AFROSCOPE: A Framework for Studying the Linguistic Landscape of Africa

Anonymous ACL submission

Figure 1: Overview of the AFROSCOPE framework. **Dataset & Coverage** constructs AFROSCOPE-DATA by aggregating multilingual datasets and metadata on language families, scripts, and domains, while applying contamination analysis to support clean evaluation. **Model & Analysis** fine-tunes baseline models to produce AFROSCOPE-MODELS, introduces AFROSCOPE-MIRROR for hierarchical disambiguation of closely related languages, and studies cross-lingual transfer and robustness across African languages.

Abstract

Language Identification (LID), the task of determining the language of a given text, is a fundamental preprocessing step that shapes the reliability of downstream NLP applications. While recent work has expanded African LID, existing systems remain limited in both language coverage and fine-grained discrimination among closely related languages and varieties. We introduce AFROSCOPE, a unified framework for African LID that includes AFROSCOPE-DATA, a dataset covering 640 languages, and AFROSCOPE-MODELS, a suite of strong LID models with broad African language coverage. To address persistent confusions among closely related languages, we propose a hierarchical classification approach that leverages AFROSCOPE-MIRROR, a specialized embedding model for targeted disambiguation. This approach improves macro-F_1 by 1.57 points on the confusable subset compared to our best base model. We further analyze cross-lingual transfer and domain effects, showing how language-family structure, script compatibility, and domain coverage shape LID performance. We position African LID as an enabling technology for large-scale measurement of Africa's linguistic landscape in digital text and release AFROSCOPE-DATA and AFROSCOPE-MODELS online.[1]

1 Introduction

Language Identification (LID), the task of determining the language of a given text, is a foundational step in curating multilingual corpora from web crawls (Penedo et al., 2025; Foroutan et al., 2025). LID errors propagate to downstream stages such as tokenization (Duvenhage et al., 2017), filtering (Grattafiori et al., 2024; Li et al., 2024), and data scheduling for multilingual pretraining (Conneau et al., 2020; de Gibert et al., 2024; Laurençon et al., 2022). Crucially, LID systems determine not only how reliably each language is predicted but also the *scope* of identifiable languages. If a language is out of scope, its text is either dropped or misattributed to an in-scope language, distorting corpus composition and downstream evaluation (Costa-Jussà et al., 2022; Adebara et al., 2022a).

[1] Anonymous_link

Although LID is often treated as largely solved (Mcnamee, 2005; Blodgett and O'Connor, 2017), recent evaluations show that performance remains uneven across languages and domains (Ojo et al., 2025; Suarez et al., 2026). LID for African languages remains especially challenging. Existing benchmarks cover only a small fraction of the continent's languages, and even within this limited coverage, performance often lags behind that of higher-resource languages. At the same time, major web-crawled corpora for African languages suffer from systematic quality issues (Kreutzer et al., 2022), including noisy or unusable text, misattributed documents (Alabi et al., 2020), and heavy concentration in religious or translated material that does not adequately reflect everyday language use (Kargaran et al., 2023). These artifacts degrade downstream performance and inflate apparent coverage—a form of *representation washing* (Burchell et al., 2023) that reinforces disparities in language technology (Blasi et al., 2022). Recent LID systems (Kargaran et al., 2023; Foroutan et al., 2025), including African-focused models (Adebara et al., 2022a), have made important progress. However, two central gaps remain: *scope*, i.e., the set of African languages that can be reliably identified, and *granularity*, i.e., the ability to distinguish closely related languages and varieties. We address these gaps with AFROSCOPE, a unified framework for African LID. As shown in Figure 1, AFROSCOPE consists of three main contributions:

(i) Coverage-oriented data and models. We introduce AFROSCOPE-DATA, a large-scale multilingual dataset curated to expand both language and domain coverage for African LID. AFROSCOPE-DATA spans 640 languages across multiple orthographies and domains (§3), enabling a more comprehensive assessment of African LID performance, including fine-grained analysis across domains. Using AFROSCOPE-DATA, we train AFROSCOPE-MODELS, a family of LID models that outperform prior African LID baselines across our evaluation setting (§4).

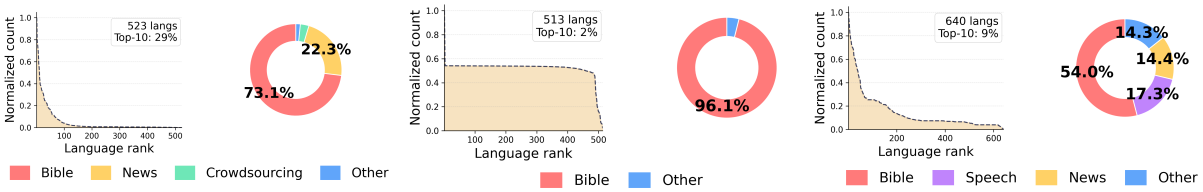
(ii) Hierarchical disambiguation of closely related languages. We show that genetically related and geographically proximate languages are a major source of false positives, especially when labels correspond to closely related varieties or macrolanguage members. To address this, we introduce AFROSCOPE-MIRROR, a lightweight contrastive embedding model specialized for frequently confused language groups, and use it in a hierarchical inference procedure that performs targeted disambiguation only when fine-grained separation is needed (§6.1.1). This design improves discrimination among closely related languages while preserving broad-coverage LID.

(iii) Transfer and robustness analysis. Leveraging AFROSCOPE-DATA, we analyze LID performance by coverage and domain (§5), and study multilingual transfer effects—including positive transfer and negative interference—as a function of language family structure and script overlap (§6.2). These analyses provide practical guidance for building, evaluating, and curating robust African LID systems.

2 Related Works

Linguistic diversity in Africa. Africa is among the most linguistically diverse regions, spanning many language families and typological profiles (Heine and Nurse, 2000; Eberhard et al., 2021a). For NLP systems, this diversity manifests in phenomena that directly stress corpus curation and LID, including rich morphology, orthographic variation, and pervasive multilingual practices such as code-switching (Ojo et al., 2025; Hussen et al., 2025). In addition, language varieties with fluid boundaries complicate labeling and evaluation (Abdulmumin et al., 2024; Alabi et al., 2025), particularly for diglossic African languages like Arabic. Recent work has responded with new resources, benchmarks, and African-focused models, which we discuss—along with related challenges—in Appendix A.1.

Data authenticity and corpus quality. Large multilingual corpora frequently contain misattributed text, ambiguous language codes, and other quality issues that disproportionately affect low-resource settings (Kreutzer et al., 2022). Prior studies of widely used multilingual resources and pipelines document systematic noise and labeling errors (Bañón et al., 2020; Schwenk et al., 2021; Xue et al., 2021), and emphasize the role of LID quality and preprocessing in mitigating such artifacts (Agarwal et al., 2023). Improving authenticity is therefore central to building reliable and culturally representative language technologies (Ojo et al., 2023; Zhong et al., 2024; Alhanai et al., 2025), motivating dataset construction that explicitly controls for coverage, domain diversity, and contamination.



(a) GlotLID-C dataset (b) AfroLID dataset (c) AFROSCOPE-DATA

Figure 2: Language-resource and domain distributions for GlotLID-C, AfroLID, and AFROSCOPE-DATA. Each pair of plots shows the language distribution of sentence counts across languages (*left*) and the corresponding domain composition (*right*). AFROSCOPE-DATA reduces language imbalance and domain concentration relative to the source corpora. Domains accounting for less than 3% of a dataset are grouped under *Other*.

Progress in African language identification. Recent evaluations show that LID coverage for African languages remains narrow and performance lags substantially (Suarez et al., 2026). Despite progress from FastText-based systems (Joulin et al., 2016; Kargaran et al., 2023; Burchell et al., 2023) to African-focused transformer-based models (Adebara et al., 2022a,b, 2024), as well as methodological advances through contrastive learning (Foroutan et al., 2025) and hierarchical approaches (Agarwal et al., 2023), targeted efforts for African LID remain necessary to address problems specific to African languages, such as domain variation and closely related varieties (Salamanca et al., 2026).

3 Addressing Coverage and Domain Skew: AFROSCOPE-DATA

Building robust LID systems for African languages requires data that is broad across two axes: *language coverage* and *domain coverage*. Existing resources cover only a small subset of African languages, leaving many underrepresented or entirely absent and making it difficult to measure true progress. Even among covered languages, sentence counts are highly concentrated in a small number of relatively high-resource languages (Figure 2a). Domain coverage is similarly skewed: religious texts, predominantly Bible translations, account for 73.10% and 94.70% of sentences in GlotLID (Figure 2a) and AfroLID (Figure 2b), respectively. This narrow distribution limits generalization to heterogeneous digital text, including web text, administrative documents, and transcribed speech.

To address these challenges, we curate AFROSCOPE-DATA (Figure 2c) using a strategy guided by two objectives: *(i) maximizing language coverage* to reduce *out-of-model cousin* errors (Caswell et al., 2020; Kreutzer et al., 2022), where text from an unsupported language is misattributed to the closest supported relative; and *(ii) increasing domain diversity* to mitigate the narrow domain concentration in available African language data.

AFROSCOPE-DATA (Table 1) spans 640 African languages across eight language families, seven scripts, and eight domains. To the best of our knowledge, AFROSCOPE-DATA provides the broadest publicly described coverage for African LID in terms of the joint number of African language labels and domain categories.

3.1 Data Curation and Sampling

We compile AFROSCOPE-DATA from publicly described sources, prioritizing datasets that provide sufficient provenance metadata for assigning coarse domain labels. To maximize coverage, we use `GlotLID-C` (Kargaran et al., 2023), `AfroLID` (Adebara et al., 2022a), and `SimbaText` (Elmadany et al., 2025) as training sources, given their breadth and metadata availability (Table 1).

To mitigate language imbalance, we sample a fixed-size training corpus using temperature smoothing, following prior multilingual practice (Arivazhagan et al., 2019; Costa-Jussà et al., 2022; Burchell et al., 2023). For a language l with corpus fraction p_l, we sample proportional to p_l^α, where $\alpha = 0.3$. We apply the same smoothing hierarchically at the domain level: within each language's budget, a domain d with intra-language fraction p_d is sampled proportional to p_d^α. This reduces the dominance of high-frequency domains, such as religious texts, and increases the representation of lower-frequency ones.

Through this sampling strategy, AFROSCOPE-DATA mitigates imbalance along both axes (Figure 2c). The top-10 languages account for only 9%

	Dataset		Sent.	Lang.	Family	Script	Domain
TRAIN	GlotLID-C (Kargaran et al., 2023)		60,682,541	523	7	5	
	AfroLID (Adebara et al., 2022a)		1,682,541	513	7	5	
	SimbaText (Elmadany et al., 2025)		382,541	101	5	4	
	AFROSCOPE-DATA	*Train*	5,952,575	640	8	7	
		Dev	463,875				
EVALUATION	FLORES+ (NLLB Team et al., 2024)		108,486	54	5	4	
	MAFAND (Adelani et al., 2022)		54,795	21	4	2	
	SmolSent (Caswell et al., 2025)		10,872	53	6	4	
	MCS-350 (Agarwal et al., 2023)		94,894	141	6	3	
	UDHR (Kargaran et al., 2023)		6,696	117	6	4	
	CommonLID (Suarez et al., 2026)		10,696	25	3	3	
	AFROSCOPE-DATA	*Test*	232,563	640	8	7	

Speech, Human Rights, Crowdsourcing, News, Bible, Web, Government, Other

Table 1: Summary statistics for the datasets used to construct and evaluate AFROSCOPE-DATA. We report the number of sentences (*Sent.*), language labels (*Lang.*), high-level language-family groupings (*Family*), scripts (*Script*), and domains (*Domain*). Colored squares show the domain categories represented in each source. Details on source composition and domain mapping are provided in Appendix B.2.

of sentences, down from 29% in GlotLID-C. Bible-derived content is reduced to 53.5%, compared to 73.1% and 94.7% in GlotLID-C and AfroLID, respectively. We provide more details regarding data curation and sampling in Appendix B.

3.2 Family, Script, and Domain Coverage

Language family. AFROSCOPE-DATA spans eight high-level groupings: *Afro-Asiatic, Austronesian, Creole, Indo-European, Khoe-Kwadi, Mixed language, Niger-Congo,* and *Nilo-Saharan*. This diversity reduces reliance on cues from dominant families such as *Niger-Congo* and supports evaluation of cross-family generalization. We use this hierarchy in our transfer analyses (§6.2).

Script. We include seven writing systems: *Latin* (`Latn`), *Arabic* (`Arab`), *Ge'ez* (`Ethi`), *N'Ko* (`Nkoo`), *Tifinagh* (`Tfng`), *Coptic* (`Copt`), and *Vai* (`Vaii`). We explicitly label scripts for each language (*language_script*), as individual languages may employ multiple writing systems (e.g., *gof, ttq*), enabling script-aware evaluation and analysis of orthographic variation.

Domain. To analyze domain effects, we first adopt the source-domain categorization from GlotLID (Kargaran et al., 2023), which groups sources into coarse categories such as bible, news, web, government, and social text. We extend this mapping to sources absent from the original GlotLID taxonomy, using the origin metadata each dataset provides. This yields a compact set of domains: *Speech, Human Rights, Crowdsourcing, News, Bible, Web, Government*, and *Other*. We assign each sentence a domain label through this source-level mapping and use the labels for controlled evaluation by domain; we discuss domain effects in detail in §5.

4 Experimental Setup

Evaluation data. We evaluate on six external EVALUATION sets spanning diverse domains and language coverage: `FLORES+` (NLLB Team et al., 2024), `UDHR` (Kargaran et al., 2023), `MAFAND` (Adelani et al., 2022), `SmolSent` (Caswell et al., 2025), `MCS-350` (Agarwal et al., 2023), and `CommonLID` (Suarez et al., 2026). We also evaluate on a held-out AFROSCOPE-DATA test split. Together, these benchmarks allow us to assess broad-coverage African LID and robustness across domains.

Contamination analysis. Because several African LID resources are derived from overlapping public sources, we explicitly measure residual overlap between TRAIN and EVALUATION data. We use a strict 4-gram contamination criterion: a test sentence is marked as contaminated if all of its 4-grams appear within a single training sentence. Table 3 reports contamination rates across three training corpora and the six external evaluation sets. AFROSCOPE-DATA exhibits no detected contamination on any of the EVALUATION sets under this criterion, whereas the AfroSimba and GlotLID-C corpora show non-trivial overlap, particularly on MCS-350 (14.15% with GlotLID-C) and MAFAND (8.13% with GlotLID-C). We

Models	FLORES+ $n = 54$		UDHR (Human Rights) $n = 117$		CommonLID (Web) $n = 25$		SmolSent (Translation) $n = 53$		MAFAND (News) $n = 21$		MCS-350 (Stories) $n = 141$		AFROSCOPE Test $n = 640$	
	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$	$F_1 \uparrow$	FPR $\downarrow$
AfroLID	61.96	0.0015	56.61	0.0014	90.51	0.0010	58.06	0.0020	79.77	0.0017	52.53	0.0011	69.04	0.0004
GlotLID-M	95.25	0.0003	74.01	0.0007	94.33	0.0003	80.21	0.0006	**90.87**	0.0006	76.95	0.0002	71.85	0.0003
ConLID	95.35	0.0003	74.76	0.0008	94.23	0.0002	79.58	0.0006	90.65	0.0005	**77.56**	0.0002	72.90	0.0002
OpenLID	83.30	0.0013	26.19	0.0034	90.94	0.0009	47.31	0.0031	82.37	0.0017	39.83	0.0020	20.13	0.2100
AfroLID$_{AS}$	62.11	0.0024	59.74	0.0020	91.13	0.0007	64.75	0.0024	79.63	0.0015	54.70	0.0011	74.38	0.0003
Serengeti$_{AS}$	52.35	0.0037	57.72	0.0023	92.22	0.0007	60.05	0.0034	73.28	0.0027	45.70	0.0016	75.07	0.0003
Cheetah$_{AS}$	58.49	0.0017	62.34	0.0016	86.89	0.0005	59.67	0.0018	69.25	0.0017	51.53	0.0008	79.49	0.0003
AfroLID$_{GL}$	94.61	0.0003	72.35	0.0009	95.70	0.0004	84.08	0.0007	83.69	0.0011	73.39	0.0006	70.09	0.0004
Serengeti$_{GL}$	95.80	0.0003	72.82	0.0009	96.39	0.0004	84.48	0.0008	83.98	0.0012	75.55	0.0005	69.24	0.0004
Cheetah$_{GL}$	92.04	0.0004	72.58	0.0008	91.81	0.0004	78.81	0.0008	79.77	0.0012	68.65	0.0006	75.72	0.0004
AfroLID$_{AF}$	95.06	0.0004	74.61	0.0009	96.51	0.0004	85.07	0.0007	87.66	0.0010	74.56	0.0005	95.82	0.0001
+ *Mirror*	96.29	0.0003	77.53	0.0007	95.47	0.0004	85.07	0.0007	88.69	0.0009	74.56	0.0004	97.74	0.0000
Serengeti$_{AF}$	95.04	0.0004	74.63	0.0009	96.51	0.0004	85.06	0.0007	87.66	0.0010	74.56	0.0005	95.82	0.0001
+ *Mirror*	**97.44**	0.0003	**77.91**	0.0006	**96.72**	0.0004	**85.70**	0.0007	90.13	0.0008	76.03	0.0004	**97.87**	0.0000
Cheetah$_{AF}$	95.62	0.0003	76.29	0.0008	93.70	0.0003	85.59	0.0007	85.42	0.009	71.77	0.0005	96.90	0.0000

Table 2: Language identification performance across seven evaluation benchmarks, where n denotes the number of African languages in each benchmark. We report macro-F_1 and false positive rate (FPR), with higher macro-F_1 and lower FPR indicating better performance. Subscripts denote the fine-tuning corpus: $_{GL}$ = GlotLID, $_{AS}$ = AfroSimba, and $_{AF}$ = AFROSCOPE. + *Mirror* adds AFROSCOPE-MIRROR for targeted disambiguation. Red underlines indicate benchmark results affected by high training–evaluation contamination, and **bold** denotes the best macro-F_1 score for each benchmark.

return to the relationship between contamination and downstream model performance in §5.

Eval Set	AfroSimba		Glotlid-C		AFROSCOPE-DATA	
	Cont.	Lang.	Cont.	Lang.	Cont.	Lang.
FLORES+	0.00%	0	0.00%	0	0.00%	0
UDHR	**1.75%**	33	0.97%	10	0.00%	0
CommonLID	1.41%	16	**7.19%**	18	0.00%	0
SMOL	0.00%	1	**0.04%**	16	0.00%	0
MAFAND	0.37%	12	**8.13%**	18	0.00%	0
MCS-350	1.47%	84	**14.15%**	97	0.00%	0

Table 3: Training–evaluation contamination across external benchmarks. We report the percentage of contaminated evaluation sentences (*Cont.*) and the number of affected languages (*Lang.*) per training corpus. Highest contamination rates are bolded.

4.1 Evaluation Protocol and Metrics

Following (Kargaran et al., 2023), we evaluate each model on the full benchmark rather than filtering examples to the model's supported language set. This setting reflects more realistic LID deployment conditions, where systems encounter text in languages outside their training label space and must avoid incorrectly assigning such text to an in-scope language. It also makes coverage differences visible: models with narrower label spaces are penalized when they force unsupported languages into related or superficially similar supported labels.

We report **macro-F_1**, which averages performance across languages, together with **False Positive Rate** (FPR), defined as: $\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}$, where FP is the number of false positives and TN is the number of true negatives. Macro-F_1 measures average per-language identification quality, while FPR captures how often a model falsely attributes text to a language it does not belong to. This distinction is important for African LID, where unsupported or closely related languages can otherwise be misclassified as higher-resource relatives.

4.2 Baselines

We compare against a diverse set of LID systems, ranging from FastText-based classifiers to transformer-based African language models.

FastText models. We evaluate three existing FastText-based LID systems: *GlotLID-M* (Kargaran et al., 2023), *OpenLID* (Burchell et al., 2023) and *ConLID* (Foroutan et al., 2025). *ConLID* incorporates supervised contrastive learning to improve robustness on out-of-domain data.

Neural models. We evaluate three transformer-based models developed for African languages: *AfroLID* (Adebara et al., 2022a), *Serengeti* (Adebara et al., 2022b), and *Cheetah* (Adebara et al., 2024). To isolate the effect of training data, we fine-tune these models on each TRAIN source where applicable, and refer to the resulting models collectively as AFROSCOPE-MODELS. Because *SimbaText* is substantially smaller than the other training sources, we merge it with *AfroLID* to form a combined training source, *AfroSimba*. Fine-tuning hyperparameters are provided in Appendix C.

5 Results

Table 2 reports macro-F_1 and FPR across all external EVALUATION sets for the full set of

baselines and AFROSCOPE-MODELS variants. AFROSCOPE-MODELS achieve the strongest overall performance, leading on five of seven benchmarks, with the largest gains appearing on benchmarks with broader language coverage. The strongest improvements occur on the two broadest evaluation sets, AFROSCOPE-DATA Test (640 languages) and UDHR (117 languages), where the best $Serengeti_{AF+Mirror}$ variant obtains 97.87 and 77.91 macro-F_1, respectively, with FPRs of 0.0000 and 0.0006. Datasets with narrower language coverage, such as FLORES+, CommonLID, and SmolSent, show the same general pattern but with smaller margins. GlotLID-M and ConLID, both trained on the GlotLID-C corpus, achieve the strongest results only on MAFAND (90.87 and 90.65 macro-F_1) and MCS-350 (77.56 macro-F_1). These are also the evaluation datasets for which the GlotLID-C corpus exhibits the highest contamination, suggesting that evaluation overlap may contribute to their advantage on these benchmarks rather than reflecting broader generalization.

Performance and coverage. Coverage is a major driver of performance differences across evaluation sets. Languages absent from a model's training label space are vulnerable to forced misattributions, where the model assigns them to a related or superficially similar supported language, thereby inflating FPR. Across the seven evaluation sets, the GlotLID-C corpus label space lacks coverage for 171 evaluation-language entries while AFROSCOPE-DATA covers all but 31. On datasets where both corpora cover nearly all evaluation languages, such as FLORES+, MAFAND, and SmolSent, model performance clusters within a relatively narrow macro-F_1 range. In contrast, on broader evaluation sets with larger coverage gaps, such as UDHR and MCS-350, AFROSCOPE-MODELS variants achieve an average macro-F_1 of 75.25 compared to 67.25 for non-AFROSCOPE-DATA baselines, while average FPR decreases from 0.00101 to 0.00067.

Performance by domain. Figure 3 shows that model performance is strongly shaped by the amount and distribution of domain-specific training data. High-resource domains such as *Bible, News, Crowdsourcing,* and *Government* generally achieve strong per-language macro-F_1 scores, with

Models	AFROSCOPE Test†	
	$F1 \uparrow$	$FPR \downarrow$
GlotLID-M	69.08	0.0002
$Serengeti_{AF}$	97.71	0.0001
Gemini-Pro	36.40	0.0069
GPT-5.4-mini	45.30	0.0043
Claude-4.7-OPUS	47.40	0.0041

Table 4: Comparison of dedicated LID systems and frontier LLMs on AFROSCOPE Test.

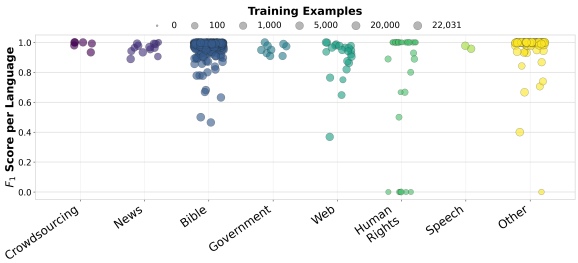


Figure 3: Per-language macro-F_1 scores across domains. Bubble size corresponds to training examples.

most languages clustered near the top of the distribution. In contrast, lower-resource or more heterogeneous domains exhibit weaker and less stable behavior. *Human Rights* contains only a few training examples across 60 languages, and its per-language results include a severe lower tail, with several languages receiving near-zero macro-F_1. This pattern helps explain the relatively low performance on UDHR, a human-rights benchmark. Similarly, the *Web* domain has moderate training coverage but noticeably weaker lower-tail performance, suggesting that domain diversity and text heterogeneity also affect robustness beyond raw example count alone. Overall, these results show that LID performance depends on both language coverage and domain coverage: domains that are well represented in training generalize more reliably, while sparse or unevenly distributed domains produce brittle performance for some languages.

Comparison with frontier LLMs. Given that LLMs now set the state of the art across many NLP tasks, we evaluate three frontier LLMs on AFROSCOPE-DATA test: Gemini-Pro, Claude-4.7-OPUS, and GPT-5.4-mini. Prompt templates, candidate-label formatting, and other evaluation details are provided in Appendix D. As shown in Table 4, all three LLMs lag substantially behind dedicated LID systems: the strongest, Claude-

†We downsample to 100 examples per language to reduce inference cost.

4.7-OPUS, reaches only 47.40 F_1 (FPR 0.0041), compared to 69.08 for GlotLID-M and 97.71 for AFROSCOPE-MODELS. This gap suggests that despite their broad multilingual capabilities, general-purpose LLMs are not a substitute for purpose-built LID models on African languages.

6 Analysis

In this section, we analyze the patterns behind the aggregate results in Table 2. We focus primarily on AFROSCOPE-DATA test because it has the broadest language coverage, while using external evaluation sets to test whether the same patterns recur across domains and benchmarks. Our analysis addresses two questions: *(i)* how does confusability among closely related languages affect performance? and *(ii)* which languages synergize or interfere most with one another, and how do language family relationships and script compatibility shape these effects?

6.1 Confusability in related languages

We find that *label confusability* is a primary failure mode: the repeated false-positive assignment between a focus language and a small set of related or geographically proximate labels. Table 5 previews representative confusion patterns identified across the evaluation sets. Such cases are especially common across ISO macrolanguage groupings and closely related languages or varieties that share substantial lexical and orthographic overlap. We find that these confusions recur in at least four of six evaluation sets, indicating the errors are stable across datasets rather than isolated benchmark artifacts. Among *macrolanguage* varieties, Arabic (*ara*: *aeb, ary, arz*) and Fulah (*ful*: *fub, fuf, fuv*) are representative. High-confusion pairs are often genetically related and geographically proximate, such as Lamnso (*lns*) and Tikar (*tik*) in Cameroon, and Anii (*blo*) and Lukpa (*dop*) in Togo. These groupings account for a disproportionate share of false positives, suggesting that models struggle to discriminate fine-grained varieties. Together, these patterns indicate that many errors reflect genuine linguistic similarity and label granularity rather than random noise. We provide further examples in Appendix E.1.

6.1.1 AFROSCOPE-MIRROR Disambiguation

To address these systematic confusions, we introduce AFROSCOPE-MIRROR, a lightweight contrastive embedding model used inside a hierarchical inference procedure. The base LID model first produces a broad-coverage prediction. When the prediction falls within a predefined confusable group and the model's confidence is below a threshold τ, AFROSCOPE-MIRROR performs a group-specific disambiguation step using specialized embeddings. We build AFROSCOPE-MIRROR on top of AfroLID$_{AF}$ and Serengeti$_{AF}$, our strongest base models on the external evaluation sets, and train it with Mirror-BERT (Liu et al., 2021), an unsupervised contrastive learning objective that pulls semantically similar representations together while pushing unrelated ones apart. Detailed training procedures and hyperparameters are in Appendix E.1.

Figure 4 compares the embedding spaces produced by Serengeti$_{AF}$ and Serengeti$_{AF+Mirror}$, vi-

	Lang.	FP.	FC.	TC.
Macro.	**ara**	0.093	2281	aeb, apd, arq, ary
	din	0.040	497	dib, dik, dip, dks
	ful	0.010	414	ffm, fub, fuc, fue
High.	**lns**	0.281	2577	tik, pfe, agq
	blo	0.199	1748	kbp, dop, bib
	kin	0.153	4323	run, swh, nnb

Table 5: Representative false-positive patterns for highly confusable languages. We group cases into macrolanguage-related confusions and other high-confusion language groups. *FP* reports the false positive rate (%), *FC* reports the false positive count, and *TC* lists the most frequent target labels to which the focus language is incorrectly assigned.

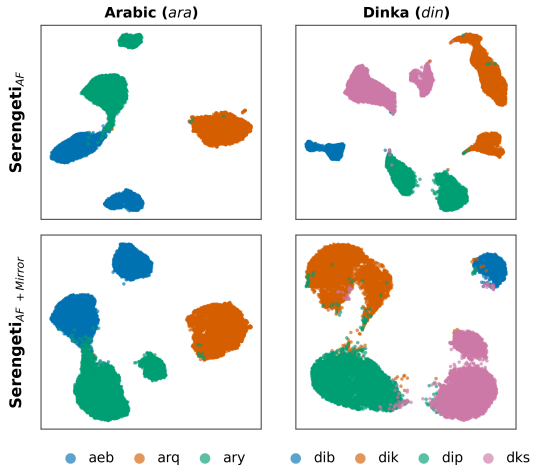


Figure 4: UMAP of embedding spaces from base *Serengeti*$_{AF}$ (top) and *Serengeti*$_{AF+Mirror}$ (bottom). We show examples from macro-language groups and high-confusion language pairs. Specialized embeddings exhibit improved separation between closely related varieties. We provide additional examples in Appendix E.1.

sually illustrating clearer separation among confusable labels and tighter within-label clustering after specialization.

We evaluate this strategy on the 30 confusable languages in Table E.2, including focus languages and their top confusion partners, across confidence thresholds. On the full set, the average macro-F_1 gain is +1.57 points. As shown in Table 2, adding AFROSCOPE-MIRROR yields consistent gains across the external benchmarks: Serengeti$_{AF + Mirror}$ improves over its base model on all seven evaluation sets (e.g., +2.40 on FLORES+, +3.28 on UDHR, +2.47 on Mafand), and AfroLID$_{AF + Mirror}$ improves on five of seven datasets. At the language level (Table 6), we observe improvements for closely related varieties within macro-language groupings, such as Arabic (*ara*) varieties *aeb*, *arq*, and *ary*, which improve by +0.63, +0.79, and +0.31, respectively, as well as Swahili (*swa*) variety *swc* (+0.55). We also observe improvements for confusable regional pairs, including Kinyarwanda (*kin*) and *run* (+0.71). However, a small number of languages decline, including *niq* (-0.10) and *kbp* (-0.14), suggesting that hierarchical routing can introduce unnecessary complexity when the base model is already reliable for a given label.

Group	Lang	Baseline	$\tau = 0.75$		$\tau = 0.85$		$\tau = 0.95$	
			Δ	F_1	Δ	F_1	Δ	F_1
ara	aeb	94.50	+0.18	94.68	**+0.63**	**95.13**	+0.50	95.00
	arq	98.89	+0.32	99.21	+0.64	99.53	**+0.79**	**99.68**
	ary	96.05	+0.11	96.16	**+0.31**	**96.35**	+0.13	96.17
swa	swc	94.21	+0.34	94.55	+0.07	94.28	**+0.55**	**94.77**
kin	run	95.34	+0.03	95.37	+0.03	95.37	**+0.71**	**96.05**
Avg		–	+0.19	–	+0.33	–	**+0.54**	–

Table 6: Sample hierarchical classification results using AFROSCOPE-MIRROR embeddings across confidence thresholds τ. Baseline F_1 shows base Serengeti performance; Δ columns show improvement in F_1 over baseline. Bold indicates the best F_1 per language. Full results are provided in Appendix E.2

6.2 Transfer Synergy and Interference

Given limited and skewed resources, effective African LID depends on cross-lingual transfer. We investigate which languages synergize or interfere most, and how language family and script compatibility shape these effects. To quantify this, we adapt the Bilingual Transfer Score (BTS) (Longpre et al., 2025) to language identification. We compute BTS using macro-F_1, and cross-entropy loss across 30 representative African languages spanning five family clusters.

Figure E.3 shows that cross-lingual transfer in African LID is governed by shared script and fine-grained language relatedness rather than coarse family membership. Co-training yields the largest F_1 gains for low-resource targets with weak monolingual baselines (e.g., *fuv*, *dyu*, *bam*; F_1 BTS up to +0.22), consistent with baseline rescue rather than family-specific transfer. Critically, the closest relationships are not uniformly beneficial: mutually intelligible varieties such as the Arabic cluster (*arz*, *ary*, *aeb*) and Mande pair *bam–dyu* improve F_1 but *degrade* probability calibration, with loss-based BTS as low as -0.24. A regression across all pairs confirms that script and this confusable-cousin relationship—not language family—are the dominant predictors of transfer; family membership alone is not significant. This accuracy–calibration trade-off motivates hierarchical disambiguation. We report full per-family analysis in Appendix E.2.

7 Conclusion

We introduce AFROSCOPE, a unified framework for African LID that combines broad-coverage data, strong LID models, targeted disambiguation, and linguistic analysis. We present AFROSCOPE-DATA, a large-scale dataset spanning 640 language labels, which we use to train AFROSCOPE-MODELS, a family of African LID models that outperform prior African-focused baselines across internal and external evaluations.

To address persistent confusions among closely related languages and varieties, we propose a hierarchical inference approach based on AFROSCOPE-MIRROR, a specialized embedding model for frequently confused language groups. On the full set of identified confusion groups, this approach improves macro-F_1 by $+1.57$ points on average.

Finally, our transfer analysis shows that African LID transfer is shaped by script, fine-grained family relatedness and confusability than by coarse family membership. These relationships can improve low-resource performance while degrading calibration, motivating targeted disambiguation with AFROSCOPE-MIRROR. We hope AFROSCOPE, AFROSCOPE-DATA, and AFROSCOPE-MODELS support reliable African NLP and future work on fine-grained varieties, domain shifts, and mixed-language text.

Limitations

We note several limitations of the current study.

1. **Mixed-language and code-switched text.** Our formulation treats each instance as belonging to a single language label. This does not fully capture important phenomena in Africa's linguistic landscape, including code-switching, mixed-language documents, translanguaging practices, and contact varieties such as pidgins and creoles. Extending AFROSCOPE to multi-label, document-level, or span-level language identification is an important direction for future work. Relatedly, the granularity at which languages are annotated is uneven across our data sources, and in some cases closely related varieties or distinct languages may be conflated under a single label. A more careful audit of these labeling decisions, ensuring that language distinctions are drawn consistently and appropriately, is needed but lies beyond the scope of this work.

2. **Language metadata and classification choices.** AFROSCOPE-DATA relies on external catalogs, primarily Ethnologue, to assign language identifiers, genealogical groupings, and script metadata. Alternative resources, such as Glottolog, may differ in classification, naming, macrolanguage treatment, and family hierarchy. These choices may affect analyses that depend on genealogical proximity or script compatibility. Future work should evaluate sensitivity to alternative metadata sources and provide mappings across catalog standards.

3. **Confidence-based routing and calibration.** Our hierarchical disambiguation method relies on model confidence to decide when to invoke the group-specific refinement step. While this improves performance on many confusable labels, gains are not uniform across languages or thresholds, and some cases exhibit degradation. Improving probability calibration and learning a routing policy, rather than relying on fixed confidence thresholds, may further increase robustness.

4. **Domain and source imbalance.** Although AFROSCOPE-DATA improves domain diversity relative to prior African LID resources, some domains remain substantially smaller or more heterogeneous than others. As a result, performance may still be brittle for domains with limited training data or high internal variation, such as web text, human-rights documents, or informal user-generated content. Future work should expand naturally occurring data in underrepresented domains and evaluate robustness under domain shift.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *arXiv preprint arXiv:2401.01053*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022a. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating llm performance on african languages. *arXiv preprint arXiv:2502.19582*.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of*

the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3053–3070, Seattle, United States. 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, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. Limit: Language identification, misidentification, and translation using hierarchical models in 350+ languages. *arXiv preprint arXiv:2305.14263*.

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David I. Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O Alabi, Michael A Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of african nlp: Mapping progress and shaping the road ahead. *arXiv preprint arXiv:2505.21315*.

Tuka Alhanai, Adam Kasumovic, Mohammad M Ghassemi, Aven Zitzelberger, Jessica M Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Moussa Doumbouya, Djibrila Diané, et al. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123.

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, Online. Association for Computational Linguistics.

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*.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017. Improved text language identification for the south african languages. In *2017 Pattern Recognition Association of South Africa and Robotics*

and Mechatronics (PRASA-RobMech), pages 214–218. IEEE.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021a. *Ethnologue: Languages of the World*, 24 edition. SIL International, Dallas, Texas.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021b. *Ethnologue: Languages of the World*, 24th edition. SIL International, Dallas, Texas.

AbdelRahim Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Mapping africa's speech technology frontier. *arXiv preprint arXiv:2505.18436*.

Negar Foroutan, Jakhongir Saydaliev, Ye Eun Kim, and Antoine Bosselut. 2025. Conlid: Supervised contrastive learning for low-resource language identification. *arXiv preprint arXiv:2506.15304*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bernd Heine and Derek Nurse, editors. 2000. *African Languages: An Introduction*. Cambridge University Press.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The state of large language models for african languages: Progress and challenges. *arXiv preprint arXiv:2506.02280*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Alex Wang, Abdul Wahab, Naman Goyal, Noah Constant, Xinying Chen, Guillaume Wenzek, Vishrav Chaudhary, Francisco Guzmán, Philipp Koehn, Ondřej Bojar, Christian Federmann, Nizar Habash, Yulia Tsvetkov, Holger Schwenk, and Alexis Conneau. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6843.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, Sayna Ebrahimi, et al. 2025. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. *arXiv preprint arXiv:2510.22037*.

Paul Mcnamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20:94–101.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Tolulope Ogunremi, Christopher D Manning, and Dan Jurafsky. 2023. Multilingual self-supervised speech representations improve the speech recognition of low-resource african languages with codeswitching. *arXiv preprint arXiv:2311.15077*.

Jessica Ojo, Zina Kamel, and David Ifeoluwa Adelani. 2025. Divers-bench: Evaluating language identification across domain shifts and code-switching. *arXiv preprint arXiv:2509.17768*.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, and Vukosi Marivate. 2025. AfroCS-xs: Creating a compact, high-quality, human-validated code-switched dataset for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33391–33410, Vienna, Austria. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Alejandro R Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, et al. 2026. Tiny aya: Bridging scale and multilingual depth. *arXiv preprint arXiv:2603.11510*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, et al. 2026. Commonlid: Re-evaluating state-of-the-art language identification performance on web data. *arXiv preprint arXiv:2601.18026*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Ansari 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.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

Appendices

The following appendices provide comprehensive supplementary material supporting the main findings of this work. We include an expanded literature review, detailed descriptions of the datasets and models, a more in-depth discussion of the experimental setup, and additional analyses and discussions.

- §A: More in depth literature review
- §B: Data Curation & Preprocessing
- §C: Baseline Models
- §E: Discussions

A Literature Review

A.1 Linguistic Diversity of Africa.

Many African languages are agglutinative (e.g., Swahili, Zulu), complicating tokenization and parsing (Abdulmumin et al., 2024; Hussen et al., 2025), while everyday communication often involves code-switching with English and other languages (Ogunremi et al., 2023).

Datasets like AfroCS-xs demonstrate that small, high-quality resources can boost LLM performance on code-switched tasks (Olaleye et al., 2025). Models such as Afrolid, SERENGETI and Cheetah show the feasibility of massively multilingual modeling for African languages (Adebara et al., 2022a,b, 2024). Benchmarks including Sahara (Adebara et al., 2025), SimbaBench (Elmadany et al., 2025), IrokoBench (Adelani et al., 2024), and AfroBench (Ojo et al., 2023).

A.2 Data Authenticity.

Studies of multilingual corpora such as ParaCrawl (Bañón et al., 2020), WikiMatrix (Schwenk et al., 2021), and mC4 (Xue et al., 2021) reveal widespread errors, mislabeled data, and ambiguous language codes (Kreutzer et al., 2022). For instance, Alabi et al. (2020) showed that many FastText embeddings were heavily mislabeled with words from other languages. Poor data authenticity directly undermines cultural representation, as inaccurate or mislabeled data fails to capture the nuanced linguistic and cultural contexts inherent in low-resource languages (Zhong et al., 2024). Addressing these quality issues is essential for building reliable language technologies (Alhanai et al., 2025; Ojo et al., 2023).

B Data Curation and Sampling

B.1 Data Curation

We deduplicate sentences across all sources to ensure that each example appears at most once in AFROSCOPE-DATA. We then partition the data into *train*, *dev*, and *test* splits, performing deduplication prior to splitting so that no sentence is shared across splits. We ensure that each sentence is written in the correct script, based on the writing system databases of Ethnologue (Eberhard et al., 2021b). When temperature sampling we set the maximum sentences to 10M sentences. The resulting splits contain 5,952,573 sentences for train, 463,875 for dev, and 232,563 for test. The full list of languages supported by AFROSCOPE, together with the number of sentences per language, is provided in Tables E.5–E.8.

Below, we describe the sources used to construct AFROSCOPE-DATA.

Train Corpus

GlotLID-C. We extract 523 African languages from `GlotLID-C`, a collection spanning 2,099 languages globally (Kargaran et al., 2023).

AfroLID. A manually curated multi-domain web dataset covering 513 African languages (Adebara et al., 2022a).

SimbaText. Speech-derived text data spanning 103 African languages, originally collected for speech and language identification (Elmadany et al., 2025).

Evaluation Datasets

FLORES+. Professionally translated sentences across 200 languages, providing a clean and standardized benchmark, particularly for low-resource languages (NLLB Team et al., 2024).

MAFAND. Manually audited data from news, Wikipedia, and religious texts across 55 African languages (Adelani et al., 2022).

SMOL. Multilingual benchmark dataset consisting of professionally translated sentence- and document-level data for 115 LRLs (Caswell et al., 2025).

MCS-350. Parallel children's stories across 151 African languages, drawn from a multilingual collection of 50k texts in over 350 languages (Agarwal et al., 2023).

UDHR. The Universal Declaration of Human Rights, translated into a wide range of languages (Kargaran et al., 2023).

CommonLID. Professionally translated parallel data for 115 under-represented languages (Suarez et al., 2026).

B.2 Domain classification.

We assign domains by matching keywords found in the metadata associated with each sentence, following the categorization scheme from (Kargaran et al., 2023). Table B.2 lists the specific keyword mappings.

Sample Metadata
`Bible-aar_line94`
`CommonVoice v11`
`Masakhanews`
`GlotStoryBook`
`CC100_zu.txt.tsv_f17_line40162`
`JW-zul_line3295`
`gov-za`
`Open Subtitles`
`KDE4`
`Vuk'uzenzele`

Table B.1: Examples of sentence-level metadata identifiers used for domain attribution.

Domain	Associated Keywords
Speech	Speech, CommonVoice, TTS, Audio
Human Rights	Human Rights
Government	Human Rights, Autshumato, Legal, GOV, Parliament, Gazette
CrowdSource	Flores, NLB, mt560, Tatoeba, UD, ai4d, lti, Benchmark, Human, Madar, iadd
News	News, xlsum, Vukuzenzele, CBC, BBC, Afriqa, Masakha, Goud
Bible	Bible, JW, Scripture, Religion
Web	Oscar, CC, CommonCrawl, Web, Dialect, Social, Forum
Other	Health, Covid, Medical, Med, Tanzil, PBC, Quran, Story, Stories, Fiction, Bloom, Lyrics

Table B.2: Keywords extracted from dataset metadata to map sources into domain categories.

C Training Hyperparameters for Baseline Models

To isolate the effect of training data, we train or fine-tune comparable model architectures on each TRAIN source where applicable. This allows us to evaluate whether the gains come from the model architecture, the training corpus, or their interaction. Because *SimbaText* is substantially smaller than the other training sources, we merge it with *AfroLID* to form a combined training source, *AfroSimba*. AfroLID and Serengeti are XLM-RoBERTa variants, and Cheetah is a T5-based model.

C.1 Neural Models

We use the hyperparameters in Table D.1 for *AfroLID* (Adebara et al., 2022a) and *Serengeti* (Adebara et al., 2022b) (both XLM-R variants), and Table D.2 for *Cheetah* (Adebara et al., 2024).

D Evaluating Open-Srouce LLMs

D.1 Prompts for Open-Source Models

Figure D.1 shows the prompt template used to elicit language predictions from the LLMs. We provide the candidate ISO 639-3 labels in the prompt and instruct the model to return only the language code. To reduce inference cost we downsample each language examples to 100 examples per language.

argument	description	value
-max_seq_length	max input sequence length	128
-per_device_train_batch_size	training batch size (per device)	64
-learning_rate	learning rate	2e-5
-num_train_epochs	number of training epochs	10
-metric_for_best_model	evaluation metric	f1

Table D.1: Training hyperparameters for Afrolid and Serengeti.

argument	description	value
-max_target_length	max target sequence length	128
-per_device_train_batch_size	training batch size (per device)	32
-learning_rate	learning rate	5e-5
-num_train_epochs	number of training epochs	10
-metric_for_best_model	evaluation metric	f1

Table D.2: Training hyperparameters for the Cheetah model.

E Discussions

E.1 Confusability in related languages

To investigate languages with high FPR, we isolate languages with scores below F_1 90 across all evaluation datasets—and identify the top three most frequent misclassifications for each to form confusion groups. This analysis yields 13 distinct groups comprising 30 languages in total. We find that these confusions primarily stem from either macrolanguage structures (e.g., *ful* vs. *fub*) or geographic proximity (e.g., *bsq* vs. *bas*). Table E.2 details the composition of all confusion groups, and we report the corresponding performance improvements for each individual language.

Field	Content

Prompt	`Classify the language of the given sentence. Return only the ISO 639-3 language code from the list below.`
	`Language options: aba, abi, abn, acd, ach, ada, ..., zul`
	`Return only the language code, nothing else.`

Input	`Sentence: {sentence}`
Output	`amh`

Figure D.1: Prompt template used for LLM-based language identification.

Mirror-BERT training procedure We follow the procedures from the official github repository for Mirror-BERT[3]. We use *Serengeti*$_{AF}$ for this experiment as it being our best performing model. Table E.1 shows the hyperparamerters we use to train AFROSCOPE-MIRROR. We see in Figure E.1 that the label spaces of *Serengeti*$_{AF\,+\,Mirror}$ shows much better separation in the embedding spaces than *Serengeti*$_{AF}$.

argument	description	value
-epoch	number of training epochs	1
-train_batch_size	training batch size	200
-learning_rate	learning rate	2e-5
-max_length	max sequence length	50
-infoNCE_tau	InfoNCE temperature (τ)	0.04
-dropout_rate	dropout rate	0.0
-drophead_rate	drophead rate	0.05
-random_span_mask	length of random span mask	5
-agg_mode	aggregation mode	cls

Table E.1: Training hyperparameters for the Mirror-BERT model.

E.2 Transfer Synergy and Interference

We adapt the Bilingual Transfer Score (BTS) (Longpre et al., 2025) to language identification. For a source s and target t, we compare a monolingual model trained on target data $\mathcal{D}_t$ against a bilingual model trained on $\mathcal{D}_t \cup \mathcal{D}s$, evaluated on a held-out target test set. For cross-entropy loss, where lower is better, we define:

$$\mathrm{BTS}^{\mathcal{L}}_{s\to t} = \frac{\mathcal{L}_t(\mathcal{D}_t) - \mathcal{L}_t(\mathcal{D}_t \cup \mathcal{D}_s)}{\mathcal{L}_t(\mathcal{D}_t) + \epsilon}. \quad (1)$$

[3]https://github.com/cambridgeltl/mirror-bert

so that positive values indicate improvement. For macro-F_1, where higher is better, we define:

$$\mathrm{BTS}^{F1}_{s\to t} = \frac{F1_t(\mathcal{D}_t \cup \mathcal{D}_s) - F1_t(\mathcal{D}_t)}{F1_t(\mathcal{D}_t) + \epsilon}. \quad (2)$$

E.2.1 Language selection

We select 30 languages spanning five clusters. To populate a range of genealogical distances, we group eligible languages (those with at least 10,000 sentences) by their shared family hierarchy and retain clusters that contain documented confusable pairs, drawn from our false-positive analysis (Appendix E.1). This yields clusters of mutually intelligible varieties—the Arabic dialects (*arz*, *ary*, *aeb*), Fula varieties (*fub*, *fuv*, *fuh*, *fuq*, *fue*), and the Mande pair (*bam*, *dyu*)—alongside more distant same-family languages and unrelated outgroups (*plt*, *afr*, *sag*). Figure E.2 shows the fine-grained relatedness of the selected languages across the family hierarchy.

E.2.2 Experimental setup

Each language contributes 1,000 target sentences and a held-out test set; bilingual models receive 200 background sentences from every other language so all models share one label space. We fine-tune *Serengeti*$_{AF}$ for the 30 monolingual baselines (median over three seeds) and the 435 language pairs (one seed), yielding a 30 X 30 directional transfer matrix per metric.

F_1 gains track headroom. The largest F_1-BTS values occur for the weakest baselines—*fuv* (0.62), *fuq* (0.53), *dyu* (0.69), *bam* (0.73)—which improve from nearly any source (*bam*→*dyu* $= +0.22$, *fub*→*fuv* $= +0.16$; Table E.3). F_1-BTS correlates with target headroom ($r = 0.60$), indicating low-resource rescue rather than family-specific transfer; strong baselines (*kin*, *xho*, *tso*, *tsn*, all > 0.98) show near-zero BTS.

Cousins degrade calibration. The clearest signal is in loss-BTS. The Arabic varieties form a block of negative loss-BTS (*aeb*→*ary* $= -0.24$, *aeb*→*arz* $= -0.18$, *ary*→*aeb* $= -0.15$), and the Mande pair replicates this (*dyu*→*bam* $= -0.17$). All show *positive* F_1-BTS: co-training mutually intelligible varieties picks the right label but spreads probability mass across the cousin, raising cross-entropy. This trade-off is visible in Figure E.4, where cousin pairs sit in the mutual-gain region under F_1 but the mutual-loss region under cross-entropy. Bambara (*bam*) and Jula (*dyu*) are adja-

cent Manding varieties spoken across Mali, Côte d'Ivoire, and Burkina Faso (Eberhard et al., 2021b), so this interference is expected.

Script gates transfer. Substantial transfer is confined to same-script pairs: 78% of F_1 changes with $|\mathrm{BTS}| \geq 0.05$ and all calibration changes with $|\mathrm{BTS}| \geq 0.10$ occur within a script (Table E.4). Controlling for headroom and cousin status, shared script independently predicts larger F_1-transfer magnitude ($\beta = +0.005$). The exception is low-resource rescue, where weak Fula targets improve from any source. Distinctive scripts with strong baselines, like the Ethiopic pair *amh–tir* (> 0.99), show negligible transfer (mean $|\mathrm{BTS}^{\mathcal{L}}| = 0.006$), as orthographic distinctiveness leaves no confusion to resolve.

Distance gates the effect. Different-family pairs show loss-BTS near zero (distance-5 mean $= +0.002$, $n = 472$). Regressing loss-BTS on genealogical distance d and d^2 yields a significant inverted-U ($\beta_d = +0.015$, $\beta_{d^2} = -0.002$): distance-0 cousins are uniquely negative (mean $= -0.029$), all greater distances near zero. F_1-BTS shows no distance pattern, masked by ceiling saturation.

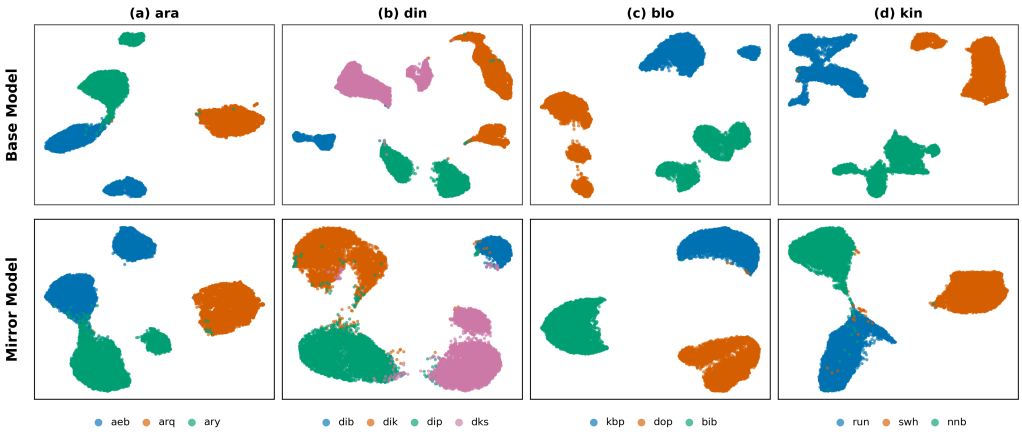


Figure E.1: UMAP visualization comparing base $Serengeti_{AF}$ (top) and $Serengeti_{AF+Mirror}$(bottom) embedding spaces. We visualize four groups representing macro-languages and confusion pairs. Specialized embeddings show improved separation between closely related language varieties.

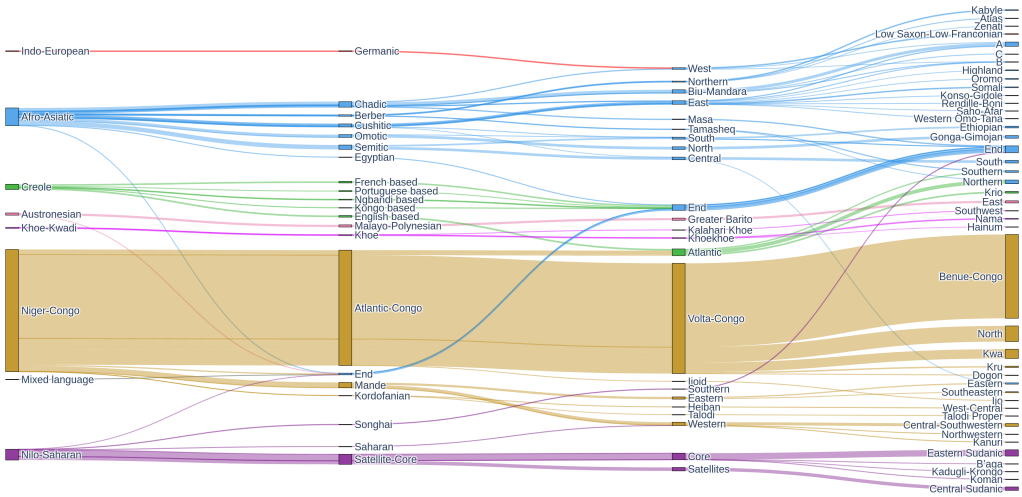


Figure E.2: Distribution of languages across major language groupings, intermediate sub-families, and finer-grained groupings, capturing their genetic relationships.

Group	Lang	Baseline	$\tau = 0.75$		$\tau = 0.85$		$\tau = 0.95$	
			F_1	Δ	F_1	Δ	F_1	Δ
aar	aar	97.10	**98.24**	**+1.14**	**98.24**	**+1.14**	**98.24**	**+1.14**
bsq	bas	94.23	**95.36**	**+1.13**	**95.40**	**+1.17**	**95.46**	**+1.23**
ara	aeb	94.50	94.68	+0.18	**95.13**	**+0.63**	95.00	+0.50
	ara	92.55	92.84	+0.29	92.99	+0.44	**93.12**	**+0.57**
	arq	98.89	99.21	+0.32	99.53	+0.64	**99.68**	**+0.79**
	ary	96.05	96.16	+0.11	**96.35**	**+0.31**	96.17	+0.13
	arz	93.46	95.70	+2.24	**96.23**	**+2.77**	98.91	+5.45
kin	kin	92.90	93.17	+0.27	93.17	+0.27	**93.57**	**+0.66**
	run	95.34	95.37	+0.03	95.37	+0.03	**96.05**	**+0.71**
	swh	99.80	**99.90**	**+0.10**	**99.90**	**+0.10**	**99.90**	**+0.10**
	nnb	98.58	**98.72**	**+0.14**	**98.72**	**+0.14**	**98.72**	**+0.14**
lns	tik	99.60	**99.70**	**+0.10**	**99.70**	**+0.10**	**99.70**	**+0.10**
	pfe	96.83	**96.96**	**+0.13**	**96.96**	**+0.13**	**96.96**	**+0.13**
	agq	95.16	**95.29**	**+0.14**	**95.29**	**+0.14**	**95.29**	**+0.14**
blo	kbp	98.51	**98.64**	**+0.13**	**98.64**	**+0.13**	**98.64**	**+0.13**
	dop	99.17	**99.31**	**+0.14**	**99.31**	**+0.14**	**99.31**	**+0.14**
	bib	97.41	**97.54**	**+0.13**	**97.54**	**+0.13**	**99.35**	**+1.94**
maf	maf	95.37	**99.50**	**+4.13**	**99.50**	**+4.13**	**99.50**	**+4.13**
	mcn	96.05	98.17	+2.12	98.37	+2.32	**99.18**	**+3.13**
	meq	95.26	**99.86**	**+4.60**	**99.86**	**+4.60**	**99.86**	**+4.60**
	mfz	95.74	97.79	+2.05	97.89	+2.15	**98.84**	**+3.10**
	mpe	96.19	98.56	+2.37	98.76	+2.57	**98.96**	**+2.77**
	mug	96.60	98.90	+2.30	99.30	+2.70	**99.70**	**+3.10**
	myx	94.96	**95.86**	**+0.90**	**95.86**	**+0.90**	**95.86**	**+0.90**
din	din	70.51	**73.75**	**+3.24**	72.96	+2.44	**73.75**	**+3.24**
ful	fub	92.28	95.62	+3.34	96.33	+4.05	**96.59**	**+4.31**
sot	sot	92.83	**94.08**	**+1.25**	**94.08**	**+1.25**	**94.08**	**+0.25**
ibo	ibo	98.58	**99.53**	**+0.95**	**99.53**	**+0.95**	**99.53**	**+0.95**
wol	wol	88.13	**90.35**	**+2.22**	**90.35**	**+2.22**	**90.35**	**+2.22**
swa	swc	94.21	94.55	+0.34	94.28	+0.07	**94.77**	**+0.55**
Average		–	–	**+1.22**	–	**+1.29**	–	**+1.57**

Table E.2: Hierarchical classification results using AFROSCOPE-MIRROR embeddings across available confidence thresholds ($\tau = 0.75, 0.85, 0.95$). Baseline F_1 shows base Serengeti performance; Δ columns show improvement over baseline. Bold indicates best performance per language.

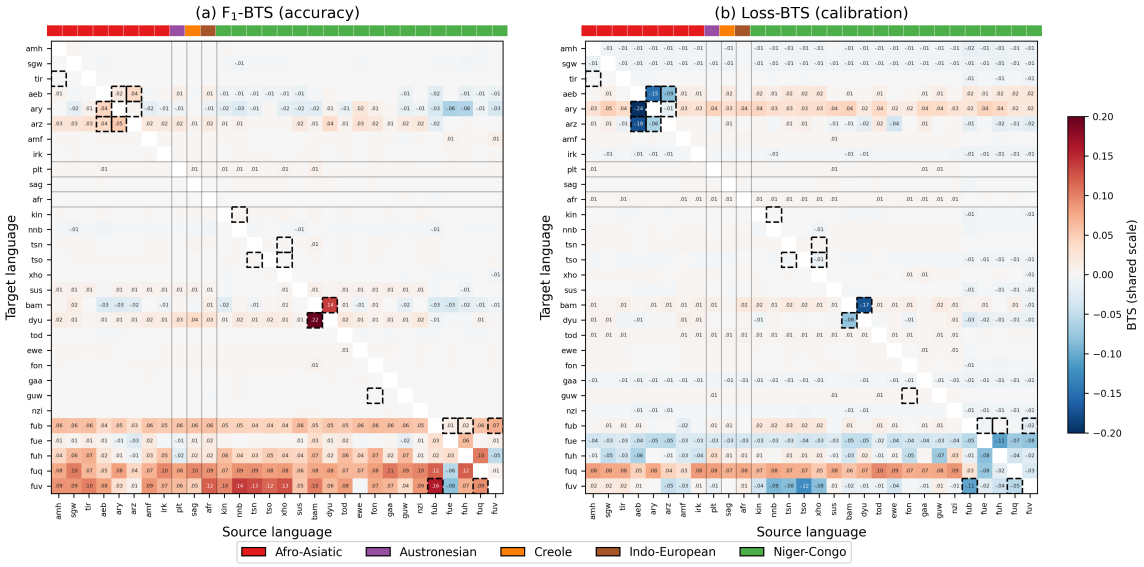


Figure E.3: Pairwise transfer effects across African languages, measured by **(a)** F_1-BTS (accuracy) and **(b)** loss-BTS (calibration). Rows indicate target languages and columns indicate source languages used for co-training; cell values show the transfer score, with warmer colors indicating positive transfer and cooler colors indicating negative transfer. Color bars along the axes indicate high-level language-family groupings. Dashed boxes highlight selected within-family or same-script clusters where transfer is especially structured, such as Arabic varieties, Southern Bantu languages, Mande languages, and Fulah varieties. Both panels share a common symmetric color scale; note that F_1-BTS and loss-BTS are distinct quantities, so the same pair may gain accuracy in (a) while losing calibration in (b).

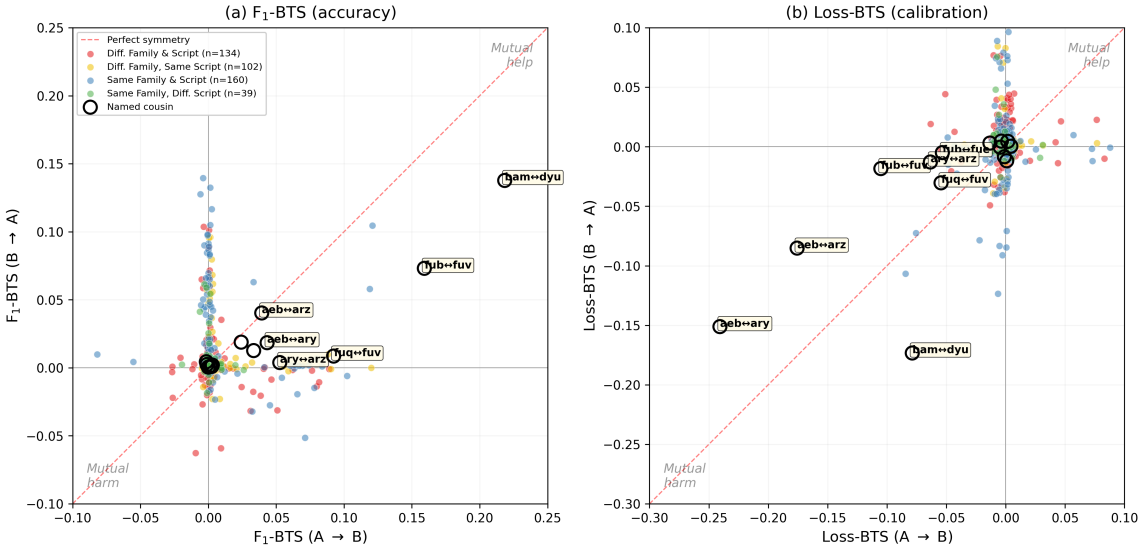


Figure E.4: Directional transfer symmetry for **(a)** F_1-BTS and **(b)** loss-BTS. Each point is a language pair; axes show BTS in each direction, and the dashed line marks perfect symmetry. Confusable cousins (circled) sit in the mutual-gain (upper-right) region under F_1 but the mutual-loss (lower-left) region under cross-entropy, illustrating the accuracy–calibration trade-off; off-diagonal points indicate asymmetric transfer.

Source	Target	BTS^{F1}	$\mathrm{BTS}^{\mathcal{L}}$	Cousin
bam	*dyu*	+0.22	-0.08	✓
fub	*fuv*	+0.16	-0.11	✓
nnb	*fuv*	+0.14	-0.09	
dyu	*bam*	+0.14	-0.17	✓
tsn	*fuv*	+0.13	-0.08	
xho	*fuv*	+0.13	-0.08	
fuh	*fuq*	+0.12	-0.02	
afr	*fuv*	+0.12	+0.01	
fub	*fuq*	+0.12	+0.03	
tso	*fuv*	+0.12	-0.12	

Table E.3: Largest F_1 transfer events. Confusable cousins improve F_1 while degrading calibration (negative $\mathrm{BTS}^{\mathcal{L}}$).

Script pair	n	$\overline{\lvert\mathrm{BTS}^{F1}\rvert}$	$\overline{\lvert\mathrm{BTS}^{\mathcal{L}}\rvert}$
Latn+Latn	506	0.015	0.012
Arab+Latn	184	0.010	0.011
Ethi+Latn	138	0.008	0.008
Arab+Ethi	24	0.007	0.010
Arab+Arab	12	0.023	0.066
Ethi+Ethi	6	0.001	0.006

Table E.4: Mean transfer magnitude by script pair. Within-script pairs (bold) dominate; the within-Arabic block carries the largest calibration effects.

ISO-3	Language	Sentences	ISO-3	Language	Sentences	ISO-3	Language	Sentences
aar	Afar	2,930	**aba**	Abé	6,902	**abi**	Abidji	10,218
abn	Abua	8,431	**acd**	Gikyode	7,102	**ach**	Acholi	8,626
ada	Dangme	11,002	**ade**	Adele	14,293	**adh**	Jopadhola	7,569
adj	Adioukrou	7,121	**aeb**	Arabic, Tunisian	12,985	**afr**	Afrikaans	45,990
agq	Aghem	2,942	**aha**	Ahanta	6,773	**ajg**	Aja	7,714
aka	Akan	2,948	**akp**	Siwu	7,118	**ald**	Alladian	13,794
alz	Alur	10,003	**amf**	Hamer-Banne	14,270	**amh**	Amharic	35,740
ann	Obolo	7,310	**anu**	Anuak	2,917	**anv**	Denya	7,126
any	Anyin	13,994	**ara**	Arabic	2,938	**arq**	Arabic, Algerian	3,693
ary	Arabic, Moroccan	24,745	**arz**	Arabic, Egyptian	25,958	**asa**	Asu	2,939
asg	Cishingini	6,884	**atg**	Ivbie North-Okpela-Arhe	7,135	**ati**	Attié	6,212
avn	Avatime	6,638	**avu**	Avokaya	7,080	**azo**	Awing	2,627
bam	Bamanankan	11,854	**bas**	Basaa	8,895	**bav**	Vengo	6,800
bba	Baatonum	9,452	**bbj**	Ghomálá'	6,605	**bbk**	Babanki	5,191
bbo	Konabéré	10,168	**bci**	Baoulé	12,880	**bcn**	Bali	2,942
bcw	Bana	7,119	**bcy**	Bacama	336	**bdh**	Baka	6,750
bds	Burunge	2,940	**bem**	Bemba	41,323	**beq**	Beembe	6,638
bex	Jur Modo	7,389	**bez**	Bena	2,946	**bfa**	Bari	2,936
bfd	Bafut	7,129	**bfo**	Birifor, Malba	6,824	**bib**	Bisa	7,133
bim	Bimoba	8,827	**bin**	Edo	9,811	**biv**	Birifor, Southern	7,124
bjv	Bedjond	7,569	**bkv**	Bekwarra	14,291	**bky**	Bokyi	2,929
blh	Kuwaa	10,202	**bmo**	Chrambo	2,940	**bmq**	Bomu	14,293
bmv	Bum	6,633	**bom**	Berom	6,676	**bov**	Tuwuli	7,117
box	Buamu	7,103	**bqc**	Boko	8,703	**bqj**	Bandial	6,635
bqp	Bisã	14,290	**bsc**	Oniyan	6,636	**bsp**	Baga Sitemu	6,332
bsq	Bassa	4,803	**bss**	Akoose	7,119	**bst**	Basketo	1,435
btt	Bete-Bendi	14,286	**bud**	Ntcham	6,825	**bum**	Bulu	10,172
bun	Sherbro	333	**bus**	Bokobaru	7,213	**buy**	Bullom So	668
bwq	Bobo Madaré, Southern	14,292	**bwr**	Bura-Pabir	2,929	**bwu**	Buli	7,114
bxk	Bukusu	2,904	**byf**	Bete	2,681	**byv**	Medumba	3,902
bza	Bandi	2,942	**bzw**	Basa	2,939	**cce**	Chopi	9,805
cgg	Chiga	5,262	**chw**	Chuwabu	9,742	**cjk**	Chokwe	21,845
cko	Anufo	6,786	**cme**	Cerma	7,105	**cop**	Coptic	9,302
cou	Wamey	6,042	**cri**	Sãotomense	2,697	**crs**	Seychelles French Creole	13,500
csk	Jola-Kasa	7,127	**cwe**	Kwere	7,120	**cwt**	Kuwaataay	14,288
daa	Dangaléat	7,104	**daf**	Dan	13,673	**dag**	Dagbani	9,656
dav	Dawida	2,944	**dbq**	Daba	10,081	**ddn**	Dendi	2,238
dga	Dagaare, Southern	6,991	**dgd**	Dagaari Dioula	2,942	**dgi**	Dagara, Northern	6,829
dhm	Dhimba	6,634	**dib**	Dinka, South Central	1,175	**did**	Didinga	5,898
dig	Chidigo	6,828	**dik**	Dinka, Southwestern	17,349	**din**	Dinka	340
dip	Dinka, Northeastern	6,051	**diu**	Gciriku	906	**dje**	Zarma	2,749
dks	Dinka, Southeastern	6,965	**dnj**	Dan	7,216	**dop**	Lukpa	15,230
dos	Dogosé	10,232	**dow**	Doyayo	6,643	**dsh**	Daasanach	6,019
dts	Dogon, Toro So	14,276	**dua**	Duala	8,520	**dug**	Chiduruma	6,824
dur	Dii	13,791	**dwr**	Dawro	7,582	**dyi**	Sénoufo, Djimini	7,105
dyo	Jola-Fonyi	3,800	**dyu**	Jula	20,860	**ebr**	Tchaman	2,939
ebu	Kiembu	2,915	**efi**	Efik	16,007	**ego**	Eggon	2,942
eka	Ekajuk	7,120	**eko**	Koti	4,636	**enb**	Markweeta	13,700

Table E.5: Languages covered in AFROSCOPE-DATA– Part I

ISO-3	Language	Sentences	ISO-3	Language	Sentences	ISO-3	Language	Sentences
eto	Eton	2,286	**etu**	Ejagham	6,635	**etx**	Iten	2,936
ewe	Éwé	42,255	**ewo**	Ewondo	6,829	**fak**	Fang	2,854
fal	Fali, South	14,292	**fan**	Fang	14,513	**ffm**	Fulfulde, Maasina	7,508
fia	Nobiin	345	**fip**	Fipa	2,944	**flr**	Fuliiru	2,924
fon	Fon	22,482	**fub**	Fulfulde, Adamawa	8,041	**fue**	Fulfulde, Borgu	6,615
fuf	Pular	7,355	**fuh**	Fulfulde, Western Niger	7,154	**ful**	Fulah	2,513
fuq	Fulfulde, Central-Eastern Niger	6,737	**fuv**	Fulfulde, Nigerian	16,142	**gaa**	Ga	16,520
gax	Oromo, Borana-Arsi-Guji	2,852	**gaz**	Oromo, West Central	37,860	**gbo**	Grebo, Northern	5,684
gbr	Gbagyi	6,617	**gde**	Gude	7,118	**gej**	Gen	10,257
gid	Gidar	6,628	**giz**	Giziga	7,947	**gjn**	Gonja	7,833
gkn	Gokana	9,191	**gkp**	Kpelle, Guinea	2,898	**gmv**	Gamo	7,987
gna	Kaansa	6,614	**gnd**	Zulgo-Gemzek	6,815	**gng**	Ngangam	7,106
goa	Guro	1,929	**gof**	Gofa	7,470	**gog**	Gogo	6,869
gol	Gola	2,941	**gqr**	Gor	7,129	**gso**	Gbaya, Southwest	7,168
gud	Dida, Yocoboué	6,629	**guk**	Gumuz	14,292	**gur**	Farefare	9,570
guw	Gun	14,422	**gux**	Gourmanchéma	7,617	**guz**	Ekegusii	6,546
gvl	Gulay	6,918	**gwr**	Gwere	6,046	**gya**	Gbaya, Northwest	8,038
hae	Oromo, Eastern	14,293	**hag**	Hanga	4,840	**har**	Harari	114
hau	Hausa	32,959	**hav**	Havu	11,013	**hay**	Haya	6,848
hbb	Nya Huba	2,939	**heh**	Hehe	7,169	**her**	Herero	10,708
hgm	Hailom	2,888	**hig**	Kamwe	3,696	**hna**	Mina	2,937
ibb	Ibibio	2,945	**ibo**	Igbo	35,306	**idu**	Idoma	8,803
ife	Ifè	14,528	**igb**	Ebira	2,942	**ige**	Igede	8,879
igl	Igala	2,947	**ijn**	Kalabari	2,932	**ikk**	Ika	8,221
ikw	Ikwere	7,128	**ilb**	Ila	13,584	**iqw**	Ikwo	6,631
iri	Rigwe	7,130	**irk**	Iraqw	14,292	**ish**	Esan	9,493
iso	Isoko	14,172	**iyx**	Yaka	853	**izr**	Izere	7,132
izz	Izii	8,063	**jbu**	Jukun Takum	14,292	**jgo**	Ngomba	2,939
jib	Jibu	2,933	**jit**	Jita	2,939	**jmc**	Machame	6,752
kab	Kabyle	36,057	**kam**	Kamba	24,505	**kao**	Xaasongaxango	14,292
kbn	Kare	2,939	**kbo**	Keliko	5,192	**kbp**	Kabiyè	26,519
kbr	Kafa	14,182	**kby**	Kanuri, Manga	2,502	**kcg**	Tyap	6,639
kck	Kalanga	7,593	**kdc**	Kutu	7,122	**kde**	Makonde	7,631
kdh	Tem	4,320	**kdi**	Kumam	7,020	**kdj**	Ng'akarimojong	6,717
kdl	Tsikimba	7,119	**kdn**	Kunda	2,924	**kea**	Kabuverdianu	21,334
ken	Kenyang	7,120	**keo**	Kakwa	2,363	**ker**	Kera	10,211
kez	Kukele	14,291	**khq**	Songhay, Koyra Chiini	10,230	**khy**	Kele	6,639
kia	Kim	8,214	**kik**	Gikuyu	28,959	**kin**	Kinyarwanda	53,340
kiz	Kisi	2,885	**kki**	Kagulu	7,132	**kkj**	Kako	7,127
kln	Kalenjin	2,871	**klu**	Klao	2,932	**kma**	Konni	6,042
kmb	Kimbundu	26,546	**kmy**	Koma	5,307	**knc**	Kanuri, Yerwa	10,991
knf	Mankanya	7,194	**kng**	Koongo	7,442	**knk**	Kuranko	6,818
kno	Kono	6,805	**kny**	Kanyok	5,261	**kon**	Kongo	2,948
koo	Konzo	11,112	**koq**	Kota	390	**kpz**	Kupsapiiny	3,741
kqn	Kaonde	13,753	**kqo**	Krahn, Eastern	13,800	**kqp**	Kimré	7,125
kqs	Kissi, Northern	6,629	**kqy**	Koorete	6,796	**kri**	Krio	10,861
krs	Gbaya	2,947	**krw**	Krahn, Western	2,932	**krx**	Karon	1,201
ksb	Shambala	6,627	**ksf**	Bafia	6,655	**ksp**	Kabba	6,383

Table E.6: Languages covered in AFROSCOPE-DATA– Part II

ISO-3	Language	Sentences	ISO-3	Language	Sentences	ISO-3	Language	Sentences
kss	Kisi, Southern	7,796	**ktb**	Kambaata	14,253	**ktj**	Krumen, Plapo	7,098
ktu	Kituba	34,578	**kua**	Oshiwambo	13,681	**kub**	Kutep	7,127
kuj	Kuria	6,616	**kus**	Kusaal	6,834	**kvj**	Psikye	6,573
kwn	Kwangali	12,072	**kwy**	Kikongo	20,333	**kxc**	Konso	14,289
kyf	Kouya	6,851	**kyq**	Kenga	7,512	**kzn**	Kokola	2,789
kzr	Karang	2,695	**lai**	Lambya	6,618	**laj**	Lango	7,356
lam	Lamba	8,235	**lap**	Laka	4,931	**las**	Lama	10,206
ldi	Laari	15,818	**lea**	Lega-Shabunda	4,750	**led**	Lendu	5,573
lee	Lyélé	7,109	**lef**	Lelemi	6,826	**leh**	Lenje	15,944
lem	Nomaande	7,128	**lgg**	Lugbara	8,842	**lgm**	Lega-Mwenga	6,631
lia	Limba, West-Central	6,832	**lik**	Lika	2,929	**lin**	Lingala	43,654
lip	Sekpele	7,126	**llb**	Lolo	14,888	**lln**	Lele	12,030
lmd	Lumun	2,656	**lmp**	Limbum	6,628	**lnl**	Banda, South Central	2,924
lob	Lobi	14,293	**log**	Logo	4,561	**lok**	Loko	10,225
lol	Mongo-Nkundu	13,799	**lom**	Loma	6,396	**loq**	Lobala	6,308
lot	Otuho	2,939	**loz**	Lozi	15,985	**lro**	Laro	2,939
lsm	Saamya-Gwe	6,815	**lth**	Thur	2,895	**lto**	Olutsotso	2,885
lua	Luba-Kasai	39,225	**lub**	Luba-Katanga	14,269	**luc**	Aringa	6,501
lue	Luvale	14,312	**lug**	Ganda	31,230	**lun**	Lunda	13,024
luo	Dholuo	32,868	**lwg**	Oluwanga	3,411	**lwo**	Luwo	7,107
maf	Mafa	6,642	**mas**	Maasai	8,400	**maw**	Mampruli	7,113
mbu	Mbula-Bwazza	2,945	**mck**	Mbunda	10,430	**mcn**	Masana	7,797
mcp	Makaa	7,356	**mcu**	Mambila, Cameroon	6,825	**mda**	Mada	7,128
mdm	Mayogo	2,932	**mdy**	Male	8,977	**men**	Mende	6,891
meq	Merey	7,120	**mer**	Kimîîru	4,678	**mev**	Maan	2,415
mfe	Morisyen	13,701	**mfg**	Mogofin	4,216	**mfh**	Matal	6,833
mfi	Wandala	7,118	**mfk**	Mofu, North	7,130	**mfq**	Moba	6,677
mfz	Mabaan	4,007	**mgc**	Morokodo	5,449	**mgh**	Makhuwa-Meetto	9,323
mgo	Meta'	6,638	**mgq**	Malila	2,939	**mgr**	Mambwe-Lungu	10,507
mgw	Matumbi	1,636	**mhi**	Ma'di	3,722	**mhw**	Mbukushu	45
mif	Mofu-Gudur	7,103	**mkl**	Mokole	7,119	**mlg**	Malagasy	2,481
mlr	Vame	887	**mmy**	Migaama	2,710	**mnf**	Mundani	6,821
mnk	Mandinka	6,832	**mny**	Manyawa	15,521	**moa**	Mwan	7,127
mor	Moro	3,695	**mos**	Moore	30,772	**moy**	Shekkacho	2,945
moz	Mukulu	2,928	**mpe**	Majang	2,932	**mpg**	Marba	7,119
mqb	Mbuko	7,122	**msc**	Maninka, Sankaran	5,067	**mse**	Musey	13,796
mua	Mundang	13,796	**mug**	Musgu	15,112	**muh**	Mündü	14,539
mur	Murle	6,814	**muy**	Muyang	6,933	**mwe**	Mwera	2,946
mwm	Sar	7,955	**mwn**	Nyamwanga	8,332	**mws**	Mwimbi-Muthambi	877
myb	Mbay	7,124	**myk**	Sénoufo, Mamara	7,155	**myx**	Masaaba	9,251
mzk	Mambila, Nigeria	14,293	**mzm**	Mumuye	7,139	**mzw**	Deg	7,089
naq	Khoekhoe	9,663	**naw**	Nawuri	7,133	**nba**	Nyemba	10,339
nbl	Ndebele	13,722	**ncu**	Chumburung	7,081	**ndc**	Ndau	11,260
nde	Ndebele	11,979	**ndh**	Ndali	2,455	**ndi**	Samba Leko	13,793
ndj	Ndamba	7,135	**ndo**	Ndonga	13,803	**ndp**	Kebu	14,293
ndv	Ndut	2,728	**ndy**	Luto	10,245	**ndz**	Ndogo	7,130
neb	Toura	14,284	**nfr**	Nafaanra	14,290	**ngb**	Ngbandi, Northern	4,833
ngc	Ngombe	6,639	**ngl**	Lomwe	10,462	**ngn**	Ngwo	2,899

Table E.7: Languages covered in AFROSCOPE-DATA– Part III

ISO-3	Language	Sentences	ISO-3	Language	Sentences	ISO-3	Language	Sentences
ngp	Ngulu	7,137	**nhr**	Naro	7,113	**nhu**	Noone	7,111
nih	Nyiha, Tanzania	2,939	**nim**	Nilamba	7,137	**nin**	Ninzo	7,253
niq	Nandi	7,786	**niy**	Ngiti	6,638	**nka**	Nkoya	2,842
nko	Nkonya	7,107	**nla**	Ngombale	2,341	**nmz**	Nawdm	14,289
nnb	Nande	18,686	**nnh**	Ngiemboon	6,324	**nnq**	Ngindo	7,133
nnw	Nuni, Southern	14,526	**nqo**	N'Ko	12,823	**nse**	Nsenga	9,637
nso	Sotho, Northern	46,564	**ntr**	Delo	7,129	**nuj**	Nyole	7,658
nus	Nuer	11,861	**nwb**	Nyabwa	6,776	**nxd**	Ngando	6,644
nya	Chichewa	51,207	**nyb**	Nyagbo	2,940	**nyd**	Olunyole	2,942
nyf	Kigiryama	6,848	**nyk**	Nyaneka	11,885	**nym**	Nyamwezi	2,934
nyn	Nyankore	11,475	**nyo**	Nyoro	6,841	**nyu**	Nyungwe	10,624
nyy	Nyakyusa-Ngonde	9,740	**nza**	Mbembe, Tigon	5,188	**nzi**	Nzema	12,609
odu	Odual	2,938	**ogo**	Khana	8,464	**oke**	Okpe	8,523
okr	Kirike	2,941	**oku**	Oku	6,048	**old**	Mochi	13,340
orm	Oromo	2,792	**ozm**	Koonzime	6,799	**pbi**	Parkwa	14,279
pcm	Pidgin, Nigerian	13,317	**pem**	Phende	6,198	**pfe**	Pere	14,395
phm	Phimbi	15,399	**pkb**	Kipfokomu	6,828	**pko**	Pökoot	2,783
plt	Malagasy, Merina	27,852	**pny**	Pinyin	13,238	**pov**	Guinea-Bissau Creole	7,251
poy	Pogolo	6,644	**rag**	Lulogooli	2,900	**rcf**	Réunion French Creole	13,722
rel	Rendille	6,056	**rif**	Tarifit	2,941	**rim**	Nyaturu	7,137
rnd	Ruund	7,481	**rng**	Ronga	8,976	**rub**	Gungu	6,636
ruf	Luguru	14,533	**run**	Rundi	44,213	**rwk**	Rwa	2,633
sag	Sango	36,544	**saq**	Samburu	2,946	**sba**	Ngambay	8,083
sbd	Samo, Southern	6,654	**sbp**	Sangu	2,939	**sbs**	Kuhane	2,540
sby	Soli	2,407	**sef**	Sénoufo, Cebaara	2,935	**seh**	Sena	19,145
ses	Songhay, Koyraboro Senni	7,372	**sev**	Sénoufo, Nyarafolo	2,930	**sfw**	Esahie	6,359
sgc	Kipsigis	11,612	**sgw**	Sebat Bet Gurage	6,828	**shi**	Tachelhit	8,390
shj	Shatt	766	**shk**	Shilluk	5,943	**shr**	Shi	14,113
shu	Arabic, Chadian	6,626	**sid**	Sidaama	9,955	**sig**	Paasaal	7,124
sil	Sisaala, Tumulung	6,827	**skg**	Malagasy, Sakalava	15,615	**sld**	Sissala	14,290
sna	Shona	45,067	**snf**	Noon	6,634	**sng**	Sanga	2,913
snw	Selee	6,863	**soe**	Ohendo	717	**som**	Somali	27,870
sop	Songe	11,004	**sor**	Soumraye	1,083	**sot**	Sotho, Southern	13,079
soy	Miyobe	7,118	**spp**	Sénoufo, Supyire	7,114	**spy**	Sabaot	14,640
srr	Serer-Sine	3,905	**ssw**	Swati	27,224	**suk**	Sukuma	7,344
sur	Mwaghavul	14,293	**sus**	Susu	8,885	**swa**	Swahili	2,918
swb	Comorian, Maore	237	**swc**	Swahili, Congo	28,449	**swh**	Swahili	37,548
swk	Sena, Malawi	5,962	**sxb**	Suba	7,121	**tap**	Taabwa	10,437
taq	Tamasheq	15,981	**tbz**	Ditammari	5,589	**tcc**	Datooga	7,128
tcd	Tafi	2,947	**tdx**	Malagasy, Tandroy-Mahafaly	8,366	**ted**	Krumen, Tepo	6,861
tem	Themne	6,840	**teo**	Ateso	8,489	**tex**	Tennet	2,942
tgw	Sénoufo, Tagwana	2,934	**thk**	Kitharaka	6,820	**thv**	Tamahaq, Tahaggart	1,946
tig	Tigré	3,105	**tik**	Tikar	14,289	**tir**	Tigrigna	36,754
tiv	Tiv	13,164	**tke**	Takwane	6,888	**tlj**	Talinga-Bwisi	7,117
tll	Tetela	13,894	**tmc**	Tumak	10,258	**tnr**	Ménik	9,951
tod	Toma	7,199	**tog**	Tonga	11,237	**toh**	Tonga	9,554
toi	Tonga	15,080	**tpm**	Tampulma	8,553	**tsc**	Tswa	12,065
tsn	Setswana	35,384	**tso**	Tsonga	38,043	**tsw**	Tsishingini	7,125

Table E.8: Languages covered in AFROSCOPE-DATA– Part IV

ISO-3	Language	Sentences	ISO-3	Language	Sentences	ISO-3	Language	Sentences
ttj	Tooro	9,664	**ttq**	Tamajaq, Tawallammat	6,624	**ttr**	Tera	2,946
tui	Tupuri	7,698	**tul**	Tula	5,200	**tum**	Tumbuka	34,918
tuv	Turkana	2,383	**tvu**	Tunen	2,942	**twi**	Twi	43,976
twx	Tewe	14,947	**tzm**	Tamazight, Central Atlas	2,472	**udu**	Uduk	4,024
umb	Umbundu	32,657	**urh**	Urhobo	10,605	**uth**	ut-Hun	5,189
vag	Vagla	7,010	**vai**	Vai	2,940	**ven**	Venda	17,236
vid	Vidunda	7,117	**vif**	Vili	2,942	**vmk**	Makhuwa-Shirima	4,231
vmw	Makhuwa	11,968	**vun**	Vunjo	7,131	**vut**	Vute	7,120
wal	Wolaytta	12,310	**wbi**	Vwanji	2,937	**wec**	Wè Western	2,691
wes	Pidgin, Cameroon	8,959	**wib**	Toussian, Southern	6,158	**wlx**	Wali	14,628
wmw	Mwani	7,125	**wob**	Wè Northern	14,529	**wol**	Wolof	17,637
won	Wongo	512	**wwa**	Waama	3,694	**xan**	Xamtanga	1,474
xed	Hdi	7,103	**xho**	Xhosa	37,847	**xmv**	Malagasy, Antankarana	15,974
xnz	Mattokki	2,939	**xog**	Soga	7,860	**xon**	Konkomba	7,511
xpe	Kpelle, Liberia	2,864	**xrb**	Karaboro, Eastern	6,933	**xsm**	Kasem	7,519
xtc	Katcha-Kadugli-Miri	2,946	**xuo**	Kuo	7,113	**yal**	Yalunka	7,086
yam	Yamba	7,124	**yao**	Yao	11,100	**yas**	Nugunu	2,168
yat	Yambeta	6,027	**yaz**	Lokaa	10,258	**yba**	Yala	2,744
ybb	Yemba	6,301	**yom**	Kiyombe	9,055	**yor**	Yoruba	48,191
yre	Yaouré	7,668	**zaj**	Zaramo	2,937	**zdj**	Comorian, Ngazidja	4,106
zga	Kinga	2,941	**zgh**	Tamazight, Standard Moroccan	5,817	**ziw**	Zigula	7,137
zne	Zande	11,952	**zul**	Zulu	37,885			

Table E.9: Languages covered in AFROSCOPE-DATA– Part V