# FFR Dataset: Documentation

**Bonaventure .F. Dossou** * **Chris .C. Emezue** * **Frejus Layele** **

* *Kazan Federal University, Kazan Russian Federation*
*e-mail: (chris.emezue, femipancrace.dossou)@gmail.com*
** *Laboratoire d'Informatique Signal et Image de la Cote d'Opale*
*e-mail:frejus.layele@gmail.com*

## 1. INTRODUCTION

FFR Dataset is an ongoing project to collect, clean and store corpora of Fon-French sentences for machine translation from Fon-French. Fon (also called Fongbe) is an African-indigenous language spoken mostly in Benin, by about 1.7million people[7]. As training data is crucial to the high performance of a machine learning model, the aim of this project is to create the largest set of training corpora for the research and design of translation and NLP models involving Fon. We have generated 117,029 parallel Fon-French sentences at the moment.

Chris Emezue, Bonaventure Dossou and Dr Frejus Layele collaborated on compiling the FFR Dataset. Dr Frejus Layele provided linguistic insight from his various researches on the Fon dialect, which was paramount to understanding how to clean the data, as well as supplying the initial dataset for analysis.

## 2. COMPOSITION

Each instance of the FFR dataset consists of a sentence in Fon and its corresponding translation in French, separated by a 'tab'. The target for each sentence is its translation in French.

The major sources of the creation of FFR Dataset are:
JW300 - *http://opus.nlpl.eu/JW300.php*
BeninLanguages - *https://beninlangues.com/*

Table 1. Contribution of websites to FFR Dataset

| Source | JW300 | BeninLanguages |
|---|---|---|
| % Sentences obtained | 20% | 80% |

JW300 is a parallel corpus of over 300 languages with around 100 thousand parallel sentences per language pair on average[2].

BeninLanguages contains (in French and Fon):

- vocabulary words
- Short expressions
- Small sentences
- Complex sentences
- proverbs
- Bible verses: Genesis 1 - Psalm 79

A week of preprocessing and cleaning techniques like removing stopwords,weblinks and wrong translations, tokenization was carried out.

The FFR Dataset, at the moment, contains 117,029 processed French-Fon sentences.

## 3. USES AND DISTRIBUTION

The FFR Dataset is open-sourced for research purposes.

*Please endeavour to reference the authors when you use this dataset.*

## 4. MAINTENANCE

The dataset will be updated once new corpora have been found and processed.

Contributions from external sources are strongly supported. Please contact any of the authors for dataset contribution requests.

## REFERENCES

[1] Africa NLP Workshop 2020. *URL: https://africanlp-workshop.github.io/cfp.html*
[2] JW300 Datasets provided by OPUS. *URL: http://opus.nlpl.eu/JW300.php*
[3] Benin languages .
[4] AI4D-Africa *URL:https://ai4d.ai/*
[5] Masakhane Machine Translation Community. *URL:https://sites.google.com/view/masakhane/home*
[6] Black in AI *URL:https://twitter.com/black_in_ai*
[7] Fon Language- Wikipedia *URL:https://en.wikipedia.org/wiki/Fon_language*
[8] Frejus Layele (PhD) - *URL: https://scholar.google.com/citations?user=rbDnH6MAAAAJ&hl=fr*