

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA

WITH AN INTRODUCTION TO THE GIZ AI4D
AFRICAN LANGUAGE DATASET CHALLENGE

3 July 2020 from 14:00 to 16:00 pm CAT/CEST/UTC+2



Implemented by
giz
Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.



IDRC | CRDI

International Development Research Centre
Centre de recherches pour le développement international

moz://a

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA
WITH AN INTRODUCTION TO THE **GIZ AI4D AFRICAN**
LANGUAGE DATASET CHALLENGE

WELCOME ADDRESS

KATHLEEN SIMINYU — REGIONAL COORDINATOR OF
AI4D



International Development Research Centre
Centre de recherches pour le développement international



AGENDA

14:00 – 14:10

Welcome Address

Kathleen Siminyu, Regional Coordinator of AI4D

14:10 – 14:30

NLP research in low-resourced languages

Cristina España-Bonet, DFKI

14:30 – 14:45

Yoruba and beyond: NLP from an African perspective

Jesujoba Alabi, DFKI

14:45 – 15:00

Ensuring good text quality in African language datasets

David Adelani, Saarland University & Masakhne

15:00 – 15:15

Coffee Break

15:15 – 15:25

Language data for African languages

Andrea Lösch, DFKI

15:25 – 15:55

Discussion round: Data collection approaches in Europe and Africa

Moderator: Andrea Lösch, DFKI

15:55 – 16:00

Closing Statement



AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA

WITH AN INTRODUCTION TO THE GIZ AI4D
AFRICAN LANGUAGE DATASET CHALLENGE

3 July 2020 from 14:00 to 16:00 pm CAT/CEST/UTC+2



Implemented by
giz
Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.



International Development Research Centre
Centre de recherches pour le développement international



Yorùbá and beyond: NLP from an African perspective

Jesujoba O Alabi

Outline

- Introduction
- Yorùbá
- Our recent work
- BERT for Low Resource Languages
- Other Initiatives

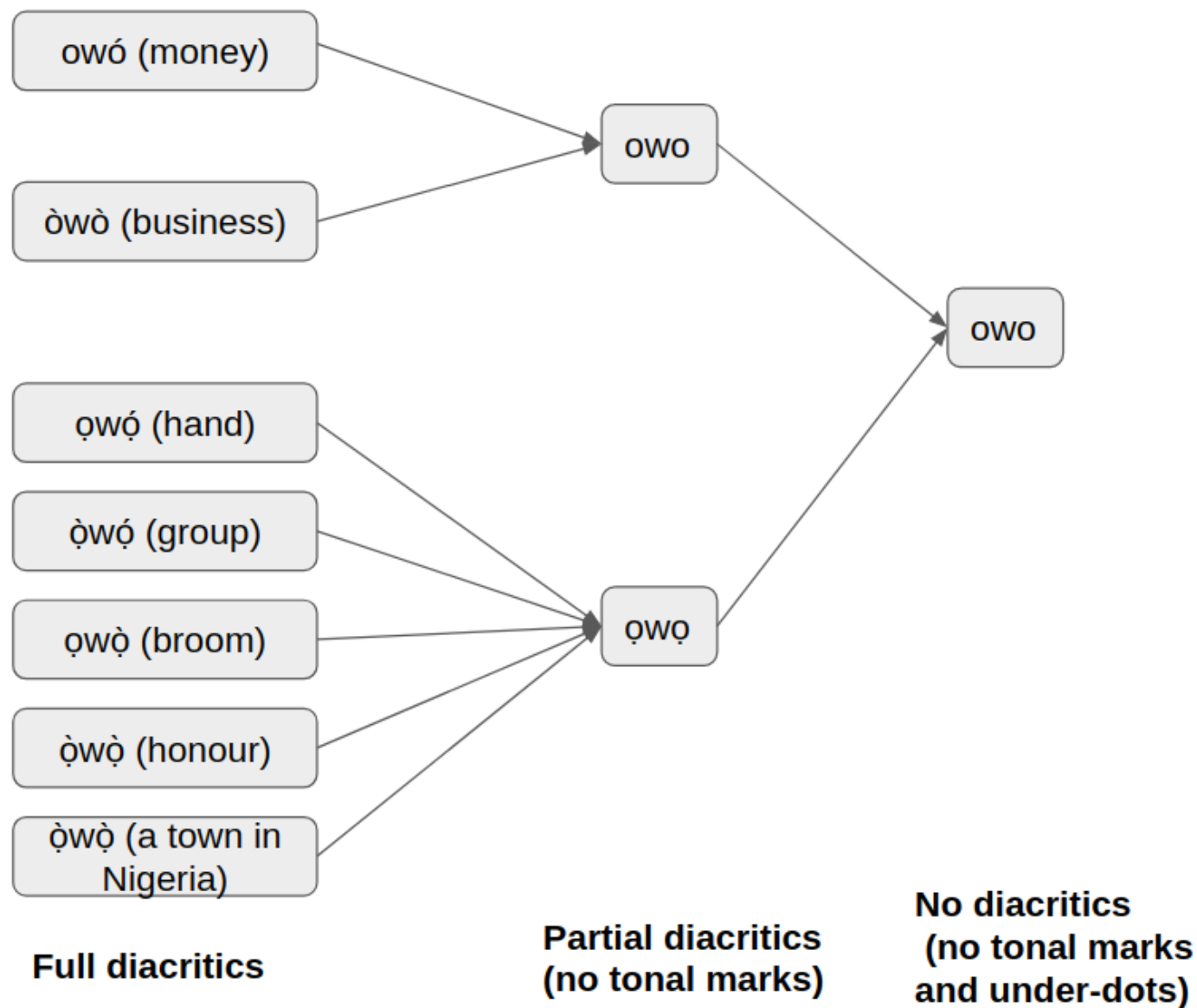
Introduction

- Africa has a lot of languages
- 1500-2000 languages
- 1/3 of world languages
- typically low resource languages
- Yorùbá is the 3rd most spoken language.
- Niger Congo – (benue-congo)
- Uses diacritics



Yorùbá

- Most of the Yorùbá texts found online either use
 - correct Yorùbá orthography or
 - replace diacritized characters Ashiah (2014)
- Yorùbá text in the public domain today is not well diacritized.
- Wikipedia is not an exception.
- No evaluation dataset
- Small Corpora



Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi

Jesujoba O. Alabi^{*†‡} Kwabena Amponsah-Kaakyire^{*†‡} David I. Adelani^{†‡} Cristina España-Bonet^{†‡}

[†]DFKI GmbH, Saarbrücken, Germany

[‡]Spoken Language Systems (LSV), Saarland Informatics Campus, [‡]Saarland University, Saarbrücken, Germany

{jesujoba.oluwadara.alabi, kwabena.amponsah-kaakyire, cristinae}@dfki.de, didelani@lsv.uni-saarland.de

Abstract

The success of several architectures to learn semantic representations from unannotated text and the availability of these kind of texts in online multilingual resources such as Wikipedia has facilitated the massive and automatic creation of resources for multiple languages. The evaluation of such resources is usually done for the high-resourced languages, where one has a smorgasbord of tasks and test sets to evaluate on. For low-resourced languages, the evaluation is more difficult and normally ignored, with the hope that the impressive capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced setting too. In this paper we focus on two African languages, Yorùbá and Twi, and compare the word embeddings obtained in this way, with word embeddings obtained from curated corpora and a language-dependent processing. We analyse the noise in the publicly available corpora, collect high quality and noisy data for the two languages and quantify the improvements that depend not only on the amount of data but on the quality too. We also use different architectures that learn word representations both from surface forms and characters to further exploit all the available information which showed to be important for these languages. For the evaluation, we manually translate the wordsim-353 word pairs dataset from English into Yorùbá and Twi. We extend the analysis to contextual word embeddings and evaluate multilingual BERT on a named entity recognition task. For this, we annotate with named entities the Global Voices corpus for Yorùbá. As output of the work, we provide corpora, embeddings and the test suits for both languages.

Keywords: Multilingual embeddings, Low-resource language, Yorùbá, and Twi

Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi

01

Investigate **quality** of word embeddings on two African languages:

- FastText & BERT

02

Compare pre-trained embeddings & our trained embeddings

03

Analyze the impact of adding noisy-texts (low-quality) to high quality curated dataset.

04

Learn representations using word and sub-word representations:

- FastText & CWE

Description	Source URL	#tokens	Status	C1	C2	C3
<i>Yorùbá</i>						
Lagos-NWU corpus	github.com/Niger-Volta-LTI	24,868	clean	✓	✓	✓
Alákòwé	alakoweyoruba.wordpress.com	24,092	clean	✓	✓	✓
Òrò Yorùbá	oroyoruba.blogspot.com	16,232	clean	✓	✓	✓
Èdè Yorùbá Rẹwà	deskgram.cc/edeyorubarewa	4,464	clean	✓	✓	✓
Doctrine \$ Covenants	github.com/Niger-Volta-LTI	20,447	clean	✓	✓	✓
Yorùbá Bible	www.bible.com	819,101	clean	✓	✓	✓
GlobalVoices	yo.globalvoices.org	24,617	clean	✓	✓	✓
Jehova Witness	www.jw.org/yo	170,203	clean	✓	✓	✓
Ìrìnkèrindò nínú igbó elégbèje	manual	56,434	clean	✓	✓	✓
Igbó Olódùmarè	manual	62,125	clean	✓	✓	✓
JW300 Yorùbá corpus	opus.nlpl.eu/JW300.php	10,558,055	clean	✗	✗	✓
Yorùbá Tweets	twitter.com/yobamoodua	153,716	clean	✓	✓	✓
BBC Yorùbá	bbc.com/yoruba	330,490	noisy	✗	✓	✓
Voice of Nigeria Yorùbá news	von.gov.ng/yoruba	380,252	noisy	✗	✗	✓
Yorùbá Wikipedia	dumps.wikimedia.org/yowiki	129,075	noisy	✗	✗	✓
<i>Twi</i>						
Bible	www.bible.com	661,229	clean	✓	✓	✓
Jehovah's Witness	www.jw.org/tw	1,847,875	noisy	✗	✗	✓
Wikipedia	dumps.wikimedia.org/twwiki	5,820	noisy	✗	✓	✓
JW300 Twi corpus	opus.nlpl.eu/JW300.php	13,630,514	noisy	✗	✗	✓

Table 1: Summary of the corpora used in the analysis. The last 3 columns indicate in which dataset (C1, C2 or C3)

i. Curated Small Dataset (clean), C1

- i. Yorùbá: 1.6 million tokens
- ii. Twi: 735k tokens

ii. Curated Small Dataset (clean + noisy), C2 (Wikipedia, BBC Yorùbá)

- i. Yorùbá: 2 million tokens
- ii. Twi: 742k tokens

iii. Curated Large Dataset, C3

Word Embeddings

- Word embeddings have been proven to be very useful for training downstream natural language processing (NLP) tasks.
- Contextualized have been shown to further improve the performance of NLP.

Task	Dataset	Model	Metric
Word Similarity	Translated WordSim-353	FastText, CWE	Spearman Correlation
Named Entity Recognition	Global Voices News Yorùbá Dataset	BERT	F1-Score

Evaluation on FastText

Model	Twi		Yorùbá	
	Vocab Size	Spearman ρ	Vocab Size	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	0.388	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	0.391

Table 2: FastText embeddings: Spearman ρ correlation between human judgements and similarity scores on the wordSim-353 for the three datasets analysed (C1, C2 and C3)

BERT Evaluation on NER Task

Entity type	Number of tokens			
	Total	Train	Val.	Test
ORG	289	214	40	35
LOC	613	467	47	99
DATE	662	452	86	124
PER	688	469	109	110
O	23,988	17,819	2,413	4,867

Table 3: Number of tokens per named entity type in the Global Voices Yorùbá corpus.

Embedding Type	DATE	LOC	ORG	PER	F1-score
Pre-trained <i>uncased</i> Multilingual-bert (Multilingual vocab)	44.6	33.9	12.1	5.7	27.1 \pm 0.7
Fine-tuned <i>uncased</i> Multilingual-bert (Multilingual vocab)	64.0	65.3	38.8	47.4	56.4 \pm 2.4
Fine-tuned <i>uncased</i> Multilingual-bert (Yorùbá vocab)	67.0	71.5	40.4	49.4	60.1 \pm 0.8

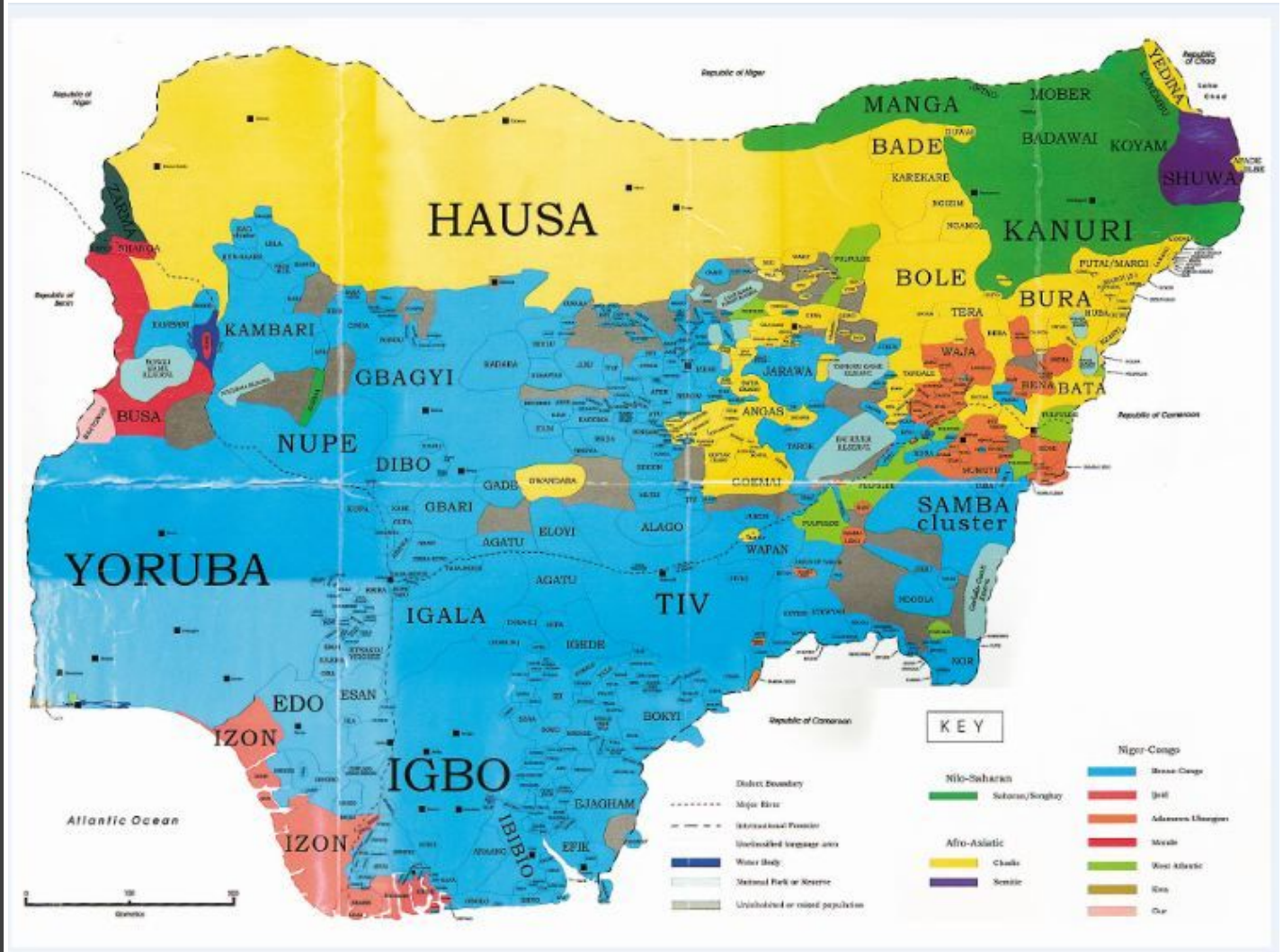
Table 4: NER F1 score on Global Voices Yorùbá corpus after fine-tuning BERT for 10 epochs. Mean F1-score computed after 5 runs

BERT and other LMs

- Lauscher et al.(2020) find that the transfer for multilingual trans-former models is less effective for resource-lean settings and distant languages.
- Fast adaptation method for obtaining a bilingual BERT of English and a target language. Tran (2020)
- Where target language could be any African Language.
- Just mono lingual data is needed.

Naija BERT

- Over 500 spoken languages
- Yoruba, Hausa and Igbo are major
- Can we have BERT for all possible Nigerian Languages?
- Can we have BERT for Nigerian Languages from the same class?
- BERT in low resource setting.



Way to go

- Replicate our work for other African languages
- For example, Xhosa
 - Xhosa has a lot of resources online (JW300, OPUS, Common crawl, etc)
 - Limited Wikipedia articles
 - No diacritics
 - No word embeddings!

Other ways to go

- Standardization of existing dataset – e.g. BBC Yoruba, Wikipedia, VON
- Automatic Diacritics Application – NN Based
- Wikipedia articles writing/translation

Thank you for Listening!

References

- Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. ArXiv, abs/2005.00633.
- Ke Tran, 2020. From English to Foreign Languages: Transferring Pretrained Language Models. ArXiv, abs/2002.07306..
- Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. 2020. Massivevs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yor`ub`a and Twi. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 2747–2755, Marseille, France. European Language Resources Association.
- <https://translatorswithoutborders.org/language-data-nigeria/>
- https://upload.wikimedia.org/wikipedia/commons/thumb/c/c4/African_language_families_en.svg/553px-African_language_families_en.svg.png
- <https://i.pinimg.com/originals/00/13/f4/0013f4b94fa1b2b85221f9e22fd43b8c.jpg>
- https://www.nationsonline.org/oneworld/african_languages.htm

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA

WITH AN INTRODUCTION TO THE GIZ AI4D
AFRICAN LANGUAGE DATASET CHALLENGE

3 July 2020 from 14:00 to 16:00 pm CAT/CEST/UTC+2



Implemented by
giz
Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

FAIR FORWARD
Artificial Intelligence for all.



International Development Research Centre
Centre de recherches pour le développement international



Ensuring good text quality in African Language Datasets

David Adelani
Saarland University / Masakhane



SIC Saarland Informatics
Campus



Outline

- Popular resources for AfricaNLP
- Known text quality issues in African languages
- Practical solutions
- Case Study: Yoruba Text quality verification

Popular Resources for AfricaNLP

Data source	Number of African Languages
Bible	>1000
JW300	101
Wikipedia	38 (~100K articles vs 6M English articles)
Common Crawl	28 (out of 160 identified languages)
VOA	13
BBC	11

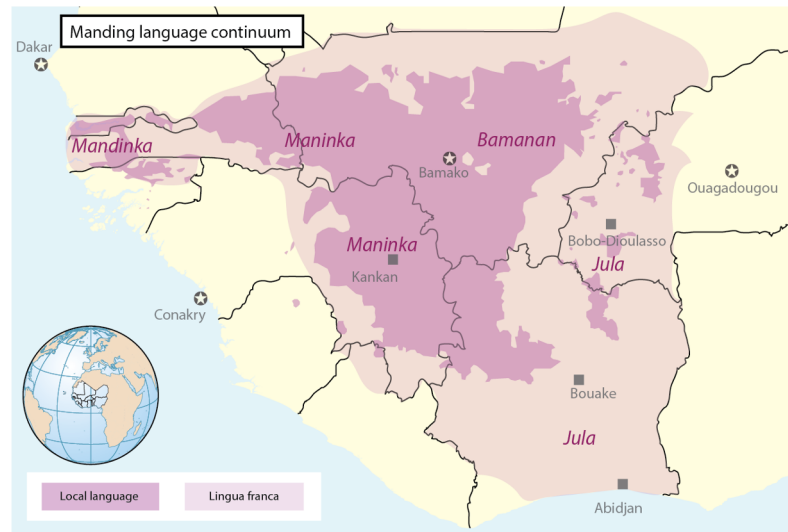
10 Most resourced African Languages

- 1) Afrikaans
- 2) Swahili
- 3) Malagasy
- 4) Somali
- 5) Amharic
- 6) Hausa
- 7) Yoruba
- 8) Kinyarwanda
- 9) Zulu
- 10) Tigrinya

https://en.wikipedia.org/wiki/Bible_translations_into_the_languages_of_Africa
<https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>

Known Text Quality Issues in African Languages

- Spoken but not often written
 - E.g. Bambara (e.g on VOA but only audio), Bible is available.
- Code Switching corpus
 - E.g. Wolof + English/French
- Mixed dialects / Languages
 - Twi (Asante & Akuapem dialects)
 - Rwanda-Rundi
 - Fula with different dialects and writing styles
- Diacritics problem
 - Yoruba



Mixed dialects and languages

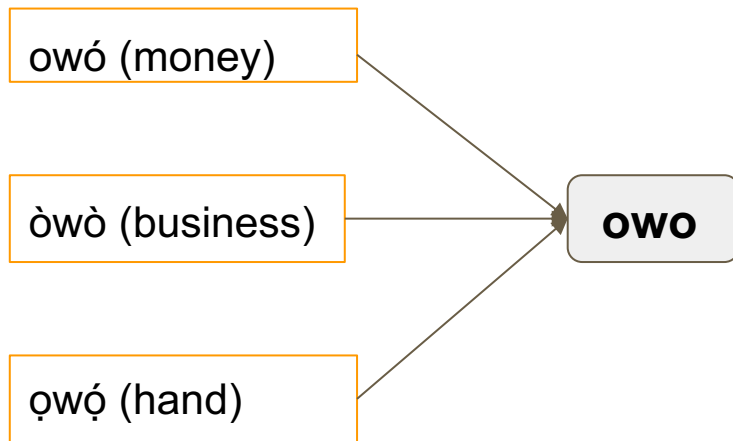
- Twi (Asante & Akuapem dialects) e.g in JW300
 - Asante: Me papa **fi**e yɛ **fɛ**.
 - Akuapem: Me papa **fi** yɛ **fɛw**.
 - English: My dad's house looks nice.
- Rwanda-Rundi e.g in VOA / BBC.
 - Kinyarwanda: Uriya **mugore yaberewe**.
 - Kirundi: Uriya **mukenyezi yarimvye**.
 - English: That woman looks good in her dress.
- Fula dialects e.g about 9 dialects
 - Sometimes **different alphabets** e.g **Nigerian Fulfulde** vs **Pular / Pulaar**
 - 4 well resourced dialects : Pular, Pulaar, Nigerian Fulfulde and Adamawa Fulfulde.

Countries that speak Fula



Yorùbá diacritics Problem

Pronunciation only depends on word context



Impact of low-quality data on word embeddings

Model	Twi		Yorùbá	
	Vocab Size	Spearman ρ	Vocab Size	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	0.388	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	0.391

Noisy data added:

Yoruba: BBC, VOA

Twí: JW300, Wikipedia

FastText embeddings: Spearman ρ correlation between human judgements and similarity scores on the wordSim-353 for the three datasets analysed (C1, C2 and C3). The comparison with massive fastText embeddings is shown in the top rows.

Automatic Diacritic Restoration for Yoruba

Training idea: Predict diacritics of a word based on its context using Seq2Seq models

source:

bi o tile je pe **egbeegberun** ti pada sile



ADR



target:

bí ó tilẹ̀ jẹ̀ pé **ẹgbẹẹgbẹ̀rún** ti padà sílé

Although **thousands** have returned home

Other Practical Solutions

1. Educating native speakers to write articles in their language e.g in Bambara
2. Standardization of language writing systems / dialects (e.g for Twi & Fula).
3. Involve native speakers in dataset collection to identify quality issues.
4. Data sheet recording quality issues should accompany dataset (Bender et al 2018).
5. Development of language identification models for African dialects/ languages.
6. Standardization of low-quality texts, e.g. adding diacritics to Yoruba text by *humans or machine learning model*.

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics (to appear) (2018)

Case Study: Yoruba Text quality verification



- Believed it will encourage people to release language data.
- A bit disturbed by the low-quality Yoruba data that may be submitted.
- Motivated a few people with the competition money.
- I joined the competition with two contributors of Yoruba Global Voices.
 - Global Voices is the only news website with high quality texts that I am aware of.
 - They verified the quality of the texts

Multi-domain Yoruba-English Parallel dataset

- In collaboration with Ìyá Yorùbá (Dámilólá Adébónòjọ) & Ọmọ Yorùbá
- Multi-domain sentences from news, proverbs, books, movie, and few sentences from technology, medicine and science terms.

Text	Number of sentences
Global Voices News	1,119
Yoruba Proverbs	2,700
Out of his mind (book)	862
Unsane movie transcript	817
Multi-domain sentences	549
Total	6047

Thanks for your attention

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA

WITH AN INTRODUCTION TO THE GIZ AI4D AFRICAN
LANGUAGE DATASET CHALLENGE

COFFEE BREAK



FAIR FORWARD
Artificial Intelligence for all.



IDRC | CRDI

International Development Research Centre
Centre de recherches pour le développement international

moz://a

Language data for African languages

Andrea Lösch
andrea.loesch@dfki.de
www.dfki.de

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA
WITH AN INTRODUCTION TO THE GIZ AI4D
AFRICAN LANGUAGE DATASET CHALLENGE

3 July 2020 from 14:00 to 16:00 pm CAT/CEST/UTC+2



  **FAIR FORWARD**  **IDRC | CRDI** 
International Development Research Center
Centre de recherches pour le développement international

*“Those who know nothing of
foreign languages know nothing
of their own.”*

Johann Wolfgang von Goethe (1749 – 1832)



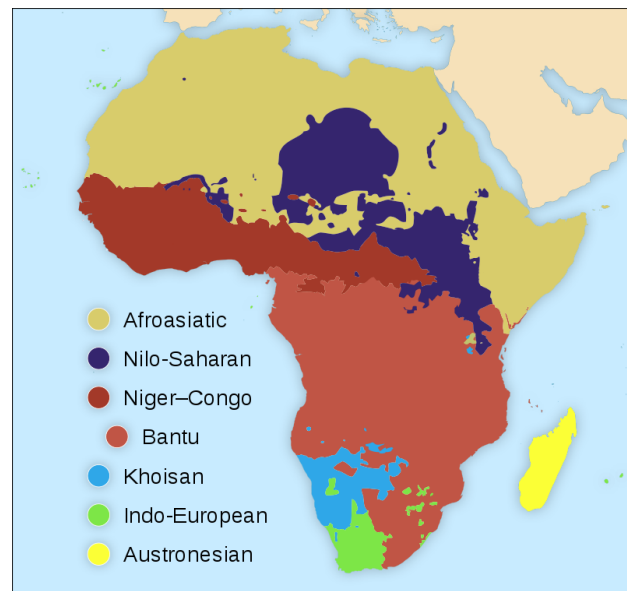
Introduction



- Language data as fundamental pre-requisite for MT and other LT
- DFKI is currently involved in the European Language Resource Coordination ([ELRC](#)) → Goal: make public services across Europe multilingual
- Support for 24 EU-official languages plus Norwegian and Icelandic
- A snapshot of resources across various domains: [ELRC-SHARE](#)
- Strong focus on under-resourced languages (Croatian, Maltese, Irish, Icelandic)

Languages in Africa

- 1.500 – 2.000 languages in Africa...
- Different language families...
- Arabic, Somali, Berber, Amharic, Oromo, Igbo, Swahili, Hausa, Manding, Fulani and Yoruba are spoken by tens of millions of people...



Source: https://en.wikipedia.org/wiki/Languages_of_Africa

A snapshot of data for African languages

- A first, non-exhaustive overview of language data and/or sources for African languages is available [here](#)
- Languages found:
 - Hausa
 - Igbo
 - Luganda
 - Luo
 - Northern Sotho
 - Setswana
 - Swahili
 - Twi
 - Xhosa
 - Xitsonga
 - Yorùbá
 - Zulu
- Central question: How and where to get data from?

Data collection approaches

- Important aspects of language data collection:
 - Identification of and collaboration with relevant language data holders
 - Identification and use of sources of language data
 - Making language data reusable!

Language data holders

- Identification of and collaboration with relevant language data holders
- Language data holders include any organisations and/or people that may create language data
- Examples of language data holders:
 - African translators and/or translation agencies (e.g. The South African Translators' Institute [SATI](#), a collection of South African translators is also available [here](#))
 - translation services in African national ministries, public services and/or governmental agencies (e.g. Language Unit of the Department of Cultural Affairs and Sport ([DCAS](#)) of Western Cape Government, South African Centre for Digital Language Resources ([SADiLaR](#)))

Language data holders

- Examples of language data holders (cont.):
 - African and/or international open data portals (e.g. [openAfrica](#)),
 - African language and/or language technology researchers and members of academia (e.g. [AfricArxiv](#), African Academy of Languages ([ACALAN](#))),
 - African and/or international language technology and language service providers (e.g. [Translate4Africa](#), [Folio Online](#))

Working with language data holders

- Retrieving language data directly from the relevant language data holders can be done in various ways, including both
 - face-to-face (e.g. through data collection workshops, focus group meetings, on-site assistance at the data holders' site) and
 - remote (e.g. through surveys among data holders or direct phone interviews).
- Surveys or phone interviews are always helpful for the identification of new data sets or for the identification of problems of the sharing of language data.
- It's a community-building effort!

Sources of language data

- Sources of language data can be any bi- or multilingual websites in the languages sought
- Examples:
 - governmental websites in the target countries/languages
 - websites of public services and academic institutions in the target countries/languages
 - websites of international, national or local organisations in the target countries/languages
- Web crawling to identify and retrieve mono-, bi- or multilingual language data from the Internet and to turn them into MT-ready language resources
➔ language profile

Making language data (re-)usable

- Two aspects:
 - Technical usability of language data
 - Legal usability of language data



Making language data (re-)usable

- Quick-check: Ensuring technical usability of language data
 - Is the format readable?
 - Is the source / are the sources copyrighted? (also see below, legal usability)
 - Have the source and target language(s) be identified correctly?
 - Is the alignment ok?
 - Are there any tokenization errors (no separator between words)?
 - Is the content machine-translated?

Making language data (re-)usable

- Quick-check: Ensuring legal usability of language data
 - Is the data protected by copyright?
 - If the data is protected by copyright can I identify the owner of the copyright or the author of the work?
 - Is the data available under a public license?
 - If no public license is clearly marked on the document, one should check the terms of use or if any documentation may help you determine the conditions of reuse of the material...

Data collection in and for Africa...

... voices and perspectives
from our experts:



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town

Discussion round: Data collection in and for Africa



Moderator: Andrea Lösch



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town

Portrait: Orevaoghene Ahia

- ❖ I am a research engineer at Instadeep.
- ❖ My research is in the area of NLP, Unsupervised Learning and Transfer Learning.
- ❖ I am very passionate about work that aims to improve the performance of complex NLP architectures on low-resourced languages.

Problems with Dataset Creation for African Languages

- ❖ Most african languages are mainly spoken.
- ❖ Some language datasets are hard to annotate for instance languages with diacritics.
- ❖ Data Statements; ethical concerns.
- ❖ Different domains, sometimes training NLP models from different domains hurt performance.
- ❖ Funding is unavailable.

Discussion round: Data collection in and for Africa



Moderator: Andrea Lösch



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town

GhanaNLP

PUTTING GHANAIAN LANGUAGES ON THE GLOBAL
MACHINE LEARNING MAP

Dr. Stephen Edward Moore



GhanaNLP

Open Source Movement of like-minded volunteers

Time and skills dedicated to building an ecosystem of :

1. Open-source datasets
2. Open source computational methods
3. An army of NLP researchers, scientists and practitioners

To revolutionize and improve every aspect of Ghanaian life through the powerful tool of Natural Language Processing.

Sponsors and Partners



Microsoft



Dataset Collection

- ❑ We set out a **google document** and requested for English to Twi translation.
- ❑ We used **Jehovah Witness (JW300)** translated English to Twi articles.
- ❑ We also used the **Christian Bible** as a main source of translation.
- ❑ Some **English to Twi books** on the market for tourists.

Challenges and Solutions

- ❑ Data contains non-twi words.
 - ❑ E.g. JW300 had over 600K sentences.
 - ❑ Several non-twi words.
 - ❑ Finally ~22K uniques words
- ❑ Involving **Ghana Linguistic Association**
- ❑ University Researchers and Professors reviewing some sentences.
- ❑ Several translations of same sentences from **google document**.

Dataset Sharing

- ❑ All our datasets are **open source**.
- ❑ Most data collected is volunteered.
- ❑ Other linguist and researchers are willing to share their data.
- ❑ Exchange is no robbery!
- ❑ For accelerated development in language development, **non-restrictive open access data exchange** is the way forward!

THANK YOU



<https://ghananlp.github.io/>



natural.language.processing.gh@gmail.com

Discussion round: Data collection in and for Africa



Moderator: Andrea Lösch



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town

AI4D Africa Webinar Series

Making NLP Work In Africa

A legal perspective on sharing language data in Africa

Webinar

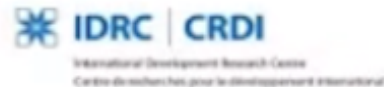
03 July 2020

Dr. Tobias Schonwetter

A/Prof, Department of Commercial Law (UCT, South Africa)

Director: UCT IP Unit

Founding Director: iNtaka Law Tech Centre (UCT)





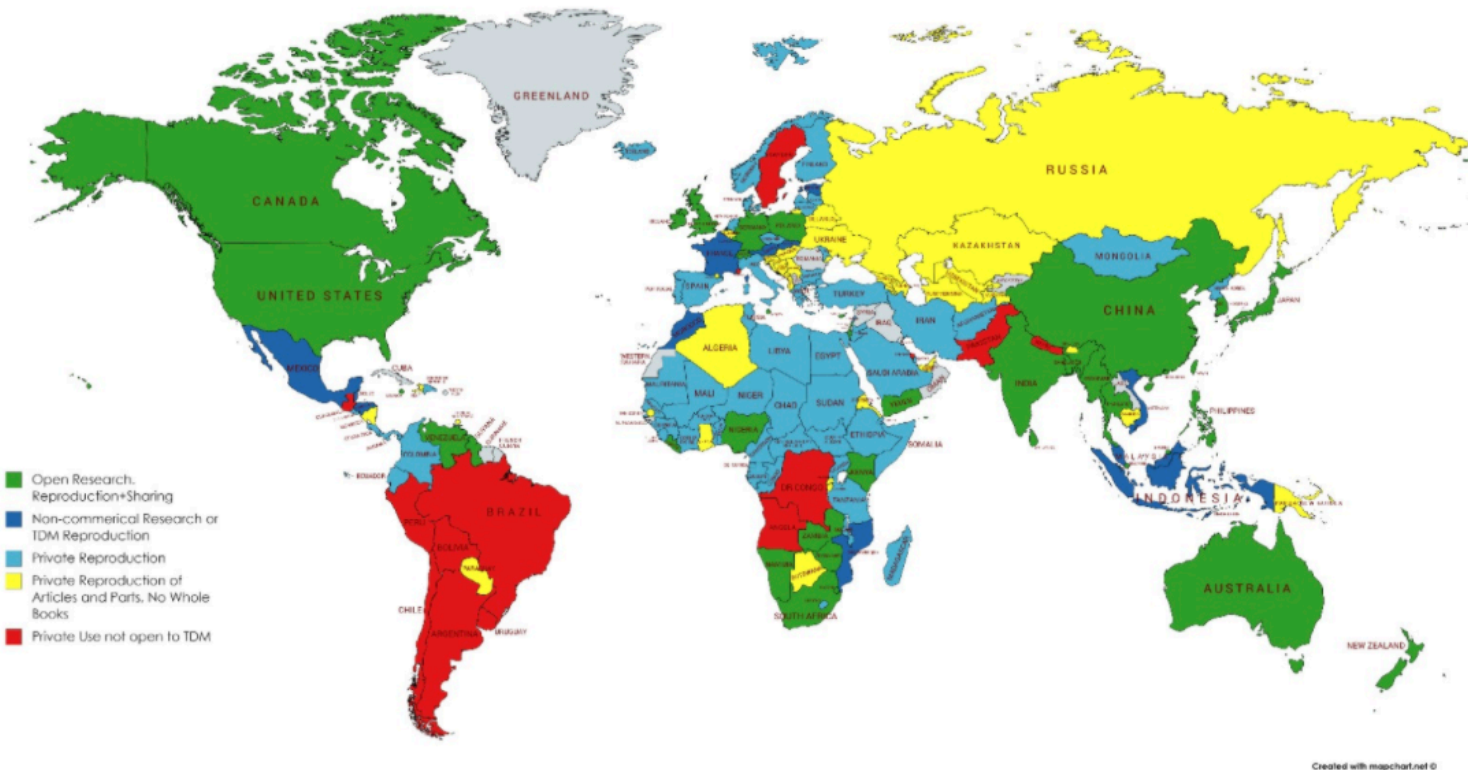
=





What is allowed in one country in Africa may or may not be allowed in other countries in Africa – and **it all depends on what domestic laws say.**

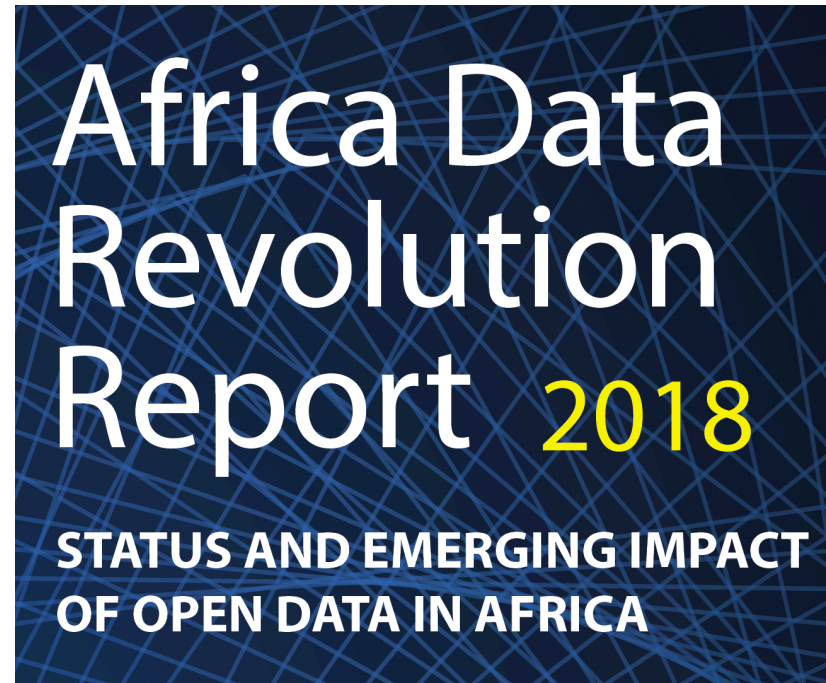
We need to **better map what is and what isn't allowed** in the different countries in Africa.



Created with mapchart.net ©

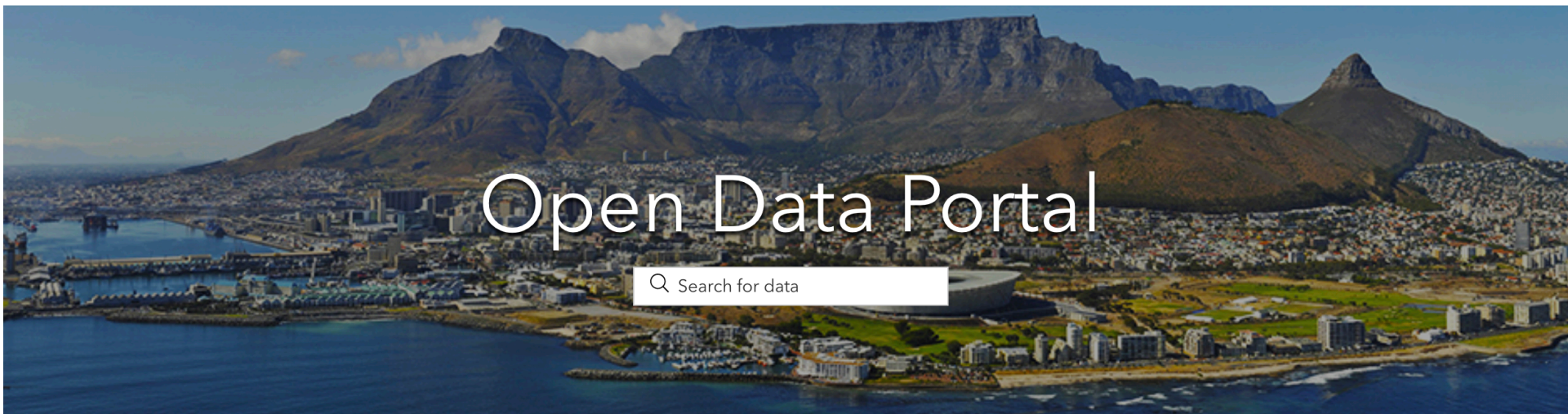
- Just because something is on the Internet or otherwise easily **accessible doesn't mean that it can freely be used**
- A lot of (but not all!) (language) data will, for instance, be **protected by domestic © laws**
 - **Permission** needs to be sought– unless a **© exception** applies (eg for research or educational purposes or through specific ***text and data mining exceptions***) or because permission is granted upfront (eg when the data is disseminated under an **open licence** such as CC)
- And even if © is not or no longer a stumbling block for access, the **collection, use and storage may still be subject to other laws**, including **data protection / privacy laws**, (like here in SA the Protection of Personal Information (POPI) Act which more or less fully commenced earlier this week.)

And...



"[Broadly,] Open data is not yet entrenched in law in the continent and the legal frameworks supporting it are either incomplete or directly absent. Implementation and resourcing are also very weak."

<http://webfoundation.org/docs/2019/03/Africa-data-revolution-report.pdf>



- Using public sector / government data may be still be the lowest hanging fruit. Several governments in Africa have now started to share their data (more) freely – as a result of Open Science or Open Government Data policy initiatives, or because newer laws actually prescribe this..
- And an increasing amount of private sector players have also begun to make their data openly available.

- ✓ **Map** legal frameworks re open data / data use
- ✓ **Harmonise** legal frameworks and **update** copyright/ data protection / right to information laws in Africa to facilitate data use
- ✓ Hold governments to their open data commitments / policies and remind them that **data produced with tax payers money should belong to the public**
- ✓ **Make your own data findable, accessible and usable**, especially through applying standard open licences
 - ✓ **Use data that is clearly marked as open**
 - ✓ Read the **Africa Data Revolution Report** 😊

Thank you!

tobias.schonwetter@uct.ac.za

www.ip-unit.org

www.lawtechlab.africa

@lawtechlab

@tobyschonwetter

@AfricanIP



Discussion round: Data collection in and for Africa



Moderator: Andrea Lösch



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town

GIZ AI4D AFRICAN LANGUAGE DATASET CHALLENGE

<https://zindi.africa/competitions/ai4d-african-language-dataset-challenge>

THANK YOU!

