Data name	Source	Language of source	No. Sentences
Jehovah Witness News	jw.org/yo/iroyin	en-yo	3,508
Voice of Nigeria News	von.gov.ng	en-yo	3,048
TED talks	ted.com/talks	en	2,945
Global Voices News	yo.globalvoices.org	en-yo	2,932
Proverbs	twitter.com/yoruba_proverbs	yo-en	2,700
Out of His Mind Book	Obtained from the author	en	2,014
Software localization	Obtained from Professional Translators	en	941
Movie Transcript ("Unsane")	youtu.be/hdWP0X5msZQ	yo-en	774
Short texts	Obtained from Professional Translators	en	687
Radio Broadcast	Transcript from Bond FM 92.9 Radio	en	258
Creative Commons License	Obtained from Professional Translators	en	193
UDHR Translation	ohchr.org	en-yo	100
Total			20,100

Table 5: Dataset collection sources with source language(s) and the number of sentences contained.

A Appendix

A.1 Dataset Collection for MENYO-20k

Table 5 summarizes the texts collected, their source, the original language of the texts and the number of sentences from each source. We collected both parallel corpora freely available on the web and monolingual corpora we are interested in translating (e.g. the TED talks) to build the MENYO-20k corpus. Some few sentences were donated by professional translators such as "short texts" in Table 5. We provide more specific description of the data sources below.

Jehovah Witness News We collected only parallel *"newsroom"* (or *"Ìròyìn"* in Yorùbá) articles from *JW.org* website to gather texts that are not in the religious domain. As shown in Table 5, we collected 3,508 sentences from their website, and we manually confirmed that the sentences are not in JW300. The content of the news mostly reports persecutions of Jehovah witness members around the world, and may sometimes contain Bible verses to encourage believers.

Voice of Nigerian News We extracted parallel texts from the VON website, a Nigerian Government news website that supports seven languages with wide audience in the country (Arabic, English, Fulfulde, French, Hausa, Igbo, and Yorùbá). Despite the large availability of texts, the quality of Yorùbá texts is very poor, one can see several issues with orthography and diacritics. We asked translators and other native speakers to verify and correct each sentence.

Global Voices News We obtained parallel sentences from the Global Voices website[14] contributed by journalists, writers and volunteers. The website supports over 50 languages, with contents mostly translated from English, French, Portuguese or Spanish.

TED Talks Transcripts We selected 28 English TED talks transcripts mostly covering issues around Africa like health, gender equality, corruption, wildlife, and social media e.g "How young Africans found a voice on Twitter" (see the Table 6 for the selected TED talk titles). The articles were translated by a professional translator and verified by another one.

[14] https://globalvoices.org

	Title	Topic
1	Reducing corruption takes a specific kind of investment	Politics
2	How young Africans found a voice on Twitter	Technology
3	Mothers helping mothers fight HIV	Health
4	How women are revolutionizing Rwanda	Gender-equality
5	How community-led conservation can save wildlife	Wildlife
6	How cancer cells communicate - and how we can slow them down	Health
7	You may be accidentally investing in cigarette companies	Health
8	How deepfakes undermine truth and threaten democracy	Politics
9	What tech companies know about your kids	Technology
10	Facebook's role in Brexit - and the threat to democracy	Politics
11	How we can make energy more affordable for low-income families	Energy
12	Can we stop climate change by removing CO2 from the air?	Climate
13	A comprehensive, neighborhood-based response to COVID-19	Health
14	Why civilians suffer more once a war is over	Human Rights
15	Lessons from the 1918 flu	Health
16	Refugees have the right to be protected	Human Rights
17	The beautiful future of solar power	Energy
18	How bees can keep the peace between elephants and humans	Wildlife
19	Will automation take away all our jobs?	Technology
20	A celebration of natural hair	Beauty
21	Your fingerprints reveal more than you think	Technology
22	Our immigration conversation is broken - here's how to have a better one	Politics
23	What I learned about freedom after escaping North Korea	Politics
24	Medical tech designed to meet Africa's needs	Health
25	What's missing from the American immigrant narrative	Education
26	A hospital tour in Nigeria	Health
27	How fake news does real harm	Politics
28	How we can stop Africa's scientific brain drain	Education

Table 6: TED talks titles.

Proverbs Yorùbá has many proverbs and culturally referred to words of wisdom that are often referenced by elderly people. We obtained 2,700 sentences of parallel yo–en texts from Twitter.[15]

Book With permission from the author (Bayo Adebowale) of the "Out of His Mind" book, originally published in English, we translated the entire book to Yorùbá and verified the diacritics.

Software Localization Texts (Digital) We obtained translations of some software documentations such as Kolibri[16] from past projects of professional translators. These texts include highly technical terms.

Movie Transcripts We obtained the translation of a Nigerian movie "Unsane" on YouTube from the past project of a professional translator. The language of the movie is Yorùbá and English, with transcription also provided in English.

Other Short Texts Other short texts like UDHR, Creative Commons License, radio transcripts, and texts were obtained from professional translators and online sources. Table 1 summarizes the number of sentences obtained from each source.

[15] Also available in `https://github.com/Niger-Volta-LTI/yoruba-text`

[16] `https://learningequality.org/kolibri`