IMPROVING AFRICAN LANGUAGE IDENTIFICATION WITH MULTI-TASK LEARNING

Ife Adebara[1] **AbdelRahim Elmadany**[1] **Muhammad Abdul-Mageed**[1,2]
[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@}ubc.ca`

ABSTRACT

We present AfroLID-$v2.0$, a multi-task neural language identification toolkit for 517 African languages and varieties. The languages that make up AfroLID-$v2.0$ belong to 14 language families spoken across 50 African countries. To ensure robustness of AfroLID-$v2.0$, we employ a multi-domain, multi-script dataset. Compared to a previous version of the tool (AfroLID), AfroLID-$v2.0$ is trained with a multi-task learning objective exploiting language family information. That is, AfroLID-$v2.0$ performs language identification as the main task and language family identification as an auxiliary task. We demonstrate how our multi-task learning setup yields better performance compared to all previous work, allowing AfroLID-$v2.0$ to reach a 96.44 F_1 on our blind test set. Language identification is a core technology in NLP, and we hope that AfroLID-$v2.0$ will be a valuable contribution to multilingual NLP in general and African NLP in particular.

1 INTRODUCTION

Language identification (LID) refers to the process of determining the language of a text or speech segment. With the increased use of social media, there is now a larger amount of multilingual data available, making automatic language identification a crucial initial step in the proper handling of human language. LID covers a variety of forms of communication, such as speech, sign language, written text, and others. It also involves identifying languages in datasets that contain a mixture of codes. Unfortunately, resources are lacking for the development of language identification tools for the majority of the world's languages, including most African languages (Abdul-Mageed et al. (2020); Thara & Poornachandran (2021)).

Due to the limited resources available for many African languages, it is crucial to explore various methods to improve accuracy of LID. In this work, we explore multitask settings to improve LID of 517 African languages and language varieties. We refer to the languages we work on as languages and language varieties because the difference between languages and dialects is not clear for many African languages. So that while some may refer to some speech forms as distinct languages, others may identify them as dialects of the same language (Wichmann (2020)). Specifically, we explore multi-task settings with LID as the primary task and language family information of the 517 languages in our LID data as the secondary task.

Our contribution is as follows:

1. We develop AfroLID-$v2.0$, 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-$v2.0$ on Twitter data.

The rest of this paper goes as follows: In Section 2, we provide a short literature review. In Section 3, we describe AfroLID-$v2.0$ including the language families, script and grammatical information of the languages we cover. In Section 4, we describe the details of our experiments and give details of our model performance and analysis in Section 5. We conclude in Section 6.

2 RELATED WORK

LID plays a crucial role in enabling communication and processing of multilingual data. LID tools allow for the automatic detection of the language used in a given text. This is important in a variety of applications including machine translation, information retrieval, and text classification. A few language identification tools provide support for some African languages. Some of those tools are CLD2 (McCandless (2010)), CLD3 (Salcianu et al. (2018)), Equilid (Jurgens et al. (2017)), FastText, Franc, LangDetect (Shuyo (2010)) and Langid.py (Lui & Baldwin (2012)) and works such as (Abdul-Mageed et al. (2020; 2021)) and (Nagoudi et al. (2022)). For many African languages, LID tools are either unavailable or have poor performance, making LID an important research area for AfricanNLP. To the best of our knowledge AfroLID-$v2.0$ is the best performing model for most African languages.

3 AFROLID-$v2.0$

AfroLID-$v2.0$ is trained using a multi-domain, multi-script LID and language family dataset that was manually curated. There is no overlap between the data for the LID and language family dataset. AfroLID-$v2.0$ consists of 517 African languages and language varieties domiciled in 50 African countries. These languages belong to 14 language families 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. We provide information about the languages supported in Tables 3, 4, and 5 and the language families in Figure 3 in the Appendix.

3.1 LANGUAGE FAMILY

We used language family classification information available in Ethnologue (Eberhard et al. (2021)). We labelled each text manually using the entire tree information with each subcategory separated with an underscore.

3.2 SCRIPT

The African languages in AfroLID-$v2.0$ are written in 5 different scripts. Majority of them use Latin scripts, including Berber Latin, while 14 languages are written in four different scripts including Arabic, Coptic, Ethiopic, and Vai. Although some of these languages are written in multiple scripts, in most cases, we were able to access only one script. For instance, Koorete (kqy) is written in Ethiopic and Latin scripts but we have only Latin texts. Harari (har) is written in Arabic, Ethiopic, and Latin scripts but we have only Latin scripts.

3.3 GRAMMATICAL INFORMATION

Sentential Word Order: AfroLID-$v2.0$ consists of 5 out of 7 word orders across human languages around the world. These are subject-verb-object (SVO), subject-object-verb (SOV), 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). The word orders not represented in our data includes object-verb-subject (OVS) and object-subject-verb (OSV).

Diacritics: The use of diacritics are pervasive in many African languages. Diacritics often have grammatical or lexical functions in these languages and may indicate length, tone, nasalization, and other linguistic features (Adebara et al. (2022)). AfroLID-$v2.0$ consists of 295 languages that use diacritics. Diacritics can be placed above, below, and across a letter and may be accent marks, punctuation marks, or other special characters (Wells (2000); Hyman (2003); Creissels et al. (2008)).

4 EXPERIMENTAL SETUP

For our LID, we used data for AfroLID (Adebara et al. (2022)) which are randomly selected $5,000$, 50, and 100 sentences for Train, Development, and Test respectively for each of the 517 languages or language varieties in our manually curated dataset [1]. In all, AfroLID-$v2.0$ contains $2,496,980$ Train , $25,800$ Dev, and $51,400$ Test examples. For the language family datasets, we have $37,147$ examples in Train, $4,643$ in Development, and $4,650$ in Test respectively.

Preprocessing. We perform minimal preprocessing to ensure that our data represent naturally occurring text. 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.

Vocabulary. We experiment with byte-pair (BPE), encodings. We used vocabulary sizes of 100K.

Hyperparameter Search and Training. We experiment with different sampling methods for training. Specifically, we experiment with the concatenation, uniform, and temperature sampling methods. For the temperature sampling method, we experimented with temperature 1, 1.5, and 2 respectively.

Implementation. We use a Transformer architecture trained from scratch. We use 12 attention layers with 12 heads in each layer, 768 hidden dimensions, making up about $200M$ parameters. All other hyperparameter settings are similar to XLMR base model (Conneau et al. (2020)). We also use temperature sampling method with a temperature of 1.5. We implemented our model in Fairseq.

Evaluation. We report our results in both macro F1-score and accuracy, selecting our best model on the Dev. All results are reported for the Test set.

5 MODEL PERFORMANCE AND ANALYSIS

We report best performance on the model that use temperature 1.5 sampling. In Table 1, we show the results for each setting and compare results with the single task AfroLID (Adebara et al. (2022)). We also show the distribution of f1 scores in Figure 1. AfroLID-$v2.0$ improves f1 scores across many languages. However, we report low scores for Xhosa and Zulu. We assume this may be due to the presence of 10 South African Languages in AfroLID-$v2.0$. These languages share vocabulary and it is not uncommon to find a lot of code-mixing in these texts (Finlayson & Slabbert (1997); Mabule (2015)). We make this assumption because a deeper investigation of the errors for Xhosa and Zulu show that AfroLID-$v2.0$ often selects one of the other South African languages when it makes errors.

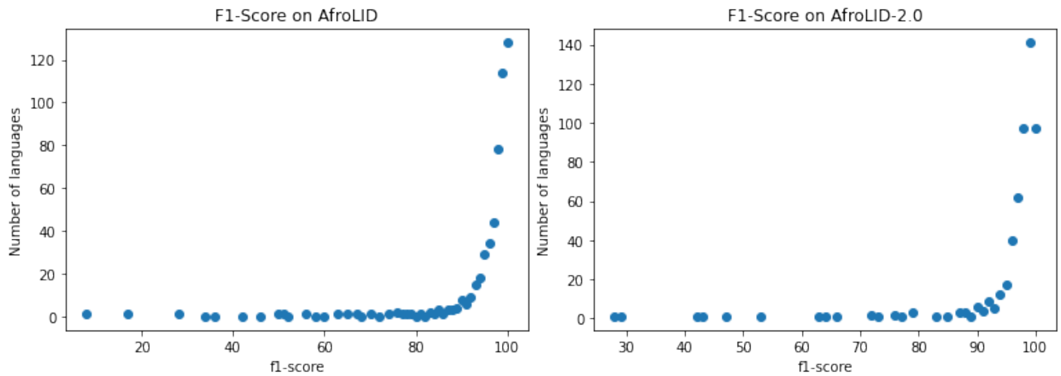
Model	Setting	$\mathbf{F_1}$-score	Accuracy
AfroLID $v2.0$	Concatenate	95.27	95.39
	Temperature = 1	95.24	95.34
	Temperature = 1.5	$\mathbf{96.44}$	$\mathbf{96.51}$
	Temperature = 2	95.17	95.28
	Uniform	93.90	94.24
AfroLID*		95.95	96.01

Table 1: Results on the different settings of AfroLID-$v2.0$ on Test dataset and results from AfroLID BPE.. **Bolded**: best result on Test. **AfroLID*** results as shown in Adebara et al. (2022)

5.1 AFROLID-$v2.0$ IN COMPARISON

We compare AfroLID-$v2.0$ with CLD2, CLD3, Franc, LangDetect, and Langid.py. We select 17 languages for this comparison based on the number of languages supported in CLD2, CLD3, langid.py and LangDetect as described in AfroLID (Adebara et al. (2022)). AfroLID-$v2.0$ outperforms all models on seven of 17 languages. We show the results of our comparison in Table 2. We also compare the performance of AfroLID-$v2.0$ with AfroLID on Naija-Senti corpus (Muhammad et al. (2022)).

[1]About 10 languages had less than these numbers due to available data.



(a) AfroLID

(b) AfroLID-$v2.0$

Figure 1: Scatter plots showing the distribution of results on AfroLID and AfroLID-$v2.0$

Lang.	CLD2	CLD3	Langid.py	LangDetect	Franc	AfroLID	AfroLID-$v2.0$
afr	94.00	91.00	69.00	88.23	81.00	97.00	**98.00**
amh	-	97.00	**100.00**	-	35.00	97.00	98.00
hau	-	83.00	-	-	77.00	88.00	**90.00**
ibo	-	96.00	-	-	88.00	97.00	**98.00**
kin	**92.00**	-	45.00	-	47.00	89.00	89.00
lug	84.00	-	-	-	64.00	**87.00**	**87.00**
mlg	-	**100.00**	98.00	-	-	**100.00**	99.00
nya	-	96.00	-	-	75.00	92.00	**97.00**
sna	-	**100.00**	-	-	91.00	97.00	97.00
som	-	92.00	-	-	89.00	**95.00**	**95.00**
sot	-	**99.00**	-	-	93.00	88.00	93.00
swa	99.00	91.00	90.00	**100.00**	-	92.00	92.00
swc	93.00	94.00	96.00	**97.02**	-	87.00	88.00
swh	89.00	**92.00**	88.23	87.19	70.00	77.00	79.00
xho	-	59.00	**88.00**	-	30.00	67.00	64.00
yor	-	25.00	-	-	66.00	**98.00**	**97.00**
zul	-	**89.00**	20.00	-	40.00	50.00	53.00

Table 2: A comparison of results on AfroLID-$v2.0$ with CLD2, CLD3, Langid.py, LangDetect, Franc and AfroLID using F_1-score on the Test set. $-$ indicates that the tool does not support the language.

We do this comparison to evaluate the performance of AfroLID-$v2.0$ in out of domain scenarios, since there was no Twitter data in the training data for AfroLID-$v2.0$. We find AfroLID-$v2.0$ outperforming AfroLID on all languages in Naija-Senti corpus. We show the results in Figure 2.

6 Conclusion

We introduced our novel African language identification tool, AfroLID-$v2.0$. A multi-task model covering 517 African languages. AfroLID-$v2.0$ is a publicly available tool that covers a large number of African languages and language varieties. AfroLID-$v2.0$ also has the advantages of wide geographical coverage (50 African countries) and linguistic diversity. Future work will explore different linguistically motivated experiments to improve model performance. We will also investigate the effect of having a large number of languages on model performance.

References

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

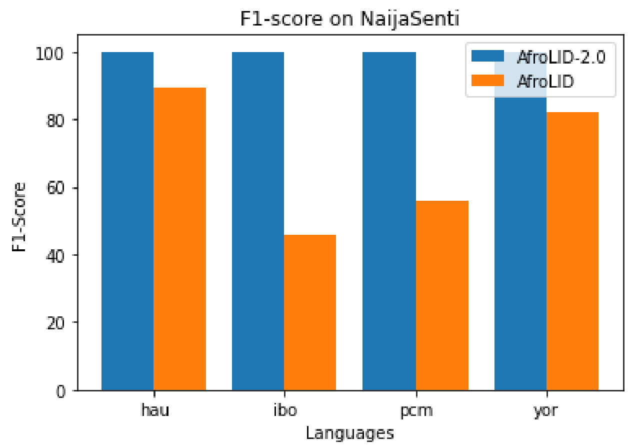


Figure 2: A comparison of AfroLID-$v2.0$ and AfroLID on naija-senti corpus

Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.472. URL https://aclanthology.org/2020.emnlp-main.472.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 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)*, pp. 7088–7105, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.551. URL https://aclanthology.org/2021.acl-long.551.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Denis Creissels, Gerrit J Dimmendaal, Zygmunt Frajzyngier, and Christa König. Africa as a morphosyntactic area. *A linguistic geography of Africa*, 86150, 2008. URL https://www.cambridge.org/core/books/abs/linguistic-geography-of-africa/africa-as-a-morphosyntactic-area/7311B4CA78BC172C656934E1498A7D96.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International, 2021. URL http://www.ethnologue.com.ezproxy.library.ubc.ca.

Rosalie Finlayson and Sarah Slabbert. "We just mix": code switching in a South African township. 1997(125):65–98, 1997. doi: doi:10.1515/ijsl.1997.125.65. URL https://doi.org/10.1515/ijsl.1997.125.65.

Larry M Hyman. African languages and phonological theory. *Glot International*, 7(6):153–163, 2003. URL https://www.researchgate.net/publication/245231198_African_languages_and_phonological_theory.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 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)*, pp. 51–57, 2017. URL https://aclanthology.org/P17-2009.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3005.

D R Mabule. What is this? is it code switching, code mixing or language alternating? *Journal of Educational and Social Research*, 5(1), 2015. ISSN 2240-0524. URL https://www.mcser.org/journal/index.php/jesr/article/view/5628.

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

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. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis, 2022.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 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)*, pp. 628–647, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.47. URL https://aclanthology.org/2022.acl-long.47.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, et al. Compact language detector v3, 2018. URL https://chromium.googlesource.com/external/github.com/google/cld_3/+/f01672272dacc4cb3409f458ed61f7d4eb0f47de/README.md.

Nakatani Shuyo. Language detection library for java, 2010. URL http://code.google.com/p/language-detection/.

S. Thara and Prabaharan Poornachandran. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837–118850, 2021. doi: 10.1109/ACCESS.2021.3104106.

John C. Wells. Orthographic diacritics and multilingual computing. *Language Problems and Language Planning*, 24:249–272, 2000. URL https://www.researchgate.net/publication/233567411_Orthographic_diacritics_and_multilingual_computing.

Søren Wichmann. How to Distinguish Languages and Dialects. *Computational Linguistics*, 45(4): 823–831, 01 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00366. URL https://doi.org/10.1162/coli_a_00366.

A Appendix

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 3: AfroLID-$v2.0$ 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 4: AfroLID-$v2.0$ 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 5: AfroLID-$v2.0$ covered Languages - Part III.

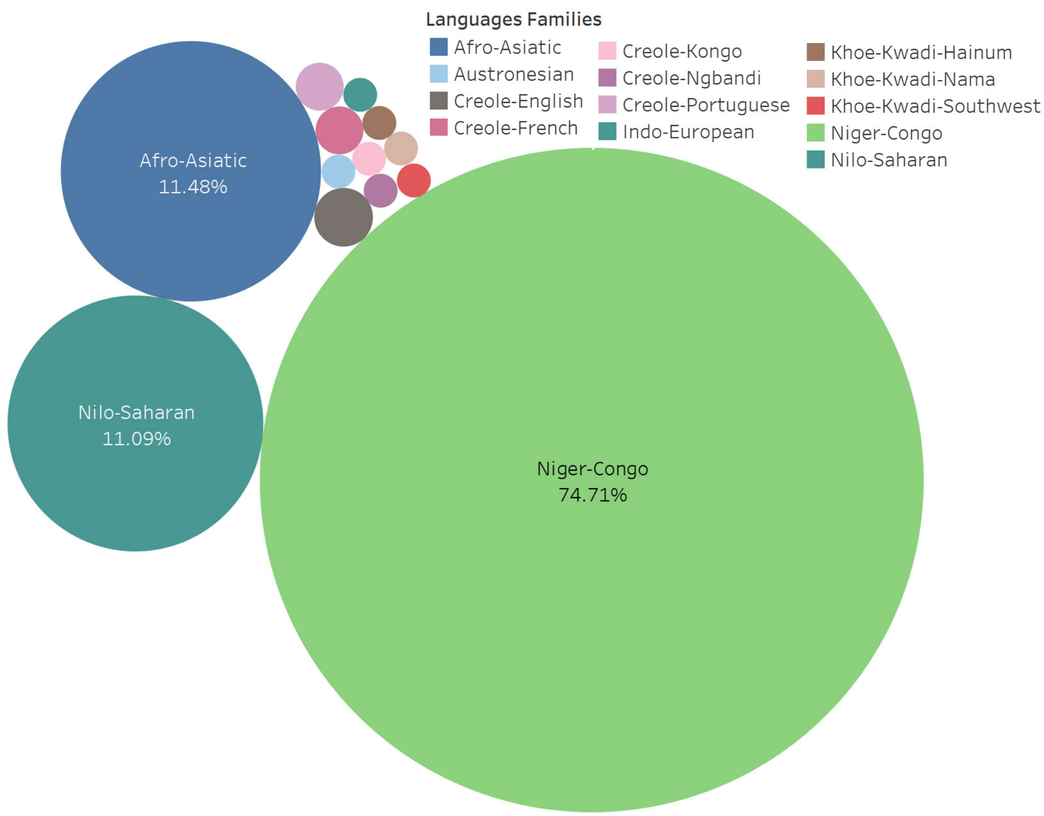


Figure 3: Families in AfroLID-$v2.0$