Evaluating the Effect of Linguistic Relatedness on Cross-Lingual Transfer in Large Multilingual Automatic Speech Recognition

Andrei Florian[1] **Cynthia Jayne Amol**[2] **Hope Kerubo Ombaba**[2] **Xiaoyu Cui**[1] **Boniface Mwau**[2]
Biatus Maina Kamau[2] **Lilian Diana Awuor Wanzare**[2] **Christiane Fellbaum**[1] **Happy Buzaaba**[1]

[1]Princeton University [2]Maseno University

Abstract

Extending automatic speech recognition (ASR) to low-resource African languages is constrained by the prohibitive demands of data collection at scale. A promising direction is to leverage linguistic relatedness to enhance cross-lingual transfer from a related auxiliary language to the low-resource target by sequentially adapting on both. Although this strategy has shown meaningful improvements in small ASR models, its effectiveness in large ASR remains unclear. We extend this framework to large multilingual ASR through a systematic controlled experimental design spanning six factors, two Africa-centric corpora, and four large ASR models, isolating whether linguistic relatedness reliably predicts cross-lingual transfer gains in this setting. Across all conditions, pre-adaptation on related auxiliary languages yields no practically meaningful transfer improvements given minimal target-language data, suggesting that linguistic relatedness alone may not reliably predict cross-lingual transfer gains in large multilingual ASR, or constitute an effective strategy for extending such models to low-resource languages.

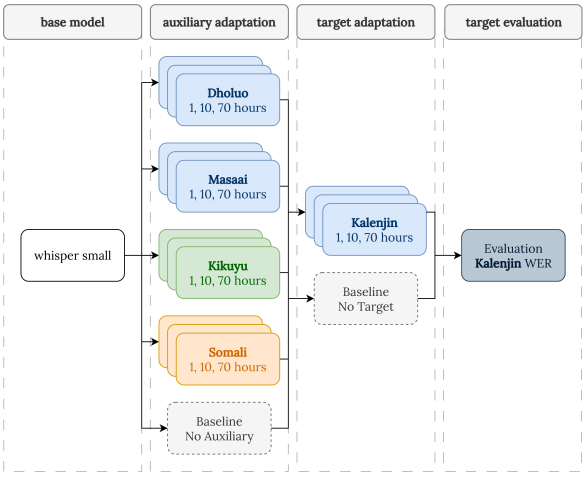

Figure 1: Experimental design for the first factor using languages from the AfriVoices KE corpus (Wanzare et al., 2026). Instances of the base Whisper Small (Radford et al., 2022) model are fine-tuned on each of four auxiliary languages, two linguistically related and two unrelated, at 1, 10, and 70 hours of labelled speech. The resulting auxiliary models, alongside the original baseline, are subsequently fine-tuned on 1, 10, and 70 hours of the target language Kalenjin, and evaluated using the Word Error Rate (WER) metric on the Kalenjin test set. Colours denote language family membership: blue (Nilotic), green (Bantu), orange (Cushitic).

1 Introduction

Automatic Speech Recognition (ASR) holds particular promise for extending language technologies to communities whose languages are primarily oral, as is the case for the majority of Africa's approximately 2,000 languages (Orife et al., 2020). Yet African languages remain profoundly underrepresented in contemporary NLP systems. Joshi et al. (2020) estimate that 88% of the world's languages have no meaningful presence in the datasets underpinning modern language models, a category encompassing most languages spoken across the African continent, and although recent large-scale models such as OpenAI's Whisper (Radford et al., 2022) and Meta's Omnilingual ASR (Keren et al., 2025) have extended ASR capabilities across increasingly larger language sets, the vast majority of African languages remain systematically absent from their training corpora, leading to poor downstream performance.

The root of this deficit lies in the nature of the data collection problem itself. Many African languages are primarily oral, with limited written materials and an even smaller fraction of those digitally available (Mbogho et al., 2025). As such, the web-scraping paradigms that enable large-scale corpus construction for high-resource languages do not transfer to this setting. Instead, building corpora for low-resource African languages typically

*Correspondence to Andrei Florian and Happy Buzaaba {andrei.florian, happy.buzaaba}@princeton.edu.

requires manual elicitation from local speakers, followed by careful transcription and linguistic annotation (Nekoto et al., 2020). At the scale demanded by modern ASR training, this process is neither financially nor logistically viable for extending ASR models to the hundreds of African languages lacking digital resources (Mbogho et al., 2025; Nekoto et al., 2020).

This challenge motivates the need for strategies that reduce the volume of target-language data required for effective ASR adaptation in low-resource settings. A promising direction is to leverage cross-lingual transfer, the tendency of multilingual language models to form language-agnostic internal representations that allow knowledge acquired in one language to generalise to others. Conneau et al. (2020) and Artetxe et al. (2020) demonstrate that large multilingual models develop shared representational subspaces across languages. Radford et al. (2022) note that Whisper similarly learns cross-lingual representations through joint multilingual training, and Babu et al. (2021) demonstrate the same capacity in wav2vec2 models, suggesting that this capacity for transfer extends to multilingual ASR models.

A substantial body of work further finds that these cross-lingual representations are especially well-aligned among linguistically related languages, where shared syntactic, morphological, and phonological structure provides a richer basis for transfer. Wu and Dredze (2019), Lauscher et al. (2020), Lin et al. (2019), and Eronen et al. (2023) collectively establish that genealogical and typological proximity reliably predicts zero-shot cross-lingual transfer performance in text-based multilingual models. Lauscher et al. (2020) and Wu and Dredze (2020) demonstrate that multilingual models whose pre-training corpora include languages related to the target can achieve comparable downstream performance with substantially less target-language fine-tuning data than would otherwise be required. Khare et al. (2021) and Nowakowski et al. (2023) extend this paradigm to small-scale ASR, demonstrating on wav2vec 2.0 that sequentially fine-tuning on a linguistically related high-resource auxiliary language before adapting to the low-resource target yields meaningful improvements over direct target-only adaptation.

We extend this paradigm to large multilingual ASR, applying the two-stage sequential fine-tuning methodology of Nowakowski et al. (2023) and Khare et al. (2021) to the 244-million-parameter Whisper Small model across two Africa-centric multilingual corpora as shown in Figure 1. We employ a systematic controlled experimental design spanning six factors, varying the auxiliary-target language pairing, fine-tuning methodology, and language corpus, assessing the effect of pre-adapting on multiple linguistically related auxiliary languages, and extending the evaluation to three additional large ASR models of varying scale and architecture. In doing so, we assess whether linguistic relatedness constitutes a reliable predictor of cross-lingual transfer effects in large multilingual ASR, with the central objective of evaluating the feasibility of relatedness-guided adaptation as a strategy for extending ASR capabilities to low-resource African languages.

2 Language Corpora and Models

We draw on five languages from each of two multilingual, Africa-centric ASR corpora: AfriVoices KE (Wanzare et al., 2026) and Google WaxalNLP (Diack et al., 2026).* Each set of five languages spans three families — Nilo-Saharan Nilotic, Niger-Congo Bantu, and Afro-Asiatic Cushitic — spoken across Central and East Africa. Both corpora were compiled through analogous processes in which scripted and free-speech recordings from recruited native speakers were transcribed and annotated by teams of trained linguists (Wanzare et al., 2026; Diack et al., 2026). Languages are selected to reflect meaningful typological contrast between within-family and cross-family pairings across syntactic, morphological, and phonological dimensions, as characterised in the subsections below. This ensures that family membership tracks multi-dimensional linguistic similarity, constituting a principled proxy for relatedness in the experimental design. Full dataset statistics for all ten languages across both corpora are provided in Table 1.

2.1 Typological Overview

Across the languages from the AfriVoices KE corpus (Wanzare et al., 2026), Dholuo, Maasai, and Kalenjin belong to the Nilotic family, sharing agglutinative morphology, where morphemes with different grammatical functions attach to a root to modify its meaning, and lexical tone, to distin-

*AfriVoices KE is released under CC BY 4.0 and Google WaxalNLP under CC BY 4.0 and CC-BY-SA 4.0.

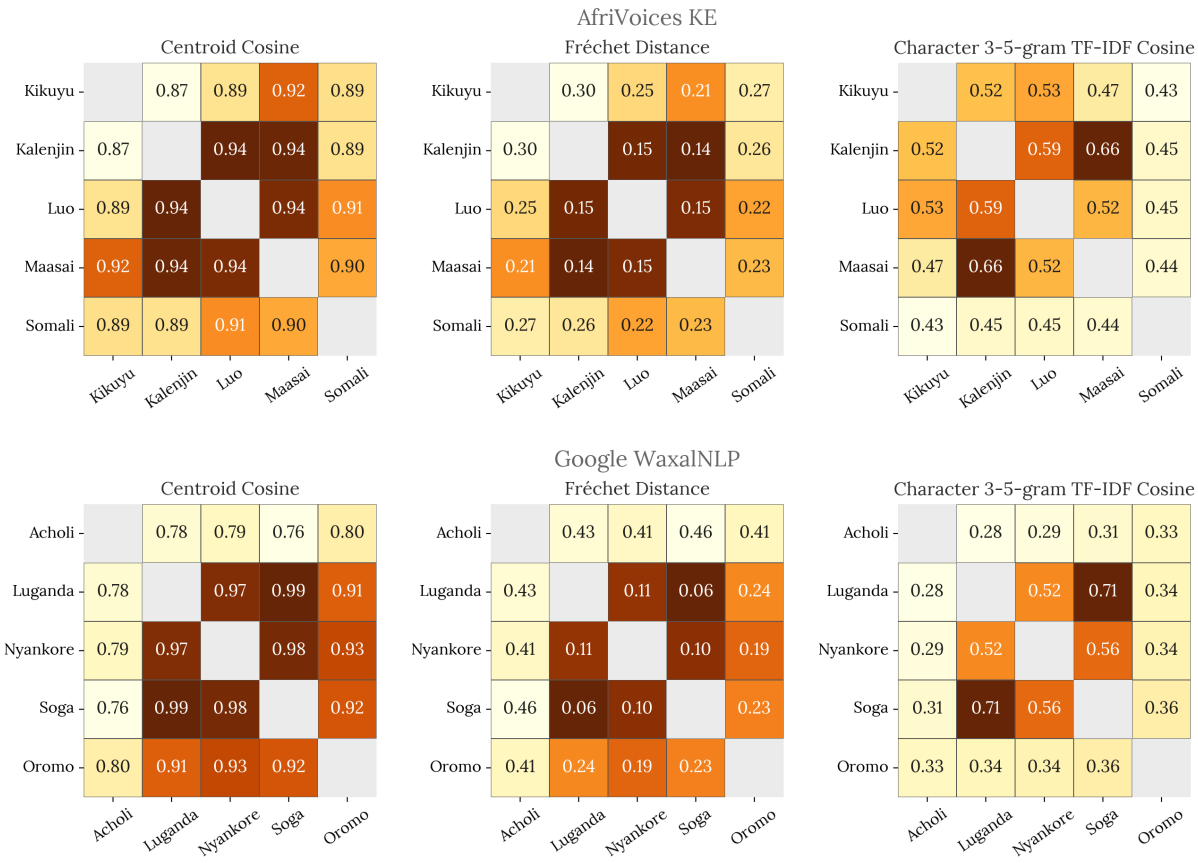

Figure 2: Corpus-level pairwise similarity results across three metrics for all languages in the AfriVoices KE (top) and Google WaxalNLP (bottom) corpora. Centroid cosine similarity is uniformly high across most language pairs in both corpora. Fréchet distance and character 3–5-gram TF-IDF cosine similarity both exhibit clear within-family clustering across both corpora.

guish different meanings of an otherwise identical root. Furthermore, Dholuo and Kalenjin show Subject-Verb-Object (SVO) word order, which is one point of distinction from the Verb-Subject-Object (VSO) order of Maasai. Although Kikuyu, a Bantu language, shares lexical tone and agglutinative morphology with the Nilotic family, it is distinct from this group through morphemes indicating noun class membership being prefixed to the root rather than added as suffixes to the end. Somali, a Cushitic language, also differs from the Nilotic set in that it has Subject-Object-Verb (SOV) syntax, relies on pitch accent, rather than tone, to distinguish word meanings, and its sound inventory includes phonemes found in Semitic languages.

The Google WaxalNLP corpus (Diack et al., 2026) spans 27 languages, of which we select five representing the same three families. From the Bantu family, we choose Luganda, Nyankole, and Soga, which are agglutinative, show lexical tone, SVO order, and prefixed noun class morphemes. Acholi represents the Nilotic language family in this corpus and similarly has SVO syntax, although differing from the Bantu set in its morphosyntactic properties and complex tonal system. The Cushitic family is represented by Oromo, which is most distinct from the Bantu and the Nilotic languages through its SOV order and use of pitch accent.

The patterns described above suggest that these syntactic, morphological, and phonological characteristics cluster with family membership, though the low-resource languages considered here lack the extensive linguistic documentation required for a fully rigorous account. The corpus-level pairwise similarity analyses that follow complement them by providing quantitative proxies for typological contrast across multiple dimensions.

2.2 Corpus-Level Analysis

We perform three pairwise similarity analyses across all languages in both corpora to assess cross-corpus comparability, and further validate the language groupings underpinning the experimental design. Prior to the analysis, we remove recordings

longer than 30 seconds, annotation tags, and utterances with fewer than 3 tokens or greater than the 99th-percentile length.

We focus on transcript-based measures because the ten languages use largely phonemic orthographies at the segmental level, making textual similarity a reasonable proxy for phonological and morphological similarity. Several languages in both corpora employ lexical tone or pitch accent as phonologically distinctive features not captured by the transcriptions, though the within-family clustering we observe remains interpretable as reflecting genuine segmental proximity. Audio-based measures were additionally considered, but pitch and formant features proved insufficiently stable across corpora to support reliable comparison.

Centroid cosine similarity encodes transcripts using the paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019) sentence transformer and computes cosine similarity between per-language mean embeddings. We use this primarily as a corpus-comparability check, since high centroid similarity indicates topically comparable corpora. Fréchet distance compares both the mean and covariance of per-language transcript embedding distributions (Heusel et al., 2017; Pillutla et al., 2021), capturing distributional shape and dispersion beyond mere topic proximity, which we interpret as a probe of morphosyntactic and lexico-semantic similarity. On the other hand, character n-gram TF-IDF cosine similarity captures recurring subword and surface-form patterns, providing a proxy for orthographic, lexical, morphological, and, to a limited extent, phonological overlap, characteristics particularly informative for the morphologically rich, agglutinative languages considered (Bojanowski et al., 2017).

As shown in Figure 2, absolute centroid cosine similarity is uniformly high across most language pairs in both corpora, indicating broad comparability in topic and elicitation structure. This reduces the risk that transfer effects are influenced by corpus-topic mismatch. By contrast, Fréchet distance and character n-gram TF-IDF similarity exhibit clear within-family clustering: in AfriVoices KE, the Nilotic languages are consistently closer to one another than to Kikuyu or Somali, while in WaxalNLP, Luganda, Nyankole, and Soga cluster more tightly together than with Acholi or Oromo.

Together, these analyses indicate cross-corpus topical comparability and corroborate meaningful typological contrast between within-family and cross-family pairings across multiple textual and embedding-based measures, lending validity to the family-based language groupings as a principled proxy for multidimensional relatedness in the experimental design.

2.3 Models

We extend the sequential fine-tuning setup of (Khare et al., 2021) and (Nowakowski et al., 2023) to four large ASR models spanning a range of parameter scales, supervision paradigms, and architectural families, with OpenAI Whisper Small (Radford et al., 2022) as the experimental base. Whisper Small is a 244-million parameter encoder–decoder Transformer trained on 680,000 hours of weakly supervised multilingual audio–text pairs across 99 languages, noted for its effectiveness in low-resource adaptation (Radford et al., 2022) and fine-tuning efficiency due to its relatively modest size. Whisper Large v3 (Radford et al., 2022) scales the same architecture to 1,550 million parameters across 32 encoder and decoder layers, trained on an expanded 5 million hours of weakly labelled audio. This model pairing allows us to isolate the effect of model scale on cross-lingual transfer while controlling for architecture and supervision regime.

Facebook XLS-R (Babu et al., 2021) and Facebook Omnilingual ASR (Keren et al., 2025) further vary both architecture and supervision regime, employing a self-supervised encoder-only wav2vec 2.0 backbone decoded via CTC rather than an autoregressive encoder–decoder. XLS-R comprises 300 million parameters trained on 436,000 hours of unlabelled speech across 128 languages, which OmniASR scales to 6.5 billion parameters and 4.3 million hours spanning around 1,600 languages.

All four models maintain a single unified parameter set shared across languages, permitting fine-tuning without architectural modification. Of the languages they are fine-tuned on in our experiments, both Whisper and XLS-R models have pre-training coverage of exclusively Somali, while OmniASR has pre-exposure to all AfriVoices KE languages except for Maasai (Radford et al., 2022; Babu et al., 2021; Keren et al., 2025).

3 Methods

We adopt the two-stage sequential fine-tuning methodology of Nowakowski et al. (2023) and Khare et al. (2021), which has been shown to yield

cross-lingual transfer gains from pre-adaptation on linguistically related auxiliary languages in small ASR models. We apply this methodology through a systematic controlled experimental design spanning six factors to isolate whether this positive effect generalises to large multilingual ASR.

We begin by filtering out all recordings exceeding 30 seconds in length from languages across both the AfriVoices KE and WaxalNLP corpora, as this is the maximum input length supported by Whisper models (Radford et al., 2022). All remaining data is partitioned into train (85%), test (10%), and validation (5%) sets for each language.

First Factor. We sample 1, 10, and 70 hours from the train set of each auxiliary language in the AfriVoices KE corpus, with 70 hours as the upper bound determined by the smallest per-language training set after filtering. Dholuo and Maasai constitute the related auxiliary set, Kikuyu and Somali the unrelated auxiliaries, and Kalenjin the target language. In the first stage, we full-model fine-tune a separate Whisper Small instance on each combination of auxiliary language and data volume, yielding 12 auxiliary models as shown in Figure 1. In the second stage, each auxiliary model and the unmodified Whisper Small baseline are further fine-tuned on 1, 10, and 70 hours of Kalenjin, producing 39 target models. All auxiliary and target models are evaluated on the Kalenjin test set.

Second Factor. We repeat the setup of the first factor, replacing full-model fine-tuning with parameter-efficient LoRA for both fine-tuning stages, to assess whether our results remain consistent across varying fine-tuning methodologies.

Third Factor. We follow the first factor setup with the roles of Dholuo and Kalenjin exchanged, making Dholuo the target language and Kalenjin an auxiliary, to assess whether our results hold under a different within-family auxiliary-target pairing.

Fourth Factor. Mirroring the setup of the first factor, we introduce an additional, intermediate fine-tuning stage, adapting the model on two related auxiliary languages — Dholuo followed by Maasai — prior to target-language fine-tuning on Kalenjin, to assess whether pre-adaptation on multiple linguistically related auxiliary languages affects cross-lingual transfer.

Fifth Factor. We sample 1, 10, and 25 hours from the train split of each language in the WaxalNLP corpus and repeat the first factor setup with Nyankole and Soga (Bantu) constituting the related auxiliary set, Acholi (Nilotic) and Oromo (Cushitic) the unrelated auxiliaries, and Luganda the target language, to assess whether our results generalize across corpora and language family groupings.

Sixth Factor. We reproduce the first factor setup across three additional models — Whisper Large v3, XLS-R, and OmniASR — to assess whether our findings generalize across large ASR models of varying scale, architecture, and supervision paradigm.

Across all factors, we report Word Error Rate (WER) as the primary evaluation metric for all auxiliary and target models. We compare the per-utterance mean WER of the group of models pre-adapted on a related language against the target-only baseline at each target-data volume across all factors, testing whether pre-adaptation on a related language improves target performance over direct target-only fine-tuning. We conduct an additional analysis, comparing per-utterance mean WER between models pre-adapted on related and unrelated auxiliary languages, testing whether linguistic relatedness between the auxiliary and target improves transfer gains over unrelated pre-adaptation. Together, these analyses isolate whether any observed transfer gains are directly attributable to linguistic relatedness. Both analyses employ a Wilcoxon signed-rank test on utterance-level mean differences alongside a non-parametric two one-sided tests (TOST) procedure (Lakens, 2017). Given the large sample ($n \approx 7{,}000$ test utterances per factor), Wilcoxon achieves near-universal significance. TOST serves as the definitive performance equivalence criterion, with an equivalence bound of $\Delta = 5$ absolute WER percentage points, an approximate threshold beyond which we consider practical utility to be materially affected. We characterise a difference as *practically equivalent* if TOST establishes equivalence and *practically meaningful* otherwise. Rank-biserial correlation r is reported as a complementary effect-size measure, with its sign indicating the direction of any difference and its magnitude reflecting the consistency with which one group outperforms the other (full statistical results in Tables 4, 5).

Full-model fine-tuning. All models are fine-tuned on a single NVIDIA H200, with a total computational budget of approximately 650 GPU hours across all experimental conditions. We use a cosine learning-rate schedule, 0.01 weight decay, and mixed precision (bf16/TF32). Hyperparameters are tuned via a small grid over learning rate and number

of epochs, with settings scaled to data volume. The final checkpoint is selected by minimum validation WER under greedy decoding, using early stopping with a patience of four evaluations. The Whisper models are fine-tuned using HuggingFace's *Seq2SeqTrainer* with an effective batch size of 16 and gradient checkpointing enabled. XLS-R is fine-tuned using the standard *Trainer* with an effective batch size of 32 and gradient checkpointing disabled, given its smaller memory footprint at 300M parameters. OmniASR is fine-tuned via a custom *CTCTrainer* with an effective batch size of 16 and gradient checkpointing disabled, as the *fairseq2* interface does not expose gradient checkpointing. The Whisper and OmniASR models use 8-bit AdamW to reduce optimizer state memory and XLS-R uses fused AdamW, sufficient at its 300M parameter footprint.

LoRA fine-tuning. LoRA fine-tuning is only applied to Whisper Small in the second factor. We apply standard flat LoRA (Hu et al., 2022) to five module types: *q_proj*, *v_proj*, *out_proj*, *fc1*, and *fc2*, targeting both attention and feed-forward sublayers following evidence that FFN modules receive disproportionately high importance scores during low-resource cross-lingual ASR adaptation (Zhang et al., 2023; Song et al., 2024). The first three encoder layers are frozen throughout, with adapters applied to encoder layers 3–11 and all decoder layers (Liu et al., 2024). Rank and learning rate are calibrated to fit fine-tuning data volume and LoRA alpha is fixed at 16 across all conditions, with adapter dropout at 0.05.

4 Results

In the first factor, WER on the Kalenjin test set decreases monotonically with additional hours of target-language fine-tuning across all auxiliary models and the unmodified Whisper Small baseline, as shown in the left panel of Figure 3, consistent with expected adaptation rates for Whisper Small (Radford et al., 2022). Prior to second-stage Kalenjin adaptation, models pre-adapted on related languages exhibit lower Kalenjin WER than the baseline, as shown in the right panel of Figure 3, a practically meaningful pre-adaptation advantage. However, from one hour of Kalenjin fine-tuning onward, models pre-adapted on a related-language, an unrelated language, and the Whisper Small baseline converge to a narrow band, with Kalenjin performance practically equivalent across all groups.

Replacing full-model fine-tuning with parameter-efficient LoRA across both fine-tuning stages produces qualitatively identical results, shown in the top-left panel of Figure 4. Prior to second-stage adaptation, all models exhibit greater variance in Kalenjin WER under LoRA than under full fine-tuning. Nonetheless, differences among auxiliary groups prior to Kalenjin fine-tuning are practically equivalent regardless of relatedness, though related auxiliary models again retain a practically meaningful advantage over the unmodified baseline. Similarly to the first factor, all models converge to a narrow performance band beginning at one hour of Kalenjin fine-tuning, with target performance practically equivalent across groups.

Exchanging the roles of Dholuo and Kalenjin — making Dholuo the target and Kalenjin an additional auxiliary — produces broadly the same pattern, as shown in the top-right panel of Figure 4. Prior to second-stage adaptation, related auxiliary models show a practically meaningful advantage over both unrelated auxiliaries and the baseline. From one hour of Dholuo fine-tuning onward, performance is practically equivalent across related and unrelated auxiliary groups. The group of related auxiliary models retains a practically meaningful advantage over the baseline at one hour before all groups converge from ten hours onward.

Pre-adapting sequentially on two linguistically related auxiliary languages — Dholuo followed by Maasai, both Nilotic — yields a practically meaningful WER advantage on the Kalenjin test set relative to both the Factor 1 unrelated auxiliaries and the Factor 1 baseline prior to any second-stage target adaptation, as shown in the bottom-left panel of Figure 4. From one hour of Kalenjin fine-tuning onward, however, performance is practically equivalent across all groups.

In the fifth factor, using the WaxalNLP corpus with Luganda as the target language, models pre-adapted on the linguistically related Bantu auxiliaries (Nyankole and Soga) show a practically meaningful advantage over both the unrelated auxiliaries (Acholi and Oromo) and the baseline prior to second-stage adaptation, as shown in the bottom-right panel of Figure 4. From one hour of Luganda fine-tuning onward, performance is practically equivalent across all groups.

The same pattern generally holds across all three additional model architectures — Whisper Large v3, XLS-R, and OmniASR — as shown in Figure 5. Prior to second-stage adaptation, Whis-

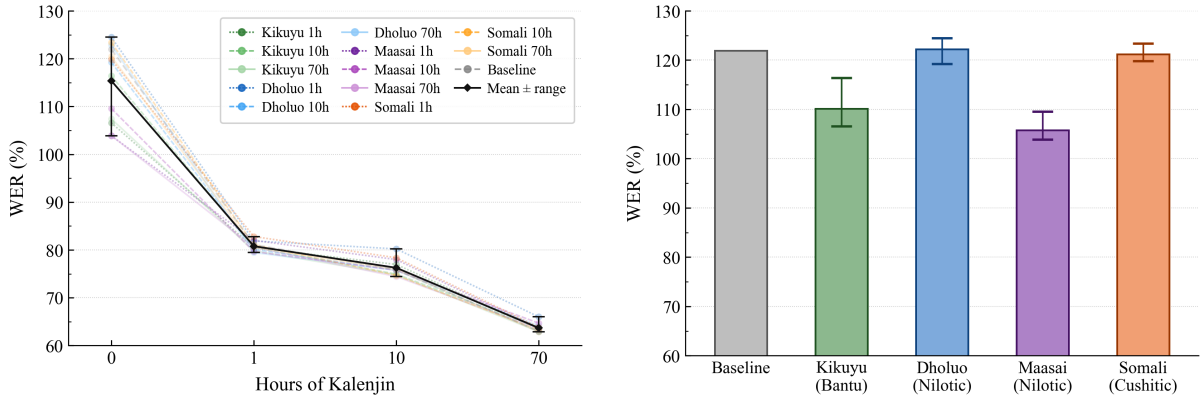

Figure 3: WER on the Kalenjin test set across all first-factor models. *Left*: target-language WER as a function of Kalenjin fine-tuning hours for all 12 auxiliary models and the unmodified baseline. The mean WER and range across auxiliary models at each volume of Kalenjin adaptation are marked in black. *Right*: Mean WER of each auxiliary model on the Kalenjin test set prior to second-stage fine-tuning, with error bars denoting the range across auxiliary language data volumes.

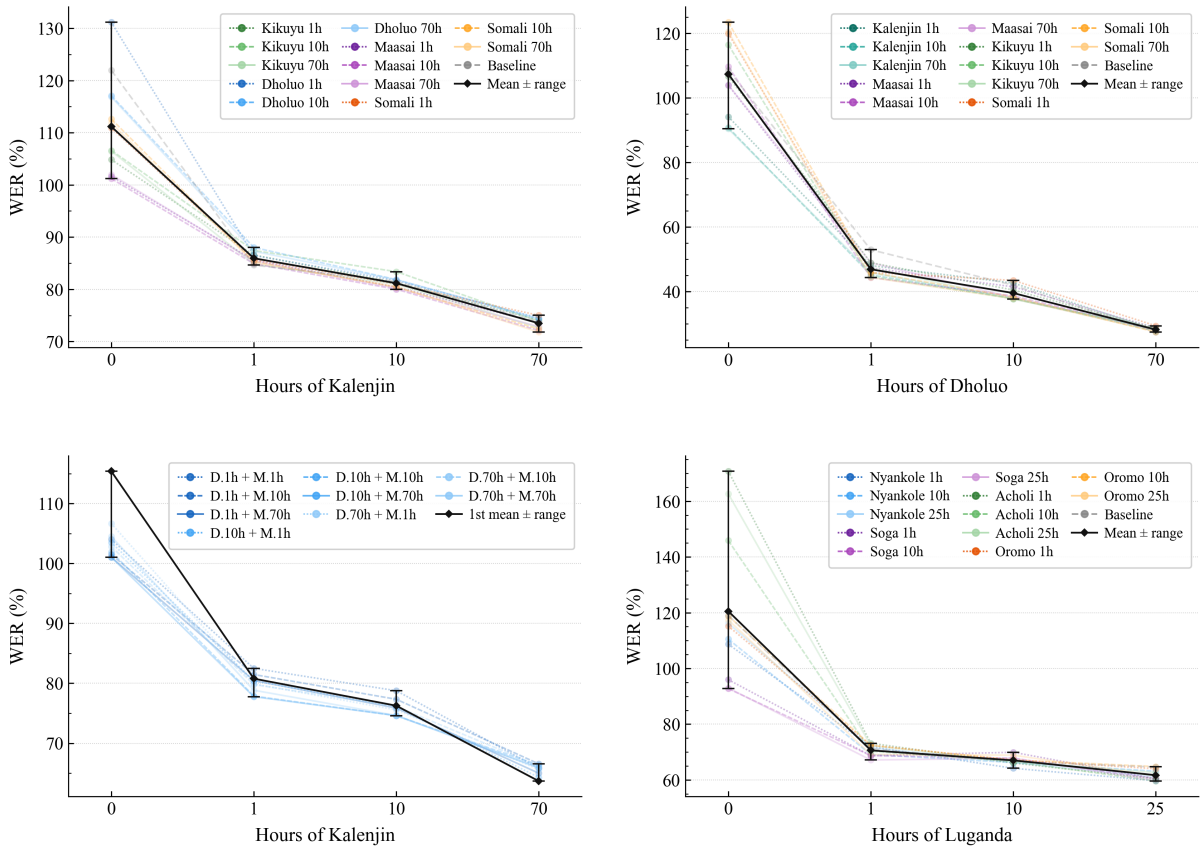

Figure 4: WER on the target language test set as a function of target-language fine-tuning hours across factors two through five. *Top left*: factor two, replicating the first-factor setup with LoRA fine-tuning. *Top right*: factor three, with Dholuo as the target language and Kalenjin reassigned to the auxiliary set. *Bottom left*: factor four, in which the model is pre-adapted sequentially on Dholuo and then Maasai prior to Kalenjin fine-tuning; in this plot, the black reference line denotes the cross-model mean from factor one. *Bottom right*: factor five, using the WaxalNLP corpus with Luganda as the target language and Nyankole, Soga, Acholi, and Oromo as auxiliaries. Colours denote language family and the cross-model mean and range across auxiliary data volumes are marked in black.

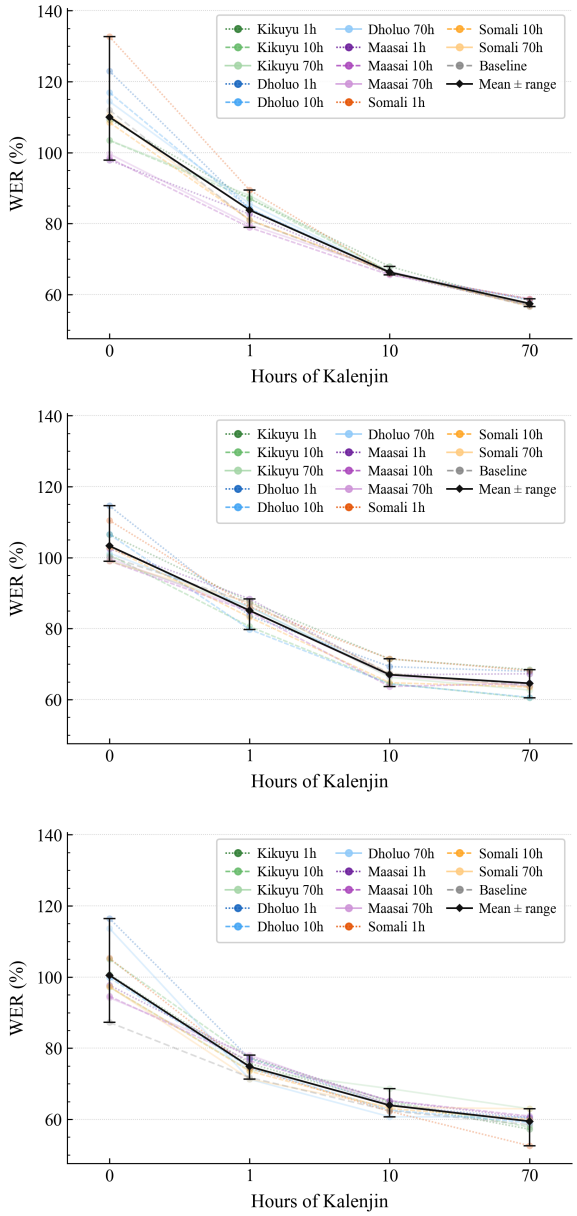

Figure 5: WER on the Kalenjin test set as a function of fine-tuning hours across all auxiliary conditions for Whisper Large v3 (top), XLS-R (centre), and OmniASR (bottom). Colours denote language family. Cross-model mean and range across auxiliary data volumes in black.

per Large v3 shows a practically meaningful pre-adaptation advantage over the baseline and OmniASR a practically meaningful disadvantage, while XLS-R shows no practically meaningful difference. Across all models, performance is practically equivalent from one hour of Kalenjin fine-tuning onward, a pattern that holds despite differences in prior exposure to the experiment languages across models.

Taken together, across all six experimental factors, all auxiliary and target data volumes, and all ASR models tested, pre-adaptation on a related auxiliary language yields no practically meaningful improvement in target language performance once second-stage adaptation begins. As little as one hour of target language fine-tuning is sufficient to establish practical equivalence across models with related, unrelated, or no pre-adaptation in all but one factor, with this sole exception reaching equivalence at ten hours (complete per-condition WER values in Table 2).

5 Conclusion

Our work evaluates whether linguistic similarity across genealogical, phonological, morphological, and syntactic dimensions can be leveraged to improve cross-lingual transfer in large multilingual ASR and reduce the volume of target-language data required to extend such models to low-resource languages, focusing on the African setting. We adopt the two-stage sequential fine-tuning framework of Khare et al. (2021) and Nowakowski et al. (2023) introduced in small-scale ASR, which leverages cross-lingual representations formed by multilingual models through pre-adaptation on a related auxiliary language prior to fine-tuning on the low-resource target, with the aim of improving target-language performance. We extend this paradigm to large multilingual ASR, employing a systematic controlled experimental design spanning six factors, two Africa-centric corpora, and four large ASR models of varying scales, architectures, and supervision paradigms to isolate whether linguistic relatedness between auxiliary and target constitutes a reliable predictor of cross-lingual transfer gains.

Across all six factors, both corpora, and all four models, we find no practically meaningful difference in target-language performance among models fine-tuned exclusively on the target language, pre-adapted on a related auxiliary language, and pre-adapted on an unrelated auxiliary language, from as little as one hour of target-language adaptation onward in all but one case. Our findings run counter to those of Khare et al. (2021) and Nowakowski et al. (2023), suggesting that linguistic relatedness alone may not reliably predict cross-lingual transfer gains in large multilingual ASR, or constitute a reliable strategy for extending such models to low-resource languages.

Limitations

Our study considers linguistic similarity across various genealogical, phonological, morphological,

and syntactic dimensions, which together capture several of the axes most commonly studied in relation to cross-lingual transfer (Lin et al., 2019; Eronen et al., 2023), though additional linguistic characteristics not explored in this work may also play a role. Relatedly, our experimental design necessarily relies on broad typological characterisation and corpus-level similarity analyses as proxies for linguistic relatedness in forming language groupings, given the limited rigorous linguistic resources available for the low-resource languages considered. More broadly, cross-lingual transfer is likely shaped by factors beyond linguistic relatedness alone — such as audio recording conditions, speaker demographics, and domain — that fall outside of this work's purview.

Our conclusions assume an equivalence bound of $\Delta = 5$ WER points as the threshold for practical meaningfulness, following the rationale outlined in Section 3. Our experimental design spans four large ASR models and two Africa-centric corpora representing three language families; broader coverage of ASR models and corpora across a wider range of African geographies and language families, alongside replication in other low-resource settings, would further strengthen the generalisability of our findings.

Ethics Statement and Broader Impact

Our work supports ASR development for low-resource African languages by studying whether linguistic relatedness improves cross-lingual transfer in large multilingual ASR, with the aim of informing more principled adaptation strategies for this setting. All experiments use the publicly available AfriVoices KE (Wanzare et al., 2026) and Google WaxalNLP (Diack et al., 2026) datasets compiled with the involvement of native speaker communities. Neither dataset contains private user data and all results are reported at the language levels. Both corpora are used solely for research purposes, consistent with their intended use and licensing terms. The accompanying code is released for research use only. We do not anticipate direct harms arising from this research.

Use of AI Assistance

Claude Code and Claude (Anthropic) were used to assist in writing and testing code for data processing, model fine-tuning, evaluation, and data analysis. All generated code was manually reviewed and verified for correctness. Claude and Elicit AI were used to support the literature review process. All scientific content, experimental design, analysis, and conclusions are entirely the authors' own.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. Association for Computational Linguistics.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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. Association for Computational Linguistics.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, and others. 2026. WAXAL: A large-scale multilingual African language speech corpus.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

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 *Proc. Interspeech 2021*, pages 1529–1533. ISCA.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4483–4499. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135. Association for Computational Linguistics.

Yunfei Liu, Xiaojia Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Audrey Mbogho, Quin Awuor, Andrew Kipkebut, Lilian Wanzare, and Vivian Oloo. 2025. Building low-resource African language corpora: A case study of Kidawida, Kalenjin and Dholuo. *Journal of the Digital Humanities Association of Southern Africa (RAIL)*, 6(2).

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, and others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuwazny. 2023. Adapting multilingual speech representation model for a new, under-resourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, and others. 2020. Masakhane – machine translation for Africa. Accepted at the AfricaNLP Workshop, ICLR 2020.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992. Association for Computational Linguistics.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-efficient and extensible multilingual ASR. In *Interspeech 2024*, pages 3934–3938. ISCA.

Lilian Wanzare, Cynthia Amol, Nelson Odhiambo, Hope Kerubo, Leila Misula, Vivian Oloo, Rennish Mboya, Edwin Onkoba, Edward Ombui, Joseph Muguro, and others. 2026. AfriVoices-KE: A multilingual speech dataset for Kenyan languages.

Shijie Wu and Mark Dredze. 2019. BETO, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 833–844. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130. Association for Computational Linguistics.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *Proceedings of the International Conference on Learning Representations*.

Corpus	Language	Family	Train (h)	Test (h)	Dev (h)
AfriVoices KE	Kalenjin	Nilotic	127.75	14.91	7.48
	Dholuo	Nilotic	123.74	14.45	7.23
	Maasai	Nilotic	75.21	8.85	4.42
	Kikuyu	Bantu	104.15	12.34	6.21
	Somali	Cushitic	92.15	10.88	5.48
Google WaxalNLP	Luganda	Bantu	26.11	3.05	1.55
	Nyankole	Bantu	37.53	4.36	2.20
	Soga	Bantu	33.17	3.86	1.93
	Acholi	Nilotic	25.08	2.94	1.48
	Oromo	Cushitic	188.68	22.30	11.17

Table 1: Dataset statistics for all languages across both corpora after filtering, showing language family membership and train/test/dev data volumes in hours across each language.

A Appendix

A Dataset Statistics

Table 1 reports the dataset statistics for all languages across both corpora after filtering recordings of a length greater than 30 seconds as outlined in Section 3. The data volumes sampled for auxiliary and target fine-tuning in each factor are capped at 70 hours for AfriVoices KE and 25 hours for WaxalNLP, corresponding to the smallest per-language training sets available after filtering: Maasai and Luganda respectively.

B WER Results

Table 2 and 3 reports WER across the full evaluation set for all experimental conditions underlying the figures and discussion in Section 4, with auxiliary model performance averaged across all auxiliary data volumes at each target fine-tuning step for clarity, given the narrow range across auxiliary volumes at each target-data level.

C Statistical Analysis

Tables 4 and 5 report the complete results of the two statistical analyses described in Section 3. Table 4 addresses whether linguistic relatedness between the auxiliary and target improves cross-lingual transfer, comparing related against unrelated auxiliary models at each volume of target-language data. Table 5 addresses whether auxiliary pre-adaptation improves target-language performance at all, comparing related auxiliary models against the target-only fine-tuning baseline.

Target	Model	None	1 hr	10 hrs	70 hrs
Kalenjin (FFT)	baseline	121.96	81.12	75.92	62.96
	luo (avg)	121.81	80.29	76.68	64.08
	mas (avg)	105.63	80.47	75.78	63.93
	kik (avg)	110.10	80.10	75.38	**63.31**
	som (avg)	122.15	81.09	76.03	63.44
Kalenjin (LoRA)	baseline	121.96	84.63	81.53	73.73
	luo (avg)	121.70	88.39	81.91	74.53
	mas (avg)	101.61	85.03	80.26	**72.65**
	kik (avg)	105.97	85.99	81.66	73.40
	som (avg)	111.79	85.33	80.63	72.88
Dholuo (FFT)	baseline	108.56	52.90	41.91	27.38
	kln (avg)	91.75	46.30	39.37	**27.96**
	mas (avg)	105.79	46.81	39.34	28.32
	kik (avg)	110.13	46.92	38.61	28.45
	som (avg)	121.22	45.32	39.81	28.11
Kalenjin (Seq.)	baseline	121.96	81.12	75.92	62.96
	luo→mas (avg)	103.00	79.89	75.96	65.76
Luganda (FFT)	baseline	118.94	71.20	66.64	64.64
	nyn (avg)	111.82	70.84	65.55	**61.31**
	sog (avg)	93.89	68.19	68.50	60.19
	ach (avg)	159.76	71.92	66.42	60.88
	orm (avg)	117.14	71.31	67.73	63.43

Table 2: WER results for Whisper Small across all conditions and target-language data volumes. Lower is better. *None* denotes pre-fine-tuning evaluation. FFT = full fine-tuning, LoRA = parameter-efficient fine-tuning, Seq. = dual sequential adaptation. Auxiliary rows report averages across auxiliary data volumes. Luganda is capped at 25 hours due to its smaller training set.

Target	Backbone	Model	None	1 hr	10 hrs	70 hrs
Kalenjin	Whisper Large	baseline	112.01	80.92	66.23	56.66
		luo (avg)	118.12	84.66	66.16	57.98
		mas (avg)	98.66	80.42	65.96	57.63
		kik (avg)	105.43	87.36	66.91	**57.05**
		som (avg)	117.29	**83.84**	**66.18**	57.47
Kalenjin	XLS-R	baseline	100.30	87.32	67.41	63.89
		luo (avg)	107.49	83.17	66.88	64.41
		mas (avg)	100.44	86.11	**65.80**	**65.18**
		kik (avg)	102.16	84.80	67.40	63.87
		som (avg)	104.13	**85.57**	67.84	65.08
Kalenjin	OmniASR	baseline	87.25	71.78	62.39	**58.49**
		luo (avg)	110.01	74.30	62.43	59.97
		mas (avg)	95.49	77.34	64.84	60.29
		kik (avg)	100.98	75.58	66.08	59.33
		som (avg)	99.97	**73.10**	**62.90**	58.50

Table 3: WER results for additional ASR backbones on Kalenjin across target-language data volumes. Lower is better. *None* denotes models evaluated before target-language fine-tuning. Baseline rows use the unmodified pretrained model, while auxiliary rows use models first adapted on the listed auxiliary language. Auxiliary rows report averages across auxiliary data volumes. Bold indicates best result per group per column.

Factor	Target (h)	n	Median Δ	r	TOST	p_{TOST}	p_W
F1	0	7,261	-1.56%	-0.168		3.21×10^{-88}	7.95×10^{-34}
	1	7,261	$+0.00\%$	-0.067		≈ 0	2.14×10^{-6}
	10	7,261	$+0.00\%$	$+0.074$		≈ 0	1.91×10^{-7}
	70	7,261	$+0.00\%$	$+0.085$		≈ 0	3.20×10^{-9}
F2	0	7,261	$+0.00\%$	$+0.136$		1.12×10^{-128}	8.09×10^{-22}
	1	7,261	$+0.00\%$	$+0.183$		8.46×10^{-292}	2.69×10^{-38}
	10	7,261	$+0.00\%$	-0.000		≈ 0	0.9745
	70	7,261	$+0.00\%$	$+0.082$		≈ 0	1.40×10^{-8}
F3	0	7,015	-14.92%	-0.957	$\times$	$1.000^{\dagger}$	≈ 0
	1	7,015	$+0.67\%$	$+0.180$		≈ 0	8.10×10^{-39}
	10	7,015	$+0.00\%$	-0.009		≈ 0	0.5511
	70	7,015	$+0.00\%$	-0.102		≈ 0	4.45×10^{-11}
F4	0	7,261	-12.22%	-0.905	$\times$	$1.000^{\dagger}$	≈ 0
	1	7,261	-0.69%	-0.132		1.02×10^{-262}	4.09×10^{-22}
	10	7,261	$+0.00\%$	$+0.014$		1.87×10^{-224}	0.3093
	70	7,261	$+2.08\%$	$+0.205$		3.69×10^{-69}	6.48×10^{-51}
F5	0	525	-28.70%	-0.939	$\times$	$1.000^{\dagger}$	1.02×10^{-76}
	1	525	-1.11%	-0.161		1.45×10^{-9}	0.0018
	10	525	-0.00%	-0.054		5.93×10^{-13}	0.2955
	25	525	$+0.00\%$	-0.093		1.82×10^{-10}	0.0708
F6 (Whisper Large)	0	7,261	-3.33%	-0.361		2.40×10^{-26}	1.09×10^{-145}
	1	7,261	-2.78%	-0.345		5.69×10^{-66}	9.10×10^{-133}
	10	7,261	$+0.00\%$	-0.068		≈ 0	2.17×10^{-6}
	70	7,261	$+0.00\%$	$+0.098$		≈ 0	1.12×10^{-11}
F6 (XLS-R)	0	7,261	$+0.00\%$	$+0.098$		5.17×10^{-304}	1.21×10^{-11}
	1	7,261	-1.67%	-0.136		≈ 0	2.57×10^{-23}
	10	7,261	-0.83%	-0.193		≈ 0	2.37×10^{-41}
	70	7,261	$+0.00\%$	$+0.075$		≈ 0	1.86×10^{-7}
F6 (OmniASR)	0	7,261	$+2.78\%$	$+0.432$		4.36×10^{-46}	5.13×10^{-208}
	1	7,261	$+1.11\%$	$+0.190$		≈ 0	3.99×10^{-41}
	10	7,261	-0.88%	-0.184		≈ 0	2.79×10^{-38}
	70	7,261	$+0.00\%$	$+0.159$		≈ 0	1.84×10^{-28}

Table 4: Per-utterance comparison of models pre-adapted on a linguistically related auxiliary language against those pre-adapted on an unrelated auxiliary, at each volume of target-language data. The equivalence criterion is TOST (Lakens, 2017) with an equivalence bound of $\Delta = 5$ WER percentage points: a row marked $\times$ indicates that equivalence was not established within this bound and the respective difference should be treated as practically meaningful. Rank-biserial correlation r reports effect size and direction: $r < 0$ indicates that related auxiliary models produce lower WER than unrelated ones, and $r > 0$ the reverse. The Wilcoxon p_W is included for completeness; near-universal significance throughout reflects the large per-factor test sample ($n \approx 7{,}261$ utterances) rather than practical effect magnitude. $p_{\mathrm{TOST}} = 1.000$ (†) indicates that the difference is strongly directional and well outside the equivalence bound. For Factor 4, models sequentially pre-adapted on Dholuo and Maasai are compared against the grouping of Kikuyu and Somali single-auxiliary models from Factor 1 on the shared Kalenjin test set.

Factor	Target (h)	n	Median Δ	r	TOST	p_{TOST}	p_W
F1	0	7,261	-7.50%	-0.429	$\times$	$1.000^{\dagger}$	2.50×10^{-214}
	1	7,261	$+0.00\%$	-0.062		1.02×10^{-142}	9.05×10^{-6}
	10	7,261	$+0.00\%$	$+0.057$		1.03×10^{-186}	5.12×10^{-5}
	70	7,261	$+0.00\%$	$+0.095$		2.94×10^{-200}	1.49×10^{-11}
F2	0	7,261	-10.00%	-0.495	$\times$	$1.000^{\dagger}$	3.84×10^{-280}
	1	7,261	$+1.56\%$	$+0.223$		9.06×10^{-73}	4.35×10^{-57}
	10	7,261	$+0.00\%$	-0.066		3.84×10^{-267}	3.53×10^{-6}
	70	7,261	$+0.00\%$	-0.020		3.01×10^{-235}	0.1617
F3	0	7,015	-5.14%	-0.532	$\times$	$1.000^{\dagger}$	≈ 0
	1	7,015	-3.89%	-0.332	$\times$	0.6773	1.74×10^{-127}
	10	7,015	-4.00%	-0.335		4.37×10^{-11}	2.63×10^{-130}
	70	7,015	-1.67%	-0.140		9.99×10^{-156}	4.06×10^{-24}
F4	0	7,261	-18.52%	-0.793	$\times$	$1.000^{\dagger}$	≈ 0
	1	7,261	-1.11%	-0.107		2.20×10^{-93}	8.10×10^{-15}
	10	7,261	$+0.00\%$	$+0.016$		1.57×10^{-126}	0.2466
	70	7,261	$+2.22\%$	$+0.206$		1.00×10^{-33}	5.06×10^{-50}
F5	0	525	-6.67%	-0.433	$\times$	$1.000^{\dagger}$	1.76×10^{-17}
	1	525	-1.11%	-0.011		1.16×10^{-6}	0.8266
	10	525	$+0.00\%$	$+0.004$		6.02×10^{-11}	0.9330
	25	525	$+0.67\%$	$+0.000$		4.27×10^{-8}	0.9999
F6 (Whisper Large)	0	7,261	-3.33%	-0.322	$\times$	$1.000^{\dagger}$	1.62×10^{-119}
	1	7,261	$+1.04\%$	$+0.092$		2.01×10^{-84}	3.77×10^{-11}
	10	7,261	$+0.00\%$	-0.012		2.58×10^{-144}	0.4053
	70	7,261	$+0.00\%$	$+0.110$		5.46×10^{-138}	5.78×10^{-15}
F6 (XLS-R)	0	7,261	$+2.47\%$	$+0.443$		4.83×10^{-17}	4.68×10^{-207}
	1	7,261	-2.63%	-0.253		2.39×10^{-29}	1.07×10^{-77}
	10	7,261	$+0.00\%$	-0.039		1.21×10^{-132}	0.0046
	70	7,261	$+0.00\%$	$+0.056$		1.89×10^{-123}	4.86×10^{-5}
F6 (OmniASR)	0	7,261	$+4.17\%$	$+0.360$	$\times$	0.6657	3.35×10^{-151}
	1	7,261	$+3.03\%$	$+0.318$		3.59×10^{-24}	7.66×10^{-116}
	10	7,261	$+1.28\%$	$+0.128$		9.42×10^{-136}	2.57×10^{-20}
	70	7,261	$+1.52\%$	$+0.187$		6.26×10^{-130}	5.83×10^{-41}

Table 5: Per-utterance WER comparison of models pre-adapted on a linguistically related auxiliary language against the target-only fine-tuning baseline, at each volume of target-language data. The primary criterion, effect size measure, and notation follow Table 4: $\times$ marks rows where equivalence was not established and $r < 0$ indicates that related auxiliary models produce lower WER than the baseline. For Factor 4, dual-Nilotic sequential models are compared against the Factor 1 baseline on the shared Kalenjin test set.