Data statements for the FFR Dataset

Publication & Citation: «FFR V1.0: Fon-French Neural Machine Translation»
Data set developers: Bonaventure Dossou & Chris Emezue
Data statement authors: Bonaventure Dossou[1] & Chris Emezue[2]
Other Contributors: Antske Fokkens[3], Zeerak Waseem

Fon is a very low-resourced language. It belongs to the group of Bantu languages spoken in Togo, Nigeria, and principally in Benin. This language is very tonal, and exists in two different forms: pure and simplified. Both forms differ from each other mostly in speaking, highly influenced by diacritics, but sometimes in writings as well. People that speak the pure fon are in majority native of Abomey, a city in the center of Benin, or have been living there for a long time. The simplified fon, however, is usually spoken by people who either do not have Fon as their first language, or who grew up in a different environment. A good example is the data science and artificial intelligence student, Bonaventure Dossou. Even though Fon is not the official language in Benin the majority of inhabitants, from north to south and east to west speak the language. Their origin, gender, age, or even socioeconomic status do not really matter. Our current FFR Dataset comes from a diverse range of genders, age, as well as socioeconomic statuses, and geographic locations.

The fact that Fon is mostly spoken than written made us include every relevant textual data found online in the dataset created under the name of **FFR (Fon-French) Dataset** which contains currently 53975 distinct pairs of Fon-French sentences. The initial dataset was compiled from the JW300 dataset (which is a relatively large dataset on biblical data, albeit with a lot of noise) (around 10%) and from Benin Langues Website (around 90%), providing daily sentences, proverbs, common expressions, and short words. The current FFR Dataset contains in balanced proportions, both forms of Fon: the pure one being mostly from the biblical data while the simplified Fon is from the other sources mentioned above and can be found on our Github page: FFR Dataset, with the correct documentation.

FFR Dataset is currently being updated with new data obtained with the collaboration of Fon and French native speakers, owning different initiatives to promote in their

[1] Email: femipancrace.dossou@gmail.com
[2] Email: chris.emezue@gmail.com
[3] Email: antske@gmail.com

way the Fon language. One of them created and monitors a Facebook Robot that shares content (part of the body, daily expression, jokes) in Fon, while the other one is the author of the label IamYourClounon, providing and promoting citations and daily sentences in Fon using pictures (an example provided below). It is important to highlight that the data are not directly retrieved from the Robot or the pictures but instead from the content creators themselves.

The data in our dataset are both conversational and non-conversational. The main domains are the bible and the daily dialogue/conversations. The topics cover various fields, from the creation of the heavens and earth (Bible) to the daily plans (occupations, activities, health, well-being, proverbs, short and common expressions). Any new raw data is pre-processed and put in the right format by the authors of the dataset.

The collaboration with the two content creators, Fon-French native speakers mentioned above, also led to the creation of a Google Form shared on Facebook, to collect more data from the users of their platforms in particular and from anybody understanding and speaking Fon in general. The idea is to ask them to write down through the mentioned form, 1-3 daily, fon sentences they would like to see translated in French. The format required ***was sentence_in_fon = french_translation***. We made the form anonymous so that people don't feel obliged to provide information like names, emails, age, and genders. However, based on the ones who provided those information we can come up with 45% women and 55% men across the participants.

So far, 25 people have filled the form and we totalize 60 responses (distinct or not: preprocessing not done yet).

The data is currently using the MIT license and Licence Creative Commons Attribution - Non Commercial Use - Sharing under the Same Conditions 4.0 International. Our dataset, while being the first of its kind (in terms of topics diversity and data domains covered), contains 10% of translations from the existing JW300 dataset. We are planning to restrict (or drop) it in order to avoid too much noise in the dataset but also to balance the topics' distribution, avoiding biased neural machine translation models.

The FFR Dataset and FFRv1.0, the current Neural Machine Translation model, have been presented at the AfricaNLP Workshop of the International Conference in Learning Representations (ICLR) 2020. Our presentation can be found here.