/No meaning preserved**: All information is lost between the translation and the source. 2 **Very poor meaning preservation**: The translation preserves little meaning from the source. 3 **Moderate meaning preservation**: The translation retains some meaning but still misses important details. 4 **Good meaning preservation**: The translation retains most of the meaning of the source. 5 **Perfect meaning preservation**: The meaning of the translation is completely consistent with the source.

Table 10: ASR Human evaluation guidelines

	BLEU ↑				AfriCOMET ↑			
	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard
Individual	16.53	16.04	12.98	30.27	0.57	0.56	0.52	0.69
Joint	21.98	22.97	18.64	37.55	0.60	0.59	0.55	0.71
Joint + MENYO-20k	19.80	20.77	17.21	31.75	0.54	0.59	0.60	0.71

Table 11: MT Finetuning Evaluation using NLLB-600M in the Yorùbá to English direction, training the dialects as individual languages, jointly under Yorùbá, and jointly along with MENYO-20k data.

	BLEU ↑				AfriCOMET ↑			
	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard	Ìjẹ̀bú	Ifẹ̀	Ìlàjẹ	Standard
Individual	8.48	8.74	5.78	18.32	0.52	0.50	0.47	0.66
Joint	8.71	8.93	6.48	18.98	0.52	0.50	0.47	0.66
Joint + MENYO-20k	7.23	7.25	5.29	17.24	0.50	0.48	0.44	0.65

Table 12: MT Finetuning Evaluation using NLLB-600M in the English to Yorùbá direction.