AfroLID: A Neural Language Identification Tool for African Languages

Ife Adebara[1,⋆] **AbdelRahim Elmadany**[1,⋆] **Muhammad Abdul-Mageed**[1,2] **Alcides Alcoba Inciarte**[1]
[1]Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{ife.adebara@,a.elmadany@,muhammad.mageed@,alcobaaj@mail.}ubc.ca

Abstract

Language identification (LID) is a crucial precursor for NLP, especially for mining web data. Problematically, most of the world's 7000+ languages today are not covered by LID technologies. We address this pressing issue for Africa by introducing AfroLID, a neural LID toolkit for 517 African languages and varieties. AfroLID exploits a multi-domain web dataset manually curated from across 14 language families utilizing five orthographic systems. When evaluated on our blind Test set, AfroLID achieves 95.89 F_1-score. We also compare AfroLID to five existing LID tools that each cover a small number of African languages, finding it to outperform them on most languages. We further show the utility of AfroLID in the wild by testing it on the acutely under-served Twitter domain. Finally, we offer a number of controlled case studies and perform a linguistically-motivated error analysis that allow us to both showcase AfroLID's powerful capabilities and limitations.[1]

1 Introduction

Language identification (LID) is the task of identifying the human language a piece of text or speech segment belongs to. The proliferation of social media have allowed greater access to multilingual data, making automatic LID an important first step in processing human language appropriately (Tjandra et al., 2021; Thara and Poornachandran, 2021). This includes applications in speech, sign language, handwritten text, and other modalities of language. It also includes distinguishing languages in code-mixed datasets (Abdul-Mageed et al., 2020; Thara and Poornachandran, 2021). Unfortunately, for the majority of languages in the world, including most African languages, we do not have the resources for developing LID tools.

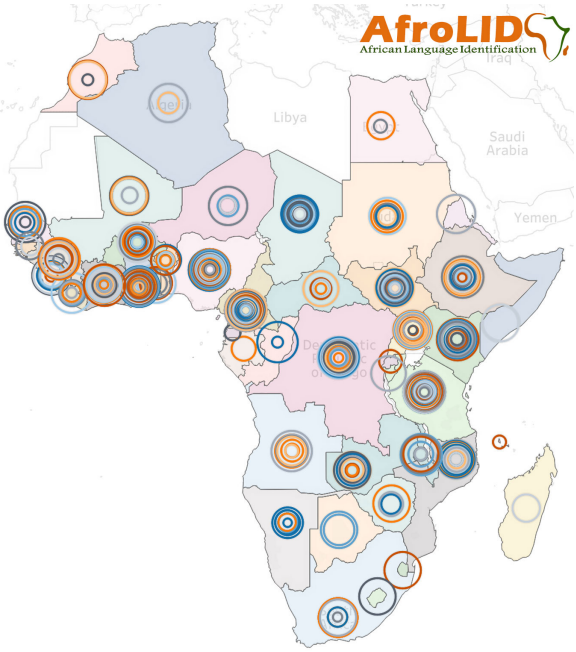

Figure 1: All 50 African countries in our data, with our 517 languages/language varieties in colored circles overlayed within respective countries. More details are in Appendix E.

This situation has implications for the future NLP technologies. For instance, LID has facilitated development of widely multilingual models such mT5 (Xue et al., 2021) and large multilingual datasets such as CCAligned (El-Kishky et al., 2020), ParaCrawl (Esplà et al., 2019), WikiMatrix (Schwenk et al., 2021), OSCAR (Ortiz Suárez et al., 2020), and mC4 (Xue et al., 2021) which have advanced research in NLP. Comparable resources are completely unavailable for the majority of the world's 7000+ today, with only poor coverage of the so-called low-resource languages (LR). This is partly due to absence of LID tools, and impedes future NLP progress on these languages (Adebara and Abdul-Mageed, 2022). The state of African languages is not any better than other regions: Kreutzer et al. (2021) perform a manual evaluation of 205 datasets involving African languages such as those in CCAligned, ParaCrawl, WikiMatrix, OSCAR, and mC4 and show that at

⋆ Authors contributed equally.
[1]AfroLID is publicly available at https://github.com/UBC-NLP/afrolid.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1958–1981
December 7-11, 2022

least 15 corpora were completely erroneous, a significant fraction contained less than 50% of correct data, and 82 corpora were mislabelled or used ambiguous language codes. These consequently affect the quality of models built with these datasets. Alabi et al. (2020) find that 135K out of 150K words in the fastText embeddings for Yorùbá belong to other languages such as English, French, and Arabic. New embedding models created by Alabi et al. (2020) with a curated high quality dataset outperform off-the-shelf fastText embeddings, even though the curated data is smaller.

In addition to resource creation, lack (or poor performance) of LID tools negatively impacts preprocessing of LR languages since LID can be a prerequisite for determining, e.g., appropriate tokenization. (Duvenhage et al., 2017a). Furthermore, some preprocessing approaches may be necessary for certain languages, but may hurt perforrmance in other languages (Adebara and Abdul-Mageed, 2022). Developing LID tools is thus vital for all NLP. In this work, we focus on LID for African languages and introduce AfroLID.

AfroLID is a neural LID tool that covers 517 African languages and language varieties[2] across 14 language families. The languages covered belong to 50 African countries and are written in five diverse scripts. We show the countries covered by AfroLID in Figure 1. Examples of the different scripts involved in the 517 languages are displayed in Figure 2. To the best of our knowledge, AfroLID supports the *largest* subset of African languages to date. AfroLID is also usable without any end-user training, and it exploits data from a variety of domains to ensure robustness. We manually curate our clean training data, which is of special significance in low resource settings. We show the utility of AfroLID in the wild by applying it on two Twitter datasets and compare its performance with existing LID tools that cover any number of African languages such as CLD2 (McCandless, 2010), CLD3 (Salcianu et al., 2018), Franc, LangDetect (Shuyo, 2010), and Langid.py (Lui and Baldwin, 2012). Our results show that AfroLID consistently outperforms *all* other LID tools for almost all languages, and serves as the new SOTA for language identification for African languages.

To summarize, we offer the following main contributions:

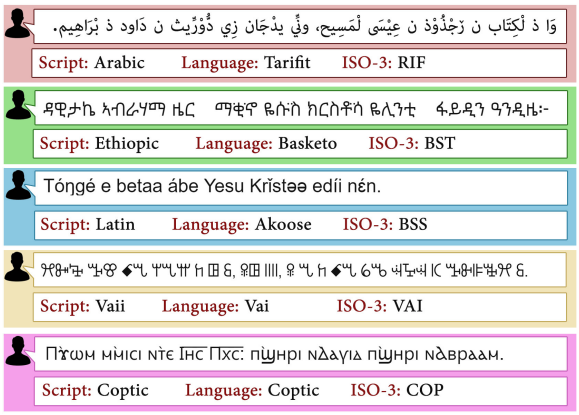

Figure 2: Examples from the five scripts in our data.

1. We develop AfroLID, a SOTA LID tool for 517 African languages and language varieties. To facilitate NLP research, we make our models publicly available.

2. We carry out a study of LID tool performance on African languages where we compare our models in controlled settings with several tools such as CLD2, CLD3, Franc, LangDetect, and Langid.py.

3. Our models exhibit highly accurate performance in the wild, as demonstrated by applying AfroLID on Twitter data.

4. We provide a wide range of controlled case studies and carry out a linguistically-motivated error analysis of AfroLID. This allows us to motivate plausible directions for future research, including potentially beyond African languages.

The rest of the paper is organized as follows: In Section 2 we discuss a number of typological features of our supported languages. We describe AfroLID's training data in Section 3. Next, we introduce AfroLID in 4. This includes our experimental datasets and their splits, preprocessing, vocabulary, implementation and training details, and our evaluation settings. We present performance of AfroLID in Section 5 and compare it to other LID tools. Our analysis show that AfroLID outperforms other models for most languages. In the same section, we also describe the utility of AfroLID on non-Latin scripts, Creole languages, and languages in close geographical proximity. Although AfroLID is not trained on Twitter data, we experiment with tweets in Section 6 in

[2]Our dataset involves different forms that can arguably be viewed as varieties of the same language such as Twi and Akan.

order to investigate performance of AfroLID in out of domain scenarios. Through two diagnostic studies, we demonstrate AfroLID's robustness. We provide an overview of related work in Section 7. We conclude in Section 8, and outline a number of limitations for our work in Section 9.

2 Typological Information

Language Families. We experiment with 517 African languages and language varieties across 50 African countries. These languages belong to 14 language families (Eberhard et al., 2021) as follows: Afro-Asiatic, Austronesian, Creole (English based), Creole (French based), Creole (Kongo based), Creole (Ngbadi based), Creole (Portuguese based), Indo-European, Khoe-Kwadi (Hainum), Khoe-Kwadi (Nama), Khoe-Kwadi (Southwest), Niger-Congo, and Nilo-Saharan. The large and typologically diverse data we exploit hence endow our work with wide coverage. We show in Figure 1 a map of Africa with the countries AfroLID covers. We also show the number of languages we cover, per country, in Figure E in the Appendix. Table E.1, Table E.2, and Table E.3 in the Appendix also provide a list of the languages AfroLID handles. We represent the languages using ISO-3 codes[3] for both individual languages and macro-languages. We use a macro-language tag when the language is known but the specific dialect is unknown. For this reason we specify that AfroLID supports 517 African languages and language varieties.

Sentential Word Order. There are seven categories of word order across human languages around the world. These are subject-verb-object (SVO), subject-object-verb (SOV), object-verb-subject (OVS), object-subject-verb (OSV), verb-object-subject (VOS), verb-subject-object (VSO), and languages lacking a dominant order (which often have a combination of two or more orders within its grammar) (Dryer and Haspelmath, 2013). Again, our dataset is very diverse: we cover five out of these seven types of word order. Table 1 shows sentential word order in our data, with some representative languages for each category.

Diacritics. Diacritic marks are used to overcome the inadequacies of an alphabet in capturing important linguistic information by adding a distinguishing mark to a character in an alphabet. Diacritics are often used to indicate tone, length, case, nasalization, or even to distinguish different letters of a language's alphabet (Wells, 2000; Hyman, 2003; Creissels et al., 2008). Diacritics can be placed above, below or through a character. Diacritics are common features of the orthographies of African languages. Out of 517 languages/language varieties in our training data, 295 use some diacritics in their orthographies. We also provide a list of languages with diacritics in our training data in Table C.3 in the Appendix.

[3] https://glottolog.org/glottolog/language.

Word Order	Example Languages
SVO	Xhosa, Zulu, Yorùbá
SOV	Khoekhoe, Somali, Amharic
VSO	Murle, Kalenjin
VOS	Malagasy
No-dominant-order	Siswati, Nyamwezi, Bassa

Table 1: Sentential word order in our data.

Script	Languages
Ethiopic	Amharic, Basketo, Maale, *Oromo, Sebat Bet Gurage Tigrinya, Xamtanga
Arabic	Fulfude Adamawa, Fulfude Caka Tarifit
Vai	Vai
Coptic	Coptic

Table 2: Non-Latin scripts in AfroLID data. *Oromo: is available in Latin script as well.

Scripts. Our dataset consists of 14 languages written in four different non-Latin scripts and 499 languages written in Latin scripts. The non-Latin scripts are Ethiopic, Arabic, Vai, and Coptic.

3 Curating an African Language Dataset

AfroLID is trained using a multi-domain, multi-script language identification dataset that we manually curated for building our tool. To collect the dataset, we perform an extensive manual analysis of African language presence on the web, identifying as much publicly available data from the 517 language varieties we treat as is possible. We adopt this manual curation approach since there are only few African languages that have any LID tool coverage. In addition, available LID tools that treat African languages tend to perform unreliably (Kreutzer et al., 2021). We therefore consult research papers focusing on African languages, such as (Adebara and Abdul-Mageed, 2022), or provide language data (Muhammad et al., 2022; Alabi et al., 2020), sifting through references to find additional African data sources. Moreover,

we search for newspapers across all 54 African countries.[4] We also collect data from social media such as blogs and web fora written in African languages as well as databases that store African language data. These include LANAFRICA, SADiLaR, Masakhane, Niger-Volta-LTI, and ALTI. Our resulting multi-domain dataset contains religious texts, government documents, health documents, crawls from curated web pages, news articles, and existing human-identified datasets for African languages. As an additional sanity check, we ask a number of native speakers from a subset of the languages to verify the correctness of the self-labels assigned in respective sources within our collections.[5] Our manual inspection step gave us confidence about the quality of our dataset, providing near perfect agreement by native speakers with labels from data sources. In total, we collect 100 million sentences in 528 languages across 14 language families in Africa and select 517 languages which had at least 2000 sentences. Again, the dataset has various orthographic scripts, including 499 languages in Latin scripts, eight languages in Ethiopic scripts, four languages in Arabic scripts, one language in Vai scripts, and one in Coptic scripts.

4 AfroLID

Experimental Dataset and Splits. From our manually-curated dataset, we randomly select 5,000, 50, and 100 sentences for train, development, and test, respectively, for each language.[6] Overall, AfroLID data comprises 2,496,980 sentences for training (Train), 25,850 for development (Dev), and 51,400 for test (Test) for 517 languages and language varieties.

Preprocessing. We ensure that our data represent naturally occurring text by performing only minimal preprocessing. Specifically, we tokenize our data into character, byte-pairs, and words. We do not remove diacritics and use both precomposed and decomposed characters to cater for the inconsistent use of precomposed and decomposed characters by many African languages in digital media.[7] We create our character level tokenization scripts and generate our vocabulary using Fairseq. We use sentencepiece tokenizer for the word level and byte-pair tokens before we preprocess in Fairseq.

Vocabulary. We experiment with byte-pair (BPE), word, and character level encodings. We used vocabulary sizes of 64K, 100K, and 2,260 for the bpe, word, and character level models across the 517 language varieties. The characters included both letters, diacritics, and symbols from other non-Latin scripts for the respective languages.

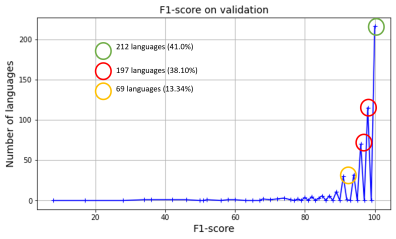

Figure 3: F_1 distribution on AfroLID Dev set.

Implementation. AfroLID is built using a Transformer architecture trained from scratch. We use 12 attention layers with 12 heads in each layer, 768 hidden dimensions, making up $\sim$ 200M parameters.[8]

Hyperparameter Search and Training. To identify our best hyperparameters, we use a subset of our training data and the full development set for our hyperparameter search. Namely, we randomly sample 200 examples from each language in our training data to create a smaller train set,[9] while using our full Dev set. We train for up to 100 epochs, with early stopping. We search for the following hyperparameter values, picking bolded ones as our best: dropout rates from the set *{**0.1**, 0.2, 0.3, 0.4, 0.5}*, learning rates from *{5e-5, **5e-6**}*, and patience from *{**10**, 20, 30}*. Other hyperparameters are similar to those for XML-R (Conneau et al., 2020). We perform hyperparameter search only with our character level model and use identified values with both the BPE and word models.

Evaluation. We report our results in both macro F_1-score and accuracy, selecting our best model on

[4] https://www.worldometers.info/geography/how-many-countries-in-africa/.

[5] We had access to native speakers of Afrikaans, Yorùbá, Igbo, Hausa, Luganda, Kinyarwanda, Chichewa, Shona, Somali, Swahili, Xhosa, Bemba, and Zulu.

[6] We remove languages with data less than 2,000 sentences, as explained earlier.

[7] A Unicode entity that combines two or more other characters may be precomposed or decomposed. For example, ä can be precomposed into $U+0061U+0308$ or decomposed into $U+00E4$. In Unicode, they are included primarily to aid computer systems with incomplete Unicode support, where equivalent decomposed characters may render incorrectly.

[8] This architecture is similar to XMLRBase (Conneau et al., 2020).

[9] This helps us limit GPU hours needed for hyperparameter search.

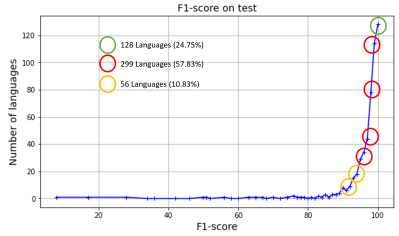

Figure 4: F_1 distribution on AfroLID Test set.

Dev based on F_1. *For all our models, we report the average of three runs*.

5 Model Performance and Analysis

As Table 3 shows, our **BPE model** outperforms both the **char** and **word** models on both Dev and Test data. On Dev, our BPE model acquires 96.14 F_1 and 96.19 acc, compared to 85.75 F_1 and 85.85 for char model, and 90.22 F_1 and 90.34 acc for word model, respectively. Our BPE model similarly excels on Test, with 95.95 F_1 and 96.01 acc. We inspect the distribution of F_1 on the entire Dev and Test sets using our BPE model, as shown in Figures 3 and 4. As annotated on Figure 3, a total of 212 languages out of the 517 ($\% = 41$) are identified with 100 F_1, 197 languages ($\% = 38.10$) identified with 95 and 99 F_1, and 69 languages ($\% = 13.30$) identified with 90–95 F_1. For Test data (Figure 4), on the other hand, 128 ($\% = 24.75$) languages are identified with 100 F_1, 299 languages ($\% = 57.83$) are between 95–99 F_1, while 56 languages ($\% = 10.83$) are between 90–95 F_1.

Model	Split	F_1-score	Accuracy	Checkpoint
Char	Dev	85.75	85.85	69
	Test	81.20	81.30	
BPE	Dev	96.14	96.19	73
	Test	95.95	96.01	
Word	Dev	90.22	90.34	65
	Test	89.04	89.01	

Table 3: Results on the BPE, word level, and character level models. **Bolded**: best result on Test. Underlined: best result on Dev.

AfroLID in Comparison Using our Dev and Test data, we compare our best AfroLID model (BPE model) with the following LID tools: CLD2, CLD3, Franc, LangDetect, and Langid.py. Since these tools do not support all our AfroLID languages, we compare accuracy and F_1-scores of our models only on languages supported by each of these tools. As Tables A.1 and 4 show, AfroLID outperforms other tools on 7 and 8 languages out of 16 languages on the Dev set and Test set, respectively. We also compare F_1-scores of **Franc** on the 88 African languages Franc supports with the F_1-scores of AfroLID on those languages. As shown in Tables 5 and 6, AfroLID outperforms Franc on 78 languages and has similar F_1-score on five languages on the Dev set. AfroLID also outperforms Franc on 76 languages, and has similar F_1-score on five languages on the Test set.

Lang.	CLD2	CLD3	Langid.py	LangDetect	Franc	AfroLID
afr	94.00	91.00	69.00	88.23	81.00	**97.00**
amh	-	97.00	**100.00**	-	35.00	97.00
hau	-	83.00	-	-	77.00	**88.00**
ibo	-	96.00	-	-	88.00	**97.00**
kin	**92.00**	-	45.00	-	47.00	89.00
lug	84.00	-	-	-	64.00	**87.00**
mlg	-	**100.00**	98.00	-	-	**100.00**
nya	-	**96.00**	-	-	75.00	92.00
sna	-	**100.00**	-	-	91.00	97.00
som	-	92.00	-	-	89.00	**95.00**
sot	-	**99.00**	-	-	93.00	88.00
swa	99.00	91.00	90.00	**100.00**	-	92.00
swc	93.00	94.00	96.00	**97.02**	-	87.00
swh	89.00	**92.00**	88.23	87.19	70.00	77.00
xho	-	59.00	**88.00**	-	30.00	67.00
yor	-	25.00	-	-	66.00	**98.00**
zul	-	**89.00**	20.00	-	40.00	50.00

Table 4: A comparison of results on AfroLID with CLD2, CLD3, Langid.py, LangDetect, and Franc using F_1-score on the Test set. – indicates that the tool does not support the language.

Effect of Non-Latin Script. We investigate performance of AfroLID on languages that use one of Arabic, Ethiopic, Vai, and Coptic scripts. Specifically, we investigate performance of AfroLID on Amharic (amh), Basketo (bst), Maale (mdy), Sebat Bet Gurage (sgw), Tigrinya (tir), Xamtanga (xan), Fulfude Adamawa (fub), Fulfude Caka (fuv), Tarif (rif), Vai (vai), and Coptic (cop).[10] Vai and Coptic, the two unique scripts in AfroLID have an F_1-score of 100 each. This corroborates research findings that languages written in unique scripts within an LID tool can be identified with up to 100% recall, F_1-score, and/or accuracy even using a small training dataset (Jauhiainen et al., 2017a). We assume this to be the reason Langid.py outperforms AfroLID on Amharic as seen in Table 4, since Amharic is the only language that employs an Ethiopic script in langid.py. AfroLID, on the other hand, has 8 languages using Ethiopic scripts. However, it is not clear why Basketo, which uses Ethiopic scripts has 100 F_1-score. We, how-

[10]We do not investigate performance on Oromo because we had both Latin and Ethiopic scripts for Oromo in our training data.

ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc
aar	**100.00**	74.50	fat	**94.11**	88.23	koo	**96.07**	86.27	nso	**84.31**	70.58	tir	98.03	**100.00**
ada	**98.03**	96.07	fon	**98.03**	86.27	kqn	**96.07**	86.27	nya	**96.07**	82.35	tiv	**100.00**	98.03
afr	**94.11**	84.31	fuf	**98.03**	60.78	kqs	**100.00**	64.70	nym	**100.00**	52.94	toi	**100.00**	68.62
amh	**98.03**	25.49	fuv	**90.19**	35.29	ktu	**96.07**	17.64	nyn	**92.15**	84.31	tsn	**70.58**	54.90
bam	**70.58**	45.09	gaa	**96.07**	**96.07**	lia	**98.03**	**98.03**	nzi	**98.03**	**98.03**	tso	**96.07**	80.39
bba	**98.03**	88.23	gaz	**96.07**	90.19	lin	**98.03**	96.07	pcm	**98.03**	78.43	twi	**90.19**	84.31
bci	76.47	**86.27**	gjn	**100.00**	94.11	lot	**100.00**	94.11	pov	**96.07**	86.27	umb	**90.19**	70.58
bem	**82.35**	64.70	gkp	64.70	**68.62**	loz	**96.07**	94.11	run	**84.31**	58.82	vai	**100.00**	**100.00**
bfa	**100.00**	90.19	hau	**94.11**	82.35	lua	**98.03**	96.07	sag	**94.11**	17.64	ven	**96.07**	**96.07**
bin	94.11	**98.03**	ibb	**98.03**	86.27	lue	**90.19**	60.78	shk	**100.00**	96.07	vmw	**88.23**	80.39
bum	**100.00**	52.94	ibo	**94.11**	90.19	lug	**86.27**	52.94	sna	**96.07**	80.39	wol	**68.62**	23.52
cjk	**98.03**	52.94	kbp	**98.03**	94.11	lun	**98.03**	90.19	som	**98.03**	96.07	xho	**82.35**	64.70
crs	**94.11**	82.35	kde	**96.07**	78.43	men	**98.03**	92.15	sot	**76.47**	90.19	xsm	**100.00**	25.49
dag	**96.07**	96.07	kdh	**100.00**	92.15	mfq	**96.07**	01.96	ssw	**90.19**	84.31	yor	**100.00**	39.21
dga	**100.00**	88.23	kea	**98.03**	3.92	mos	**94.11**	84.31	suk	**100.00**	31.37	zdj	**100.00**	62.74
dip	**98.03**	84.31	kin	**80.39**	52.94	nba	**100.00**	56.86	sus	**100.00**	96.07	zul	**58.82**	37.25
dyu	**98.03**	01.96	kmb	**100.00**	80.39	nbl	**80.39**	64.70	swh	**74.50**	72.54			
ewe	94.11	**96.07**	kng	**98.03**	66.66	ndo	**90.19**	82.35	tem	**96.07**	84.31			
AfroLID Average F_1-score: **93.21**												**Franc Average F_1-score**: 72.85		

Table 5: F_1-scores on our Dev dataset for languages in AfroLID and Franc for 88 languages.

ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc	ISO-3	AfroLID	Franc
aar	**96.00**	74.00	fat	**98.00**	94.00	koo	**96.00**	**96.00**	nso	**83.00**	59.00	tir	**99.00**	97.00
ada	**100.00**	98.00	fon	**97.00**	92.00	kqn	**98.00**	84.00	nya	**92.00**	75.00	tiv	**100.00**	99.00
afr	**97.00**	81.00	fuf	**93.00**	52.00	kqs	**95.00**	73.00	nym	**99.00**	54.00	toi	**98.00**	80.00
amh	**97.00**	36.00	fuv	**94.00**	61.00	ktu	**93.00**	19.00	nyn	**92.00**	**92.00**	tsn	**76.00**	33.00
bam	**70.00**	30.00	gaa	95.00	**97.00**	lia	97.00	**100.00**	nzi	97.00	**98.00**	tso	**99.00**	94.00
bba	**100.00**	83.00	gaz	94.00	**96.00**	lin	**99.00**	98.00	pcm	**96.00**	82.00	twi	**100.00**	87.00
bci	**98.00**	92.00	gjn	98.00	**99.00**	lot	**99.00**	93.00	pov	**93.00**	82.00	umb	**99.00**	76.00
bem	**94.00**	90.00	gkp	63.00	**69.00**	loz	**95.00**	92.00	run	**91.00**	68.00	vai	**100.00**	**100.00**
bfa	**99.00**	91.00	hau	**88.00**	77.00	lua	**99.00**	87.00	sag	**100.00**	30.00	ven	**95.00**	85.00
bin	**99.00**	97.00	ibb	**98.00**	84.00	lue	**95.00**	68.00	shk	**100.00**	93.00	vmw	**97.00**	95.00
bum	**97.00**	72.00	ibo	**97.00**	88.00	lug	**87.00**	64.00	sna	**97.00**	91.00	wol	**81.00**	21.00
cjk	**96.00**	56.00	kbp	**100.00**	98.00	lun	**97.00**	86.00	som	**95.00**	89.00	xho	**67.00**	30.00
crs	**96.00**	83.00	kde	**95.00**	60.00	men	98.00	**99.00**	sot	88.00	**93.00**	xsm	**99.00**	53.00
dag	**100.00**	**100.00**	kdh	**99.00**	95.00	mfq	**95.00**	88.00	ssw	**86.00**	68.00	yor	**98.00**	66.00
dga	**100.00**	78.00	kea	**96.07**	0.00	mos	**97.00**	90.00	suk	**99.00**	34.00	zdj	**96.00**	63.00
dip	**93.00**	86.00	kin	**89.00**	47.00	nba	**99.00**	61.00	sus	**99.00**	96.00	zul	**50.00**	40.00
dyu	**96.00**	00.00	kmb	**94.00**	71.00	nbl	**74.00**	47.00	swh	**77.00**	70.00			
ewe	**97.00**	**97.00**	kng	**98.00**	58.00	ndo	**96.00**	76.00	tem	**99.00**	88.00			
AfroLID Average F_1-score: **91.63**												**Franc Average F_1-score**: 74.81		

Table 6: F_1-scores on our Test dataset for languages in AfroLID and Franc for 88 languages.

ever, found errors in Amharic, Sebat Bet Gurage, and Xamtanga (which use Ethiopic scripts) as well as Fulfude Adamawa, and Fulfude Caka (which use Arabic scripts). We find that languages using Ethiopic scripts are often confused with those using Ethiopic scripts (except for 2% of the time when Amharic is labelled as Wolof). We categorize this example under "others" in Figure 5 and B.1. On the other hand, Fulfude languages are wrongly labelled as other dialects of Fulfude that use Latin scripts. We visualize further details of the errors in Figure B.1 (in Appendix) and 5 for our Dev and Test sets.

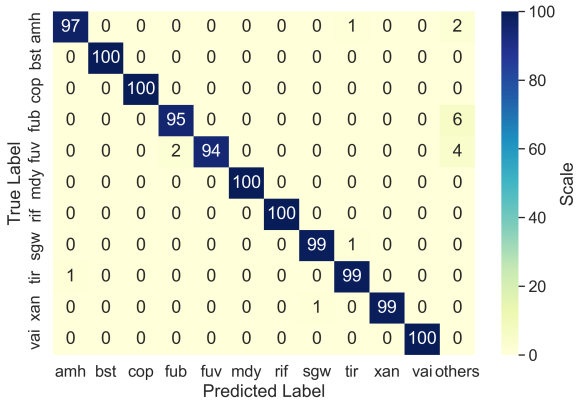

Figure 5: Errors on the different script in AfroLID Test set. We use ISO-3 codes to represent the languages. "Others" refers to languages AfroLID identifies as outside the list of languages selected for analysis.

Creole Languages. We investigate performance of AfroLID on Creole languages. Creole languages are vernacular languages that emerged as a result of trade interactions between speakers of mutually unintelligible languages (Lent et al., 2022). A Creole language therefore shares lexical items and grammatical structures with one or more different, unrelated languages. As a result, Creole languages appear to be *code-mixed*. AfroLID is trained on nine Creole languages: Krio, Nigerian

Pidgin, Cameroonian Pidgin, Seychelles Creole, Mauritian Creole, Kituba, Sango, Kabuverdianu, and Guinea-Bissau Creole. Krio, Cameroonian Pidgin, and Nigerian Pidgin are English based. Seychelles Creole and Mauritian Creole are French based. Kituba is Kongo based and Sango is Ngbadi based. Kabuverdianu and Guinea-Bissau Creole are Portuguese based. Evaluating AfroLID on Creoles thus demonstrates the robustness of our model, since (as mentioned above) Creoles can be viewed as a type of *code-mixed* language. We show performance of AfroLID on the nine Creole languages in Figure B.2 (in Appendix) and 6 for Dev and Test sets respectively.

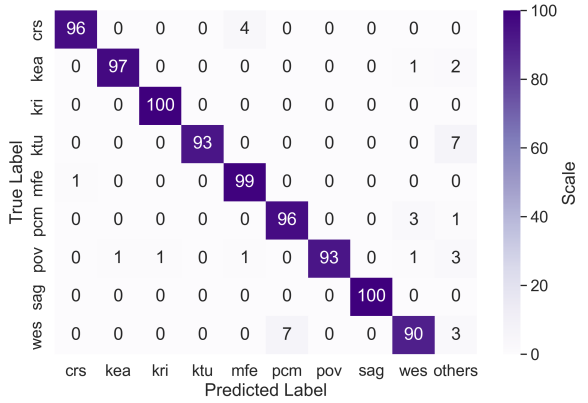

Figure 6: Errors on the different Creoles in AfroLID. We use ISO-3 codes to represent the languages. "Others" refers to languages AfroLID identifies as outside the list of languages selected for analysis.

We find that Guinea-Bissau Creole (pov), which is Portuguese based, is wrongly labelled as Kabuverdianu (kea) another Portuguese based Creole 1% of the time. Cameroonian pidgin (wes) is also wrongly labelled as Nigerian pidgin (pcm) 7% of the time. Since both Cameroonian and Nigerian Pidgin are English based, we assume lexical and/or grammatical similarities are responsible for these errors. It is also interesting to find cases where the wrong labels are languages spoken in the same geographical regions as the Creoles. For example, Kituba is wrongly labelled as Yombe, and both languages are spoken in Congo. Mauritian Creole (mfe), which is French based, is also wrongly labelled as Seychelles Creole (crs, another French based Creole) and two Indigenous languages spoken in Francophone Africa Ngiemboon, and Masana. We now further investigate the role of geographical proximity in our results.

Effect of Geographic Proximity. We evaluate performance of AfroLID on languages that share a large number of lexical items, or those that are spoken within the same country. In this analysis, we focus on 10 South African languages: Afrikaans (afr), Ndebele (nbl), Sepedi (nso), Sotho (sot), Swati (ssw), Tswana (tsn), Tsonga (tso), Tsivenda (ven), Xhosa (xho), and Zulu (zul). We select South Africa because most South Africans are multi-lingual, and it is not uncommon to find code-mixing using a combination of Indigenous languages within the same text (Finlayson and Slabbert, 1997; Mabule, 2015). Figures B.3 (in Appendix) and 7 show the types of errors AfroLID makes in identifying these languages on our Dev and Test datasets respectively. We find that about $\sim$ 70% of the errors are with other South African languages. Another 16% are with dialects from neighbouring countries including Tswa, a dialect of Tsonga, Ndebele (Zimbabwe) similar to Zulu, and Ronga, a dialect of Tsonga.[11] We now provide a number of case studies we carry out to further probe AfroLID performance.

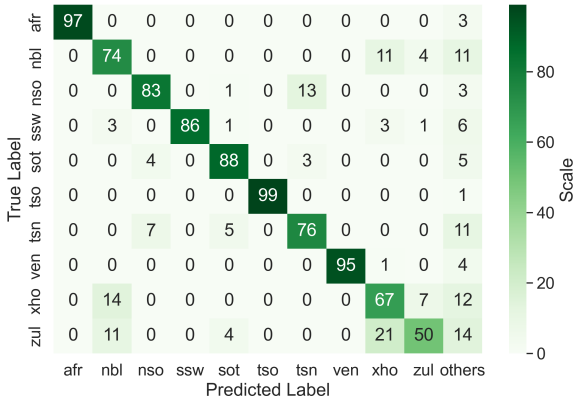

Figure 7: Errors on Indigenous South African languages in AfroLID Test data. "Others" refers to languages AfroLID identifies as outside the list of languages selected for analysis.

6 Diagnostic Case Studies

Although AfroLID is not trained on Twitter data, we evaluate its performance on Twitter to investigate the robustness of our models in out of domain scenarios. Namely, we carry out two diagnostic case studies using Twitter data. In the first study, which we refer to as Twitter in the wild, we use unannotated Tweets crawled from the web. In the the second, we use annotated tweets. We now turn to the details of these studies.

[11] A total of 14% of the errors are for other languages not related to South African languages.

Tool	Covered/All	Training Data	Methodology
Langid.py	7/97	GDoc, SDoc, News, ENC, IC	Naive Bayes, n-gram
Langdetect	3/49	Wikipedia	Naive Bayes, char n-gram
CLD2	4/80	Unknown	Naïve Bayes
CLD3	13/107	Unknown	Neural network, char n-gram
Equilid	1/70	Several GDoc, SDoc, RDoc, News, ENC, IC, Twitter	Neural seq2seq
Fasttext	5/176	Wiki, Tatoeba, Settimes	Classifier+hierarch. softmax, n-grams
Franc	88/403	UDHR	N-grams
AfroLID	517/517	Several GDoc, SDoc, RDoc, News, ENC, IC	Transformer

Table 7: AfroLID in comparison. **Covered/All:** # of African lgs compared with covered lgs, **GDoc:** Gov docs, **SDoc:** Software docs, **RDoc:**Religious docs, **News:** Newswire, **ENC:** online encyclopedia, **IC:** Internet crawl.

6.1 Case Study I: AfroLID in the Wild

In order to evaluate the utility of AfroLID in a real-world scenario, we collect 700M tweets from Africa. For this, we use Twitter streaming API from $2021-2022$ with four geographical bounding boxes (central, eastern, western, and southern of Africa). We extract a random sample of 1M tweets from this larger Twitter dataset for our analysis. As is known, Twitter currently automatically labels a total of 65 languages. Only one of these languages, i.e., Amharic, is an African language in our 517 languages. In the 1M sample, 110 tweets were tagged as "Amharic" and $6,940$ as "undefined" by Twitter. We run our model on the "undefined" data. In all, the $6,940$ tweets were identified as belonging to 242 African languages by AfroLID. Since the Tweets we used were unannotated, we are not able to determine the number of tweets wrongly classified by AfroLID for each language. For this reason, we only evaluate a subset of the predicted languages: we ask native speakers of three languages (Yorùbá, Hausa, and Nigerian Pidgin) to help identify each tweet that was classified by AfroLID as belonging to their language. We provide details of this annotation study and examples of annotated samples in Table D.1 (Appendix D). We find that AfroLID is able to correctly identify Yorùbá both with and without diacritics and code-mixed examples. A total of 16 tweets are classified as Yorùbá by AfroLID, of which 7 are correct (43.75%), 2 are mixed with English, and 7 are wrongly labelled. Of the wrongly labelled tweets, one is identified as Nigerian Pidgin, while the others are unknown languages. For Nigerian Pidgin, of the 28 tweets predicted, 2 are correct (12.50%), 1 is mixed with an unknown language, and the others are wrongly classified. We find that in most cases, tweets classified as Nigerian pidgin are code-mixed with English and another Indigenous language. This gives us indication that AfroLID identifies Nigerian Pidgin as an English-based Creole. Finally, a total of 333 tweets are classified as Hausa. Of these, 105 examples are correct (37.50%), 18 are mixed, while the others are wrongly labeled.

6.2 Case Study II: AfroLID on AfriSenti

We also test performance of AfroLID on the recently released AfriSenti Twitter dataset of African languages. AfriSenti (Muhammad et al., 2022; Yimam et al., 2020) contains $\sim 56,000$ tweets annotated for sentiment in Amharic, Hausa, Igbo, Nigerian Pidgin, Swahili, and Yorùbá. We run AfroLID and Franc tool on AfriSenti. As Figure 8 shows, AfroLID outperforms Franc on all languages except Nigerian Pidgin. We assume this is because Franc supports English and may have learnt some lexical / grammatical information from English to aid the identification of Nigerian Pidgin (although AfroLID outperforms Franc on Nigerian Pidgin on our Dev and Test as shown in Table 5 and 6.

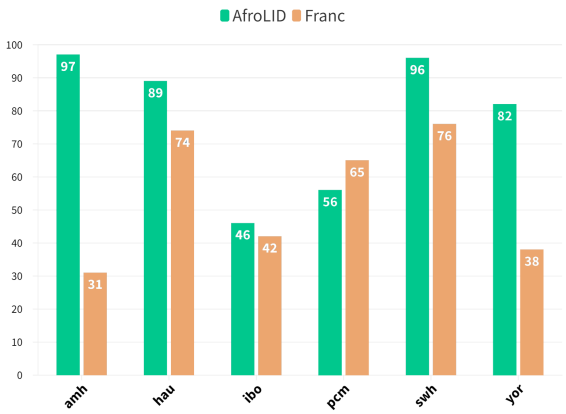

Figure 8: Performance of AfroLID and Franc on Afri-senti using F_1-score.

7 Related Work

LID tools are often used to select data to pre-train language models (Buck et al., 2014a) and, more generally, develop multilingual corpora (Buck et al., 2014b; Dunn, 2020; Scannell, 2007; Ortiz Suárez et al., 2019). For many languages, including African languages, LID tools are either not available or perform poorly (Kreutzer et al., 2021; Caswell et al., 2020). A few works, however, have already focused on African language identification. For example, Asubiaro et al. (2018) cover Yorùbá, Hausa, and Igbo. Similarly, Duvenhage et al. (2017b); Dube and Suleman (2019) treat 10 Indigenous South African official languages. In addition, a handful of other African languages are covered in LID tools such as CLD2 (McCandless, 2010), CLD3 (Salcianu et al., 2018), Equilid (Jurgens et al., 2017), FastText, Franc, LangDetect (Shuyo, 2010) and Langid.py (Lui and Baldwin, 2012) and works such as Abdul-Mageed et al. (2020, 2021) and Nagoudi et al. (2022). We provide an extended literature review of language identification, related tools, as well as data and methods employed in Appendix C. We also provide a comparison between available LID tools in terms of training data, methodology, and number of covered African languages in Table 7. To the best of our knowledge, AfroLID is the first publicly available LID tool covering a large number of African languages and varieties (n=517).

8 Conclusion

We introduced our novel African language identification tool, AfroLID. To the best of our knowledge, AfroLID is the first publicly available tool that covers a large number of African languages and language varieties. AfroLID also has the advantages of wide geographical coverage (50 African countries) and linguistic diversity. We demonstrated the utility of AfroLID on non-Latin scripts, Creoles, and languages with close geographical proximity. We also empirically showed AfroLID's superiority to five available tools, including in performance in the wild as applied to the much-needed Twitter domain. In the future, we plan to extend AfroLID to cover the top 100 most popular languages of the world as well as code-switched texts.

9 Limitations

We can identify a number of limitations for our work, as follows:

- AfroLID does not cover high-resource, popular languages that are in wide use by large populations. This makes it insufficient as a stand-alone tool in real-world scenarios where many languages are used side-by-side. Extending AfroLID to more languages, however, should be straightforward since training data is available. Indeed, it is our plan to develop AfroLID in this direction in the future.

- AfroLID recognizes only Indigenous African languages in monolingual settings. This limits our tool's utility in code-mixed scenarios, (although Creoles are like code-mixed languages). This is undesirable especially because many African languages are commonly code-mixed with foreign languages due to historical reasons (Adebara and Abdul-Mageed, 2022). Again, to improve accuracy in the future, it would be beneficial to add foreign languages support in code-mixed settings such as with English, French, and Portuguese.

- Although we strive to test AfroLID in real-world scenarios, we were not able to identify native speakers except from a small number of languages. In the future, we plan to work more with the community to enable wider analyses of our predictions.

10 Ethical Considerations

Although LID tools are useful for a wide range of applications, they can also be misused. We release AfroLID hoping that it will be beneficial to wide audiences such as to native speakers in need of better services like health and education. Our tool is also developed using publicly available datasets that may carry biases. Although we strive to perform analyses and diagnostic case studies to probe performance of our models, our investigations are by no means comprehensive nor guarantee absence of bias in the data. In particular, we do not have access to native speakers of most of the languages covered in AfroLID. This hinders our ability to investigate samples from each (or at least the majority) of the languages. We hope that future users of the tool will be able to make further investigations to uncover AfroLID's utility in wide real-world situations.

Acknowledgements

We gratefully acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[12] UBC ARC-Sockeye,[13], Advanced Micro Devices, Inc. (AMD), and Google. Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of CRC, NSERC, SSHRC, CFI, CC, AMD, Google, or UBC ARC-Sockeye.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

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

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8, Valencia, Spain. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David 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.

[12] https://alliancecan.ca

[13] https://arc.ubc.ca/ubc-arc-sockeye

Toluwase Asubiaro, Tunde Adegbola, Robert Mercer, and Isola Ajiferuke. 2018. A word-level language identification strategy for resource-scarce languages. *Proceedings of the Association for Information Science and Technology*, 55(1):19–28.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2017. A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Ralf D. Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In *Text, Speech, and Dialogue*, pages 475–483, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014a. N-gram counts and language models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014b. N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

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.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

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.

Denis Creissels, Gerrit J Dimmendaal, Zygmunt Frajzyngier, and Christa König. 2008. Africa as a morphosyntactic area. *A linguistic geography of Africa*, 86150.

N. Dongen. 2017. Analysis and prediction of Dutch-English code-switching in Dutch social media messages.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Meluleki Dube and Hussein Suleman. 2019. Language identification for South African Bantu languages using rank order statistics. In *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings*, page 283–289, Berlin, Heidelberg. Springer-Verlag.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Language Resources and Evaluation*, 54(4).

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017a. 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.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017b. 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.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). 2021. Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Rosalie Finlayson and Sarah Slabbert. 1997. "We just mix": code switching in a South African township. 1997(125):65–98.

Spandana Gella, Kalika Bali, and Monojit Choudhury. 2014. "ye word kis lang ka hai bhai?" testing the limits of word level language identification. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 368–377, Goa, India. NLP Association of India.

Spandana Gella, Jatin Sharma, and Kalika Bali. 2013. Query word labeling and back transliteration for Indian languages: Shared task system description. In *Working Notes - Forum for Information Retrieval Evaluation (FIRE) 2013 Shared Task*. Best Performing System at FIRE-2013.

Helena Gomez, Ilia Markov, Jorge Baptista, Grigori Sidorov, and David Pinto. 2017. Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 137–145, Valencia, Spain. Association for Computational Linguistics.

Lena Grothe, Ernesto William De Luca, and Andreas Nürnberger. 2008. A comparative study on language identification methods. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Larry M Hyman. 2003. African languages and phonological theory. *Glot International*, 7(6):153–163.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online).

International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017a. Evaluation of language identification methods using 285 languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 183–191, Gothenburg, Sweden. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017b. Evaluating heli with non-linear mappings. pages 102–108.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *J. Artif. Int. Res.*, 65(1):675–682.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Chris van der Lee and Antal van den Bosch. 2017. Exploring lexical and syntactic features for language variety identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 190–199, Valencia, Spain. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022. Ancestor-to-creole transfer is not a walk in the park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

D R Mabule. 2015. What is this? is it code switching, code mixing or language alternating? *Journal of Educational and Social Research*, 5(1).

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Matej Martinc, Iza Skrjanec, Katja Zupan, and Senja Pollak. 2017. Pan 2017: Author profiling - gender and language variety prediction. In *CLEF*.

Michael McCandless. 2010. Accuracy and performance of google's compact language detector. *Blog post*.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Muntsa Padró and Lluís Padró. 2004. Comparing methods for language identification. *Proces. del Leng. Natural*, 33.

Iria del Río Gayo, Marcos Zampieri, and Shervin Malmasi. 2018. A Portuguese native language identification dataset. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, et al. 2018. Compact language detector v3.

Younes Samih. 2017. *Dialectal Arabic processing Using Deep Learning*. Ph.D. thesis.

Kevin P. Scannell. 2007. The Crúbadán project: Corpus building for under-resourced languages.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 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, Online. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language detection library for java.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

S. Thara and Prabaharan Poornachandran. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837–118850.

Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. 2021. Improved language identification through cross-lingual self-supervised learning.

Erik Tromp. 2011. Multilingual sentiment analysis on social media.

John Vogel and David Tresner-Kirsch. 2012. Robust language identification in short, noisy texts: Improvements to liga. In *Proceedings of the 3rd international Workshop on Mining Ubiquitous and Social Environments*, pages 1–9.

John C. Wells. 2000. Orthographic diacritics and multilingual computing. *Language Problems and Language Planning*, 24:249–272.

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

Yonghong Yan and E. Barnard. 1995. An approach to automatic language identification based on language-dependent phone recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3511–3514 vol.5.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

Marc A Zissman and Kay M Berkling. 2001. Automatic language identification. *Speech Communication*, 35(1):115–124. MIST.

Appendices

A Results of AfroLID on Dev Set

We report results from comparing AfroLID with CLD2, CLD3, Langid.py, LangDetect, and Franc on our Dev set in Table A.1.

Lang.	CLD2	CLD3	Langid.py	LangDetect	Franc	AfroLID
afr	94.11	88.23	70.58	92.15	84.31	**94.11**
amh	-	98.03	**100.00**	-	25.49	98.03
hau	-	86.27	-	-	82.35	**94.11**
ibo	-	92.15	-	-	90.19	**94.11**
kin	**88.23**	-	56.86	-	52.94	80.39
lug	74.50	-	-	-	52.94	**86.27**
mlg	-	**98.03**	92.15	-	-	96.07
nya	-	**96.07**		-	82.35	**96.07**
sna	-	86.27	-	-	80.39	**96.07**
som	-	96.07	-	-	96.07	**98.03**
sot	-	**90.19**	-	-	**90.19**	76.47
swa	92.15	90.19	86.27	**96.07**	-	92.15
swc	90.19	96.07	**98.03**	**98.03**	-	74.50
swh	88.23	**96.07**	90.19	90.19	72.54	74.50
xho	-	90.19	**94.11**	-	64.70	82.35
yor	-	50.82	-	-	39.21	**100.00**
zul	-	**86.27**		-	37.25	58.82

Table A.1: A comparison of results on AfroLID with CLD2, CLD3, Langid.py, LangDetect, and Franc using F_1-score on the Dev set. A dash ("−") indicates that the tool does not support the language.

B Analysis of AfroLID

We perform the experiments on non-Latin scripts, Creoles, and languages in close geographical proximity on the Dev set, as in Subsection 5. We show the results on the performance of AfroLID on non-Latin scripts in Table B.1, Creole languages in Table B.2 and geographical proximity in Table B.3 respectively.

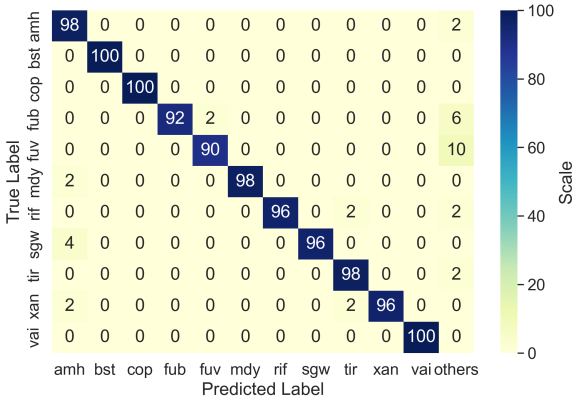

Figure B.1: Errors on the different script in AfroLID Dev set. We use ISO-3 codes to represent the languages. "Others' refers to languages AfroLID identifies as outside the list of languages selected for analysis.

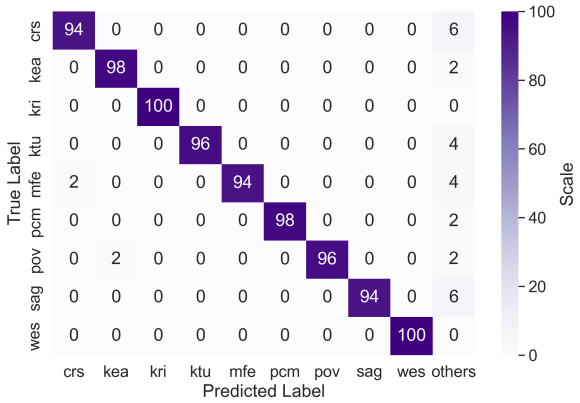

Figure B.2: Errors on the different Creoles in AfroLID. We use ISO-3 codes to represent the languages. "Others" refers to languages AfroLID identifies as outside the list of languages selected for analysis.

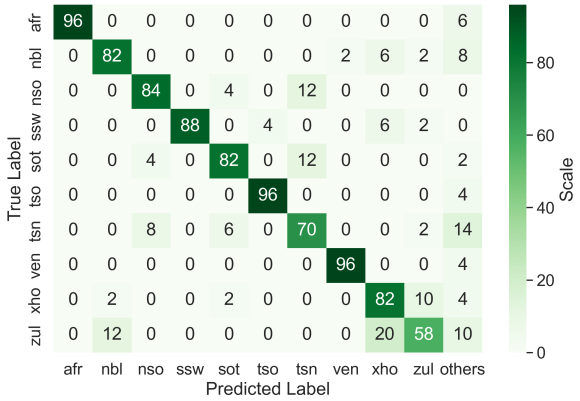

Figure B.3: Errors on Indigenous South African languages in AfroLID Dev data. "Others' refers to languages AfroLID identifies as outside the list of languages selected for analysis.

C Extended Literature Review

C.1 Datasets

Datasets for LID are often created using various genre of data for one or more languages. For multilingual LID, which is the focus of our work, documents are gathered from web pages containing multiple languages. Web pages for multilingual organizations are also often desirable because the same text is translated into various languages. Most datasets for multilingual LID cover European languages and many other high resource languages, making AfroLID dataset a significant contribution to AfricaNLP. To the best of our knowledge, AfroLID dataset is the first publicly available dataset for multilingual language identification for African languages. We provide details of some other publicly available corpora for LID.

DSL Corpus Collection (Tan et al., 2014; Malmasi et al., 2016; Zampieri et al., 2015, 2014) is a multilingual collection of short excerpts of jour-

	COPLE2	LEIRIA	PEAPL2	TOTAL
Sents	1,058	330	480	1,868
Tokens	201,921	57,358	121,138	380,417
Types	9,373	4,504	6,808	20,685
TTR	0.05	0.08	0.06	0.05

Table C.1: Distribution of the dataset: Number of texts, tokens, types, and type/token ratio (TTER) per source corpus.

nalistic texts. It has been used as the main data set for the DSL shared tasks organized within the scope of the workshop on NLP for Similar languages, Varieties and Dialects (VarDial). It covers 22 languages.

NLI-PT (del Río Gayo et al., 2018) is a dataset collected from three different learner corpora of Portuguese including COPLE2; Leiria corpus, and PEAPL. The three corpora contain written productions from learners of Portuguese with different proficiency levels and native languages. The dataset included all the data in COPLE2 and sections of PEAPL2 and Leiria corpus with details of the dataset in Table C.1. Therefore, the dataset include texts corresponding to the following 15 languages: Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Romanian, Russian, Swedish, Spanish, and Tetum.

Wanca 2017 Web Corpora (Jauhiainen et al., 2020) is made up of re-crawls performed by the SUKI project. The target of the re-crawl was to download and check the availability of the then current version of the Wanca service of about 106,000 pages. This list of 106,000 http addresses was the result of several earlier web-crawls, in which they had identified the language in a total of 3,753,672,009 pages.

EUROGOV, TCL, and WIKIPEDIA (Baldwin and Lui, 2010) consist of documents with a single encoding across 10 European languages; shorter documents across different encodings for 60 languages, and wikipedia web crawls for 67 languages respectively. These collection cover different genres with Eurogov collected from government documents, TCL from online news sources and Wikipedia dumps.

The UMass Global English on Twitter Dataset (Blodgett et al., 2017) contains 10,502 tweets, randomly sampled from all publicly available geotagged Twitter messages, annotated for being in English, non-English, or having code switching, language ambiguity, or having been automatically generated. It includes messages sent from 130 different countries.

C.2 Features

Different features can be used for training a LID system including:

- Bytes and Encoding: Some encodings use a fixed number of bytes e.g ASCII while some others use variable length encoding. Some languages also use specific encodings (GuoBiao 18030 or Big5 for chinese) while the same encoding can be used for different languages (e.g UTF-8).

- Characters: Non-alphabetic, alphabets, capitalization, the number of characters in words and word combinations, the number of characters in words and word combinations have been used as features. Non-alphabetic characters has been used to detect languages like Arabic, emojis, and other languages that use non-alphabetic characters (Samih, 2017; Bestgen, 2017; Dongen, 2017). Alphabets can also be used to exclude languages where a unique character is absent in the test document.

- Character combination: co-occurrences of some characters can be used to detect some languages. Linguistically, some languages abhor certain combination of characters which some other languages allow. For example some Niger-Congo languages abhor vowel hiatus and every consonant must be followed by a vowel. This feature has been found useful for developing LID systems (van der Lee and van den Bosch, 2017; Dongen, 2017; Martinc et al., 2017).

- Morphemes, Syllables and Chunks: different morphological features including prefixes, suffixes, and character n-grams (Gomez et al., 2017). Syllables, chunks, and chunks of syllables / ngrams have also been used for LID. This also has linguistic significance in that the prefix, suffixes and morphological information embedded in a language can provide information about the etymology of a language.

- Words: The position of words (Adouane and Dobnik, 2017), the string edit distance and n-gram overlap between the word to be identified and words in dictionaries, dictionary of unique words in a language, basic dictionary

of a language, most common words, word clusters among others are some discriminating features used for LID.

- Combination of words: Here, length of words, the ratio to the total number of words of: once-occurring words, twice-occurring words, short words, long words, function words, adjectives and adverbs, personal pronouns, and question words are some features used here (van der Lee and van den Bosch, 2017). This feature is linguistically significant since the ratio of certain categories of words can be useful for identifying some languages.

- Syntax and Part of speech (POS) tags: Syntactic features can be used to identify languages. Identifying an adjective before a noun for instance may be a good indication for some languages and even the tags available can be a useful feature. Syntactic parsers together with dictionaries and morpheme lexicons, n-grams composed of POS tags and function words have all been used as features (Adouane and Dobnik, 2017) for LID.

- Languages identified for surrounding words in word-level LID: The language of surrounding words can also be a useful feature since there may be a higher likelihood of having some languages used together. This is especially true in the case of codeswitching where some languages are more likely to be used together than some others (Dongen, 2017).

- Feature smoothing: Feature smoothing is required in order to handle the cases where not all features in a test document have been attested in the training corpora. Feature smoothing is used in low resource scenarios and when the frequency of some features are high. Different types of feature smoothing is possible. Some of them are additive smoothing where an extra number of occurrences is added to every possible feature in the language model (Jauhiainen et al., 2019).

C.3 Methods

Algorithms for LID work by first using one or more features before using a classification algorithm to determine the appropriate language for a text(Grothe et al., 2008; Jauhiainen et al., 2019).

Hidden Markov Models (HMM) Hidden Markov Models (HMM) are commonly used in spoken language identification (Zissman and Berkling, 2001; Yan and Barnard, 1995) as well as for written language (Guzman et al., 2016). Language models are first trained for each language that the system must know about using a text corpora, and stored for later comparison with unidentified text. In these models the parameters of the HMM are the transition probability and the initial probability. Probabilities are calculated using the relative frequency of each transition or initial state of the training data. After training, the system calculates the sequence probability using each language model that has been trained (Padró and Padró, 2004).

N-Gram-Based Text Categorization This method introduced by (Cavnar and Trenkle, 1994; Grothe et al., 2008) is based on comparing unique n-gram frequency profiles. These frequencies are sorted in decreasing order for all unique n-grams. N-gram profiles are created for each language to be trained with $n = 1$ to 5. To classify a piece of text, the n-gram frequency for that text is built and compared to the n-gram profiles calculated during the training phase. This is done by computing the distance between the n-gram profiles of the text and that for each language model. The computation also penalizes the total score of the language for each missing n-gram. The language with the lowest score is selected as the identified language (Jauhiainen et al., 2017a; Padró and Padró, 2004).

LIGA This uses a graph-based n-gram approach called LIGA which was originally used for sentiment analysis (Tromp, 2011) and adopted for LID (Vogel and Tresner-Kirsch, 2012). The language models use the relative frequencies of character trigrams and those of 4-grams. To identify the language in a text, the relative frequency of each trigram and 4-gram found in a language model is added to the score of the language. The language with the highest score is selected as the language of the text.

HELI Method The HeLI method (Jauhiainen et al., 2017b) uses character n-grams based language models for each language. The n-gram values are hyperparameters from one to a specific maximum number N_{max}. The model then selects one language model when classifying the language of a text. The selection is based on the most applicable model to the specified text. The model then gradually backs off to a lower order n-gram if the n-

gram with the N_{max} is not applied until an n-gram can be applied. The validation set is used during evaluation to determine the best values for N_{max}, the maximum number of features to be included in the language models, and the penalty for languages without the selected feature. The penalty functions like a smoothing parameter by transferring some of the probability mass to unseen features in the language model (Jauhiainen et al., 2017a).

Whatlang program This uses language models built with n-grams of variable byte lengths between $3 - 12$ (Brown, 2013). The K most frequent n-grams and their relative frequencies are then extracted and calculated for each language. Once the first model is generated, substrings of larger n-grams are filtered out if the larger n-gram has a frequency not less than 62% of the frequency of the shorter n-grams. The model weights are computed for each language such that shorter n-grams with the same relative frequency have lower weights than those with larger n-grams. This is because larger n-grams are more informative but less common.

C.4 Language Identification Tools

Several tools have been developed for multilingual LID. We provide details of different tools which has representation for African languages including CLD2 (McCandless, 2010), CLD3 (Salcianu et al., 2018) EquiLID (Jurgens et al., 2017), fastText (Joulin et al., 2017), Franc, Langid.py (Lui and Baldwin, 2012), and LangDetect (Shuyo, 2010).

C.4.1 CLD2[14]

CLD2 (McCandless, 2010) covers 83 languages and trained on web pages text, using one of three different token algorithms. CLD2 probabilistically detects over 86 languages including Afrikaans and Swahili. Unicode UTF-8 text, either plain text or HTML/XML. It requires that legacy encodings be converted to valid UTF-8. For mixed-language input, CLD2 returns the top three languages found and their approximate percentages of the total text bytes (e.g. 80% English and 20% French out of 1000 bytes of text means about 800 bytes of English and 200 bytes of French). Optionally, it also returns a vector of text spans with each language identified.

[14] https://github.com/CLD2Owners/cld2

C.4.2 CLD3

CLD3 (Salcianu et al., 2018)[15], the latest updated version of *CLD2* (2020) covers 106 languages including Afrikaans, Amharic, Hausa, Malagasy, Shoma, Somali, Swahili, Xhosa, Yoruba, and Zulu. CLD3 uses a neural network model for language identification. It contains the inference code and a trained model.

C.4.3 EquiLID

EquiLID (Jurgens et al., 2017)[16] is a character based DNN $encoder - decoder$ model (Cho et al., 2014; Sutskever et al., 2014) with an attention mechanism (Bahdanau et al., 2015). Equilid is a general purpose language identification library and command line utility built to identify a broad coverage of languages, recognize language in social media, with a particular emphasis on short text, recognizing dialectic speech from a language's speakers, identify code-switched text in any language pairing at least at the phrase level, provide whole message and per-word. EquiLID covers 70 languages including Amharic.

C.4.4 FastText

FastText (Joulin et al., 2016) supports 176 languages including 5 African languages. The model uses a classifier with hierachical softmax with n-grams.

C.4.5 Franc

Franc supports 403 languages including 88 African languages. It is built using Universal Declaration of Human Rights UDHR documents translated into multiple languages. Details of the model architecture is not available, however there is indication that n-grams are used in the model.

C.4.6 LangDetect

LangDetect (Shuyo, 2010) covers 49 languages including Afrikaans and Swahili. LangDetect uses a huge dictionary of inflections and compound words over a Naive Bayes model with character n-grams.

C.4.7 Langid.py

`Langid.py` (Lui and Baldwin, 2012) covers 97 languages including Afrikaans, Amharic, Malagasy, Kinyarwanda, Swahili, and Zulu. The model is trained over a naive Bayes classifier with a multinomial event model using a mixture of byte n-

[15] https://github.com/google/cld3

[16] https://github.com/davidjurgens/equilid

grams. `langid.py` was designed to be used off-the-shelf. It comes with an embedded model using training data drawn from 5 domains - government documents, software documentation, newswire, online encyclopedia, and an internet crawl, though no domain covers the full set of languages by itself, and some languages are present only in a single domain. Different aspects of `langid.py` are evaluated in different ways. For cross-lingual feature selection evaluation, each dataset is partitioned into two sets of equal sizes. The first partition is used for training a classifier while the second is used for evaluation. Since each dataset covers a different set of languages, there may be languages in the evaluation dataset that are not present in the training dataset (Lui and Baldwin, 2011). The `langid.py` module on the other hand is evaluated on different datasets and the accuracy is compared with those for CLD, Textcat, and LangDetect. The accuracy of Langid.py exceeded those from other tools on two twitter datasets (Lui and Baldwin, 2012). `Langid.py` can be used as a command line tool, python library, or web service tool.

LID Tool	African Languages
CLD2	afr, lug, kin, swa
CLD3	afr, amh, hau, ibo, mlg, nya, sna, som, sot, swa, xho, yor, zul
Langid.py	afr, amh, kin, mlg, swa, xho, zul
EquiLID	amh
LangDetect	afr, swh
FastText	afr, amh, mlg, som, swh, yor

Table C.2: African languages represented in different LID tools.

Other LID tools without representation of African languages include LDIG, and Microsoft LID-tool (Gella et al., 2013, 2014) which is a word level language identification tool for identifying code-mixed text of languages (like Hindi etc.) written in roman script and mixed with English.

D Twitter Analysis

For the Twitter in the wild analysis, we ask for annotations of *yes*, *no* or *mixed* on each tweet, where *yes* indicates agreement with the predicted label, *no* indicates disagreement, and *mixed* indicates that the tweet contains one or more other language than the predicted. We also ask for further annotations if the tweet is not in the predicted language, or is mixed with another/other language(s). In these cases, respondents are asked to identify the correct language (or mixed language[s]) if they know the language(s). We provide example annotation in the wild analysis in Table D.1 .

E Languages Covered in AfroLID

AfroLID supports 517 African languages and language varieties. We show a large map indicating the countries and languages represented in Figure E.1. Figure E.2 and E.3 show the number of languages covered in each country and the language family information for the languages. We also show the languages and language codes in Table E.1, E.2, and E.3.

aar	bez	cou	eza	ife	khy	lem	mfi	nga	rif	ssc	uth
abn	bfa	csk	fia	igb	kia	lik	mgc	ngb	rim	suk	vag
ada	bfd	daa	fip	ige	kik	lip	mgo	ngn	rub	sus	vif
adj	bfo	daf	flr	igl	kkj	lmd	mgq	nhr	run	taq	vun
afr	bib	dga	fon	ijn	klu	lmp	mkl	nhu	rwk	tcd	vut
agq	biv	dgi	gaa	ikk	kmb	lnl	mlr	nim	sag	tem	wbi
akp	bjv	dhm	gbo	ikw	knf	log	mnf	nin	sba	tex	wib
ann	bky	dib	gid	iqw	koq	lol	mnk	niq	sbd	tgw	wmw
anu	bmo	did	giz	iri	kqp	lom	mos	niy	sbp	thk	xed
anv	bmv	dik	gkp	iso3	kqs	loq	moz	nko	sef	thv	xpe
asg	bom	dip	gna	izr	krs	lot	mpg	nla	ses	tiv	xrb
atg	bov	dnj	gnd	izz	krw	loz	mqb	nnh	sev	tlj	xsm
avn	box	dow	gng	jgo	krx	lro	mua	nnw	sfw	tod	xtc
avu	bqc	dsh	gol	jib	ksb	luc	muh	nse	shi	tog	xuo
azo	bqj	dug	gqr	kam	ksf	lwo	muy	nso	shj	tsw	yam
bav	bsc	dyi	gso	kbn	ksp	maf	mwm	nus	shk	ttq	yao
bba	bss	ebr	gur	kbo	kss	mbu	mws	nyb	sig	ttr	yat
bbj	bud	ebu	guw	kbp	kub	mcp	myb	nyy	sil	tui	yba
bbk	bum	efi	gux	kcg	kuj	mcu	myk	nza	snf	tul	yor
bci	bus	ego	gvl	kde	kyq	mda	mzm	odu	snw	tum	zga
bcp	buy	eka	gya	kde	kzr	mdm	mzw	okr	sop	tvu	zne
bcy	bza	etu	hna	kdh	lam	meq	naq	oku	sor	udu	
bdh	bzw	etx	ibb	kdl	lap	mer	ncu	ozm	sot	umb	
bds	cko	ewe	ibo	ken	lee	mev	ndv	pkb	soy	urh	
bex	cme	ewo	idu	ker	lef	mfh	ndz	pko	spp	uth	

Table C.3: Language varieties that use diacritics in our training data.

ISO-3	Tweet	Representative?	No	Mixed
yor	Don't be on my TL supporting a rapist, a o ní s'oriburubuku o	Mixed		English
	USER Omo ilorin Nile Adeleke ti Binu	Yes		
	Oproblema opo openi ne	No	Unknown	
	USER On top Iron Konji na Bastard	No	Nigerian Pidgin	
ibo	USER Mana ima na ife any i na-ekwu bu eziokwu	Yes		
	USER Mo je ri e	No	Yorùbá	
	USER Hamna namna mzee	No	Unknown	
hau	USER Kaji dadinka brother ka huta	Mixed		English
	USER Su Umar danbarade	Yes		
	USER Good nkosazana Cathy	No	English + unknown	
	ovo ra mbuti USER Sesi Gladys mani	No	Unknown	
pcm	USER Gompieno o bone dust !	Mixed		Unknown
	USER Wey I travel from Ilesa to Ipetumodu	Yes		
	USER Ende zwotoralo ngoho ngoho	No	Unknown	
	Despacito! beyaudkrnkwudh despacito, daueiejrb despacitoo! goose bumps	No	English + unknown	

Table D.1: Some example annotations for the Twitter in the wild analysis. We show for each language the 4 possible annotations.

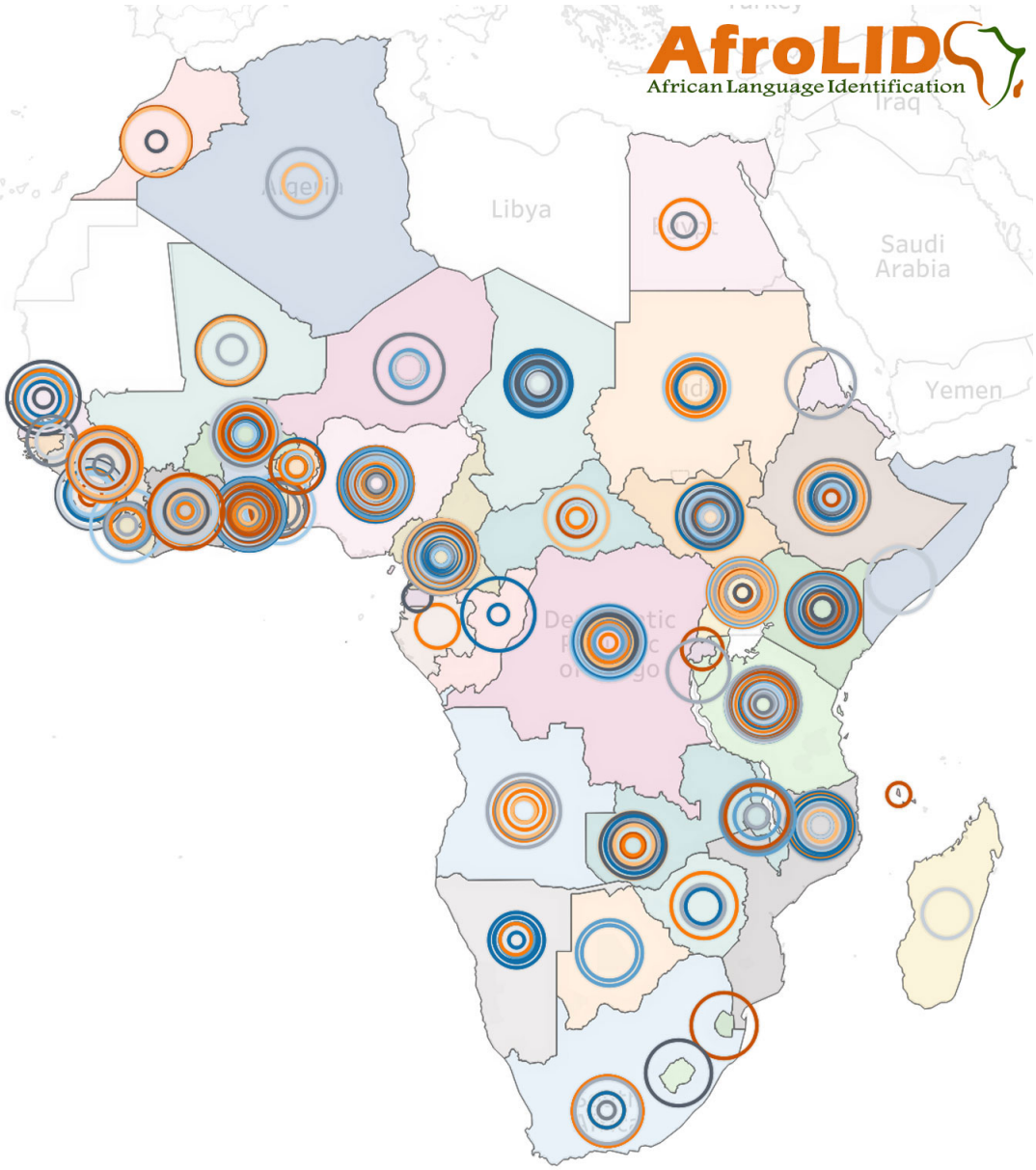

Figure E.1: All 50 African countries in our data, with our 517 languages/language varieties in colored circles overlayed within respective countries.

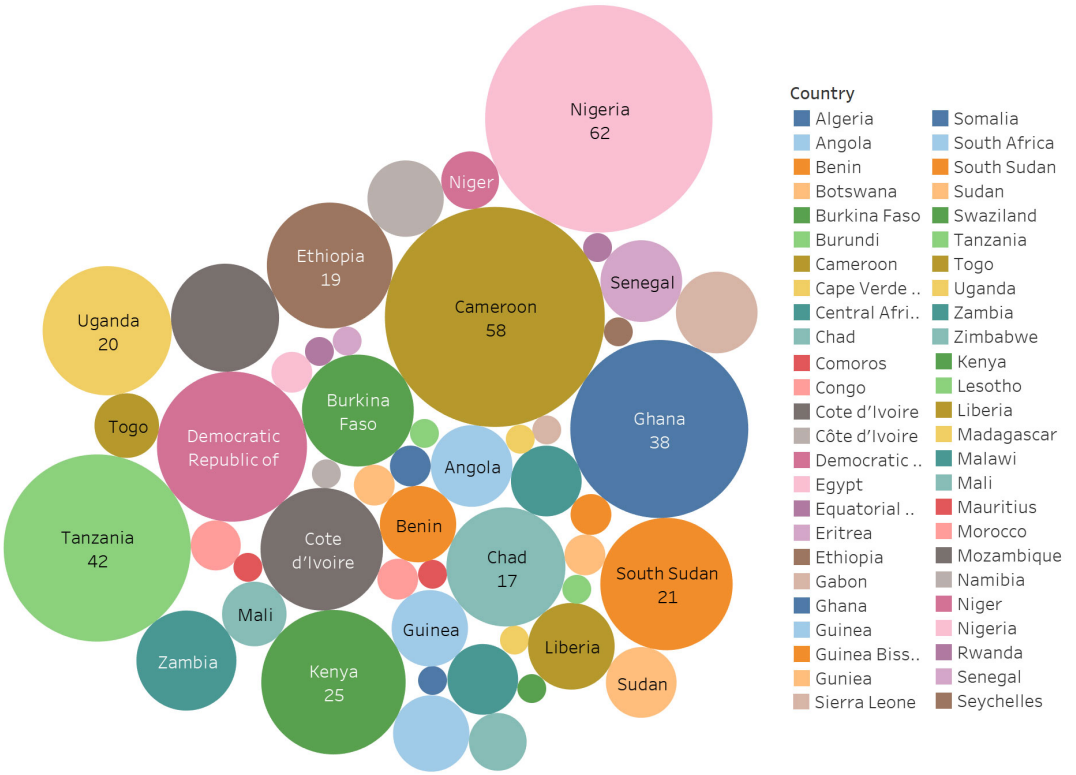

Figure E.2: AfroLID's Covered languages.

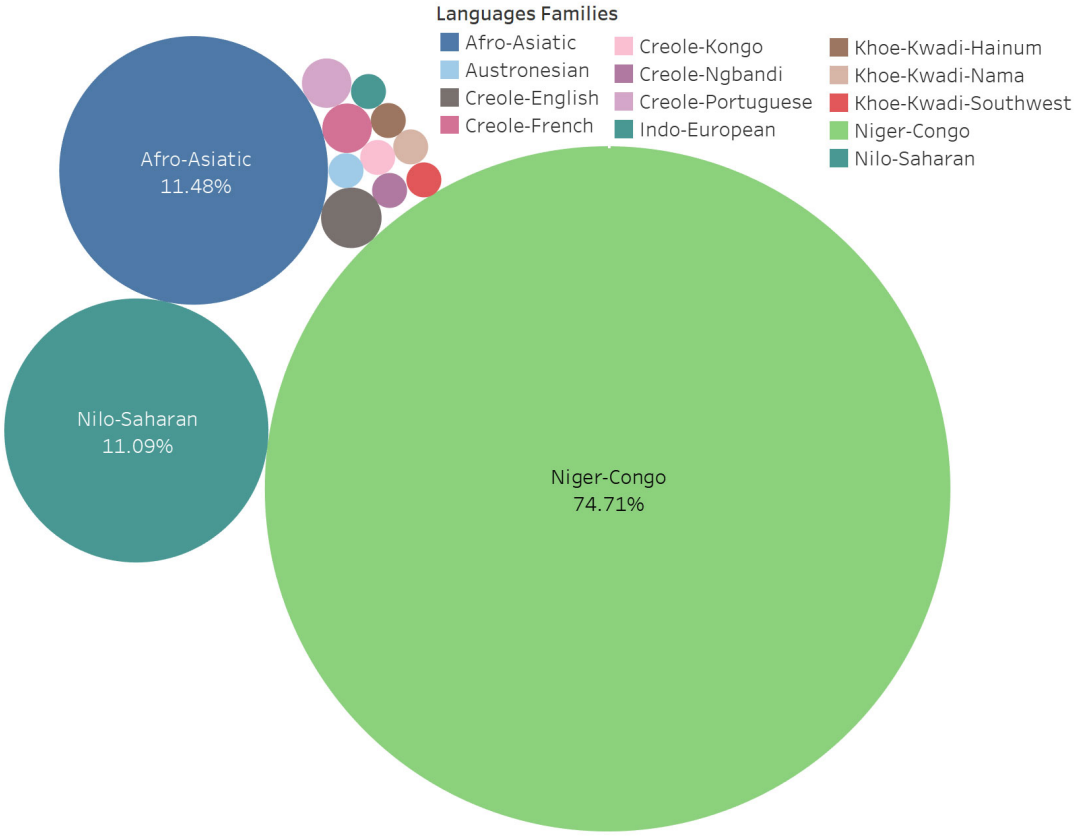

Figure E.3: Percentage of languages per family on training dataset.

ISO-3	Language	ISO-3	Language	ISO-3	Language	ISO-3	Language
aar	Afar / Qafar	bky	Bokyi	dow	Doyayo	gol	Gola
aba	Abe / Abbey	bmo	Bambalang	dsh	Daasanach	gqr	Gor
abn	Abua	bmv	Bum	dua	Douala	gso	Gbaya, Southwest
acd	Gikyode	bom	Berom	dug	Chiduruma	gud	Dida, Yocoboue
ach	Acholi	bov	Tuwuli	dwr	Dawro	gur	Farefare
ada	Dangme	box	Bwamu / Buamu	dyi	Sénoufo, Djimini	guw	Gun
adh	Jopadhola / Adhola	bqc	Boko	dyu	Jula	gux	Gourmanchema
adj	Adjukru / Adioukrou	bqj	Bandial	ebr	Ebrie	guz	Ekegusii
afr	Afrikaans	bsc	Oniyan	ebu	Kiembu / Embu	gvl	Gulay
agq	Aghem	bsp	Baga Sitemu	efi	Efik	gwr	Gwere
aha	Ahanta	bss	Akoose	ego	Eggon	gya	Gbaya, Northwest
ajg	Aja	bst	Basketo	eka	Ekajuk	hag	Hanga
akp	Siwu	bud	Ntcham	eko	Koti	har	Harari
alz	Alur	bum	Bulu	eto	Eton	hau	Hausa
amh	Amharic	bun	Sherbro	etu	Ejagham	hay	Haya
ann	Obolo	bus	Bokobaru	etx	Iten / Eten	hbb	Nya huba
anu	Anyuak / Anuak	buy	Bullom So	ewe	Ewe	heh	Hehe
anv	Denya	bwr	Bura Pabir	ewo	Ewondo	her	Herero
asa	Asu	bwu	Buli	fak	Fang	hgm	Haillom
asg	Cishingini	bxk	Bukusu	fat	Fante	hna	Mina
atg	Ivbie North-Okpela-Arhe	byf	Bete	ffm	Fulfulde, Maasina	ibb	Ibibio
ati	Attie	byv	Medumba	fia	Nobiin	ibo	Igbo
avn	Avatime	bza	Bandi	fip	Fipa	idu	Idoma
avu	Avokaya	bzw	Basa	flr	Fuliiru	igb	Ebira
azo	Awing	cce	Chopi	fon	Fon	ige	Igede
bam	Bambara	chw	Chuabo	fub	Fulfulde, Adamawa	igl	Igala
bav	Vengo	cjk	Chokwe	fue	Fulfulde, Borgu	ijn	Kalabari
bba	Baatonum	cko	Anufo	fuf	Pular	ikk	Ika
bbj	Ghomala	cme	Cerma	fuh	Fulfulde, Western Niger	ikw	Ikwere
bbk	Babanki	cop	Coptic	ful	Fulah	iqw	Ikwo
bci	Baoule	cou	Wamey	fuq	Fulfulde Central Eastern Niger	iri	Rigwe
bcn	Bali	crs	Seychelles Creole	fuv	Fulfude Nigeria	ish	Esan
bcw	Bana	csk	Jola Kasa	gaa	Ga	iso	Isoko
bcy	Bacama	cwe	Kwere	gax	Oromo, Borana-Arsi-Guji	iyx	yaka
bdh	Baka	daa	Dangaleat	gaz	Oromo, West Central	izr	Izere
bds	Burunge	dag	Dagbani	gbo	Grebo, Northern	izz	Izii
bem	Bemba / Chibemba	dav	Dawida / Taita	gbr	Gbagyi	jgo	Ngomba
beq	Beembe	dga	Dagaare	gde	Gude	jib	Jibu
ber	Berber	dgd	Dagaari Dioula	gid	Gidar	jit	Jita
bex	Jur Modo	dgi	Dagara, Northern	giz	South Giziga	jmc	Machame
bez	Bena	dhm	Dhimba	gjn	Gonja	kab	Kabyle
bfa	Bari	dib	Dinka, South Central	gkn	Gokana	kam	Kikamba
bfd	Bafut	did	Didinga	gkp	Kpelle, Guinea	kbn	Kare
bfo	Birifor, Malba	dig	Chidigo	gmv	Gamo	kbo	Keliko
bib	Bisa	dik	Dinka, Southwestern	gna	Kaansa	kbp	Kabiye
bim	Bimoba	dip	Dinka, Northeastern	gnd	Zulgo-gemzek	kby	Kanuri, Manga
bin	Edo	diu	Gciriku	gng	Ngangam	kcg	Tyap
biv	Birifor, Southern	dks	Dinka, Southeastern	gof	Goofa	kck	Kalanga
bjv	Bedjond	dnj	Dan	gog	Gogo	kdc	Kutu

Table E.1: AfroLID covered Languages - Part I.

ISO-3	Language	ISO-3	Language	ISO-3	Language	ISO-3	Language
kde	Makonde	laj	Lango	mfh	Matal	ngb	Ngbandi, Northern
kdh	Tem	lam	Lamba	mfi	Wandala	ngc	Ngombe
kdi	Kumam	lap	Laka	mfk	Mofu, North	ngl	Lomwe
kdj	Ng'akarimojong	lee	Lyélé	mfq	Moba	ngn	Bassa
kdl	Tsikimba	lef	Lelemi	mfz	Mabaan	ngo	Ngoni
kdn	Kunda	lem	Nomaande	mgc	Morokodo	ngp	Ngulu
kea	Kabuverdianu	lgg	Lugbara	mgh	Makhuwa-Meetto	nhr	Naro
ken	Kenyang	lgm	Lega-mwenga	mgo	Meta'	nhu	Noone
khy	Kele / Lokele	lia	Limba, West-Central	mgq	Malila	nih	Nyiha
kia	Kim	lik	Lika	mgr	Mambwe-Lungu	nim	Nilamba / kinilyamba
kik	Gikuyu / Kikuyu	lin	Lingala	mgw	Matumbi	nin	Ninzo
kin	Kinyarwanda	lip	Sekpele	mif	Mofu-Gudur	niy	Ngiti
kiz	Kisi	lmd	Lumun	mkl	Mokole	nka	Nkoya / ShiNkoya
kki	Kagulu	lmp	Limbum	mlg	Malagasy	nko	Nkonya
kkj	Kako	lnl	Banda, South Central	mlr	Vame	nla	Ngombale
kln	Kalenjin	log	Logo	mmy	Migaama	nnb	Nande / Ndandi
klu	Klao	lom	Loma	mnf	Mundani	nnh	Ngiemboon
kma	Konni	loq	Lobala	mnk	Mandinka	nnq	Ngindo
kmb	Kimbundu	lot	Latuka	moa	Mwan	nse	Chinsenga
kmy	Koma	loz	Silozi	mos	Moore	nnw	Nuni, Southern
knf	Mankanya	lro	Laro	moy	Shekkacho	nso	Sepedi
kng	Kongo	lsm	Saamya-Gwe / Saamia	moz	Mukulu	ntr	Delo
knk	Kuranko	lth	Thur / Acholi-Labwor	mpe	Majang	nuj	Nyole
kno	Kono	lto	Tsotso	mpg	Marba	nus	Nuer
koo	Konzo	lua	Tshiluba	mqb	Mbuko	nwb	Nyabwa
koq	Kota	luc	Aringa	msc	Maninka, Sankaran	nxd	Ngando
kqn	Kikaonde	lue	Luvale	mur	Murle	nya	Chichewa
kqp	Kimré	lug	Luganda	muy	Muyang	nyb	Nyangbo
kqs	Kisi	lun	Lunda	mwe	Mwera	nyd	Olunyole / Nyore
kqy	Koorete	luo	Dholuo / Luo	mwm	Sar	nyf	Giryama
kri	Krio	lwg	Wanga	mwn	Cinamwanga	nyk	Nyaneka
krs	Gbaya	lwo	Luwo	mws	Mwimbi-Muthambi	nym	Nyamwezi
krw	Krahn, Western	maf	Mafa	myb	Mbay	nyn	Nyankore / Nyankole
krx	Karon	mas	Maasai	myk	Sénoufo, Mamara	nyo	Nyoro
ksb	Shambala / Kishambala	maw	Mampruli	myx	Masaaba	nyu	Nyungwe
ksf	Bafia	mbu	Mbula-Bwazza	mzm	Mumuye	nyy	Nyakyusa-Ngonde / Kyangonde
ksp	Kabba	mck	Mbunda	mzw	Deg	nza	Mbembe, Tigon
ktj	Krumen, Plapo	mcn	Masana / Massana	naq	Khoekhoe	nzi	Nzema
ktu	Kikongo	mcp	Makaa	naw	Nawuri	odu	Odual
kua	Oshiwambo	mcu	Mambila, Cameroon	nba	Nyemba	ogo	Khana
kub	Kutep	mda	Mada	nbl	IsiNdebele	oke	Okpe
kuj	Kuria	mdm	Mayogo	ncu	Chunburung	okr	Kirike
kus	Kusaal	mdy	Maale	ndc	Ndau	oku	Oku
kvj	Psikye	men	Mende	nde	IsiNdebele	orm	Oromo
kwn	Kwangali	meq	Merey	ndh	Ndali	ozm	Koonzime
kyf	Kouya	mer	Kimiiru	ndj	Ndamba	pcm	Nigerian Pidgin
kyq	Kenga	mev	Maan / Mann	ndo	Ndonga	pem	Kipende
kzr	Karang	mfe	Morisyen / Mauritian Creole	ndv	Ndut	pkb	Kipfokomo / Pokomo
lai	Lambya	mfg	Mogofin	ndz	Ndogo		

Table E.2: AfroLID covered Languages - Part II

ISO-3	Language	ISO-3	Language	ISO-3	Language
pov	Guinea-Bissau Creole	tcd	Tafi	won	Wongo
poy	Pogolo / Shipogoro-Pogolo	ted	Krumen, Tepo	xan	Xamtanga
rag	Lulogooli	tem	Timne	xed	Hdi
rel	Rendille	teo	Teso	xho	Isixhosa
rif	Tarifit	tex	Tennet	xnz	Mattokki
rim	Nyaturu	tgw	Senoufo, Tagwana	xog	Soga
rnd	Uruund	thk	Tharaka	xon	Konkomba
rng	Ronga / ShiRonga	thv	Tamahaq, Tahaggart	xpe	Kpelle
rub	Gungu	tir	Tigrinya	xrb	Karaboro, Eastern
run	Rundi / Kirundi	tiv	Tiv	xsm	Kasem
rwk	Rwa	tke	Takwane	xtc	Katcha-Kadugli-Miri
sag	Sango	tlj	Talinga-Bwisi	xuo	Kuo
saq	Samburu	tll	Otetela	yal	Yalunka
sba	Ngambay	tog	Tonga	yam	Yamba
sbd	Samo, Southern	toh	Gitonga	yao	Yao / Chiyao
sbp	Sangu	toi	Chitonga	yat	Yambeta
sbs	Kuhane	tpm	Tampulma	yba	Yala
sby	Soli	tsc	Tshwa	ybb	Yemba
sef	Sénoufo, Cebaara	tsn	Setswana	yom	Ibinda
ses	Songhay, Koyraboro Senni	tso	Tsonga	yor	Yoruba
sev	Sénoufo, Nyarafolo	tsw	Tsishingini	yre	Yaoure
sfw	Sehwi	ttj	Toro / Rutoro	zaj	Zaramo
sgw	Sebat Bet Gurage	ttq	Tawallammat	zdj	Comorian, Ngazidja
shi	Tachelhit	ttr	Nyimatli	zga	Kinga
shj	Shatt	tui	Toupouri	ziw	Zigula
shk	Shilluk	tul	Kutule	zne	Zande / paZande
sid	Sidama	tum	Chitumbuka	zul	Isizulu
sig	Paasaal	tuv	Turkana		
sil	Sisaala, Tumulung	tvu	Tunen		
sna	Shona	twi	Twi		
snf	Noon	umb	Umbundu		
sng	Sanga / Kiluba	urh	Urhobo		
snw	Selee	uth	ut-Hun		
som	Somali	vag	Vagla		
sop	Kisonge	vai	Vai		
sor	Somrai	ven	Tshivenda		
sot	Sesotho	vid	Chividunda		
soy	Miyobe	vif	Vili		
spp	Senoufo, Supyire	vmk	Makhuwa-Shirima		
ssw	Siswati	vmw	Macua		
suk	Sukuma	vun	Kivunjo		
sus	Sosoxui	vut	Vute		
swa	Swahili	wal	Wolaytta		
swc	Swahili Congo	wbi	Vwanji		
swh	Swahili	wec	Guere		
swk	Sena, Malawi	wes	Pidgin, Cameroon		
sxb	Suba	wib	Toussian, Southern		
taq	Tamasheq	wmw	Mwani		
tcc	Datooga	wol	Wolof		

Table E.3: AfroLID covered Languages - Part III.