
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

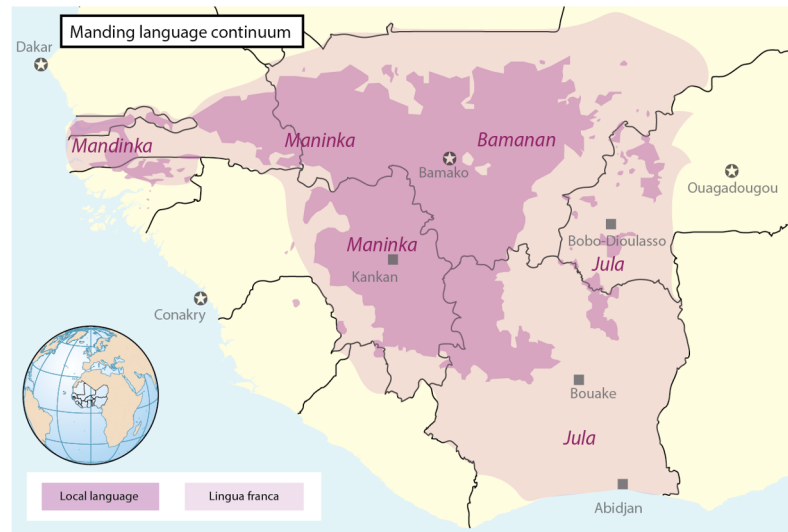
10 Most resourced African Languages

- 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

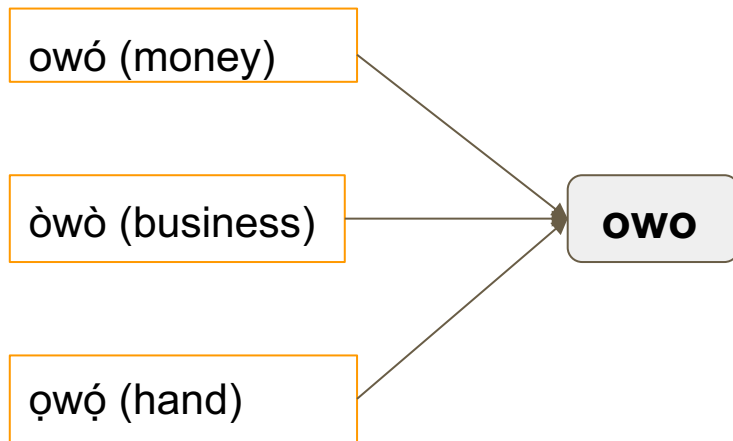
- Twi (Asante & Akuapem dialects) e.g in JW300
 - Asante: Me papa **fi**e yɛ **fɛ**.
 - Akuapem: Me papa **fi** yɛ **fɛw**.
 - English: My dad's house looks nice.
- Rwanda-Rundi e.g in VOA / BBC.
 - Kinyarwanda: Uriya **mugore yaberewe**.
 - Kirundi: Uriya **mukenyezi yarimvye**.
 - English: That woman 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.

Jesujoba O Alabi, Kwabena Amponsah-Kaakyire, David I Adelani, and Cristina España-Bonet. Massive vs. curated word embeddings for low-resourced languages. the case of Yorùbá and Twí. In LREC, 2020.

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