Koumankan: A Scalable And Cost Efficient Way To Extend Common Voice For Dyula And Other African Languages

Ismaël Koné
Department of Artificial Intelligence
Virtual University of Cote d'Ivoire
Abidjan, Cote d'Ivoire
ismael21.kone@uvci.edu.ci

Monsia Dougban
data354
Abidjan, Côte d'Ivoire
monsia.dougban@data354.co

Fabrice Zapfack
data354
Abidjan, Côte d'Ivoire
fabrice.zapfack@data354.co

Abstract

The field of automatic processing of African languages has witnessed significant progress in recent years, thanks to the efforts of researchers and communities such as Masakhane. This progress has resulted in the creation of various machine learning models, including machine translation, automatic speech recognition, and named entity recognition models. These tools have the potential to facilitate technological, social, and financial inclusion by eliminating language barriers. However, it's really important to remember that without rigorously collected, inclusive data, it's impossible to move research forward. We therefore turned our attention to data collection for low-resource languages. That's why, in this paper, we present the Koumankan project, which proposes a scalable and cost-efficient method of extending the CommonVoice dataset for Dyula and other African languages. We discuss our approach and provide an update on the current state of construction of this dataset. The project aims to improve the quality and quantity of speech data available for African languages and promote the development of speech recognition models, machine translation models for these languages. The scalability and cost-effectiveness of our approach make it suitable for gathering large amounts of speech data in a relatively short period. It should be noted that upon completion of this project, all data will be made available to the public.

1 Context and Motivation

The world is witnessing a profound transformation thanks to technological advances. We feel this wave most keenly through mobile devices, which have a deep penetration in Africa. Increasingly, interaction with these devices is facilitated by voice-based systems. Enormous progress is being made in the design of these voice-based systems and their use for Western languages. However, this considerable progress has not had the same impact in the African context. Indeed, the majority of existing applications do not support several African languages. Language remains one of the major barriers to digital and technological inclusion for African populations. Reducing these barriers remains crucial to the development of many African business sectors. Consequently, the development

5th Deep Learning Indaba Conference (DLI 2023).

of automatic speech recognition (ASR) technologies for our languages can bridge this gap and enable more inclusive technological progress that benefits everyone. On the other hand, the lack of voluminous data useful for training such a system is a major obstacle to be resolved. This is why, in the Koumankan project, we are proposing a scalable and cost-efficient method of extending the CommonVoice dataset for Dyula and other African languages. For this first phase, we will focus on the collection of audio data in Dyula. Dyula is a language spoken in several West African countries, notably Mali, Senegal, Burkina Faso and Côte d'Ivoire. The total number of speakers is around 16.4 million[1] . Due to the scarcity of available numerical data, this language is considered to be poorly endowed.

2 Our Contribution

Our aim is to build a large corpus of Dyula audio and transcriptions, with the corresponding French and English translations, which will be used as a benchmark database for the design of machine learning models. This database will be used to solve the tasks of automatic translation between Dyula and French or English, automatic recognition of Dyula speech, automatic translation of speech and also direct audio-to-audio translation between Dyula and French. To achieve this goal, we use CoVost 2[2] and commonvoice v4[1] to extract a French audio and transcription database with English translation. We then create the audio and Dyula transcription of the French texts. We use a cost-effective method of collecting large quantities of audio data in a relatively short space of time. We also propose an adapted web application to facilitate audio collection.

3 Methodology and Results

3.1 Methodology

As part of the Koumankan project, we are building a database of audio recordings and transcriptions in Dyula, as well as French and English translations of the texts. To achieve this objective, our methodological approach is based on three main stages.

Extraction step

The first step is to use the CoVost 2 and CommonVoice v4 resources to extract an audio database and transcriptions in French, accompanied by their English translations[2]. These pre-existing resources offer a variety of French phrases and their English equivalents, providing a solid foundation for our project. Thanks to this initial step, we obtained a set of 236,894 pairs of French-English sentences.

Audio collection step

The second stage is dedicated to collecting audio recordings in Dyula corresponding to the French phrases in our corpus. We chose a participatory approach, recruiting several collectors from families fluent in Dyula. These collectors came mainly from Burkina Faso, but also from Côte d'Ivoire. They were asked to record their relatives, family members or community members fluent in Dyula, in order to obtain audio recordings in Dyula corresponding to the French sentences in our corpus. This stage was carried out over a period of 01 months and resulted in the collection of 12,734 audio recordings (15h 10m 32s) in Dyula. To facilitate collaborative data collection, we have created an ergonomic, easy-to-use and responsive web application 1 for mobile devices[3]. Once the collector logs in, it records the speaker's information (*age*, *gender*, *country*, *city*). It then records the speaker repenting in Dyula the sentence displayed on the application and moves on to the next sentence.

Audio transcription step

Finally, the third stage involved transcribing the audio recordings into Dyula. We hired three skilled linguists to validate and transcribe the collected recordings. Thanks to their expertise, we have so far obtained a set of 3,817 audio transcriptions (3h 57m 58s) in Dyula. This step is essential to guarantee

[1]https://www.worlddata.info/languages/bambara.php - consulted on June 04, 2023.

[2]https://github.com/facebookresearch/covost - building the covost dataset.

[3]https://audioset-app-m3okm5mmqa-od.a.run.app/

the quality of the Dyula data in our database. In order to facilitate the linguist's work during the transcription phase, we have also developed a user-friendly web application[4]2. This interface displays the audio in Dyula to be transcribed, followed by the sentence in French to enable verification and a better understanding of the context for the translation. This approach enables verification on two levels: the quality of the audio and the accuracy of the translation.

By combining these three steps, we were able to build a rich and diversified database containing audios from speakers of different genders, ranging in age from 15 to 70, living in Burkina Faso and Côte d'Ivoire

3.2 Results

Although the project is ongoing, the tables below show the current composition of the database.

Text	Dyula	French	English
Lines	3,851	32,184	32,184
Tokens	22,322	2,124,128	2,102,393
Avg. line token	19	47	44

Table 1: Basic statistics of koumankan text set.

Dyula Audio	Time
Total length of recordings	15h 10m 32s
Avg. length of recordings	4s
Female length of recordings	7h 52m 57s
Male length of recordings	7h 17m 34s
Human verified instances	All
Recording format	wav', 44 kHz or 48 kHz

Table 2: Basic statistics of koumankan audio set.

Speakers	Counts
Number of participant	249
Number of female	120
Number of male	129
Avg. number of lines per participant	52

Table 3: Basic statistics of recorded speakers.

4 Future Work

The future work we are planning is motivated by the need to fill certain gaps in the existing database and to ensure a more accurate and complete representation of the Dyula language and culture. Indeed, the CoVoST sentences based on commonvoice v4 from sources such as wikipedia in French, address directories, parliamentary session reports from the national assembly and certain books are of a fairly high language level, even uncommon for Dyula-speaking communities. The aim of further work will be to take into account factors such as:

- **Cultural diversity**: It's important to collect audio conversations between native Dyula speakers on various topics related to their culture, traditions and history. This will help capture idiomatic expressions, specific linguistic nuances and cultural references important to the Dyula language.
- **Global news**: It's essential to include current events, such as the Covid pandemic, the war in Ukraine and others, to reflect the evolution of society and enable natural language processing applications to address contemporary issues.

[4]https://transcriptor.koumankan.com/

- **Reducing bias**: By broadening thematic coverage and collecting authentic conversations between native Dyula speakers, we help to reduce potential bias in linguistic data. Pre-existing sources may be influenced by cultural preferences, stereotypes or geographical boundaries. By actively collecting audio recordings on a variety of topics and contexts, we enable a more balanced and faithful representation of the Dyula language.

By undertaking this future work, we aim to improve the quality and diversity of the Dyula database, which will have a positive impact on automatic natural language processing applications by promoting a more accurate and culturally adapted understanding of this language. In addition, these efforts will contribute to the preservation and promotion of Dyula culture in a globalized context.

References

[1] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[2] Changhan Wang, Juan Miguel Pino, Anne Wu, and Jiatao Gu. Covost: A diverse multilingual speech-to-text translation corpus. *CoRR*, abs/2002.01320, 2020.

A Appendix

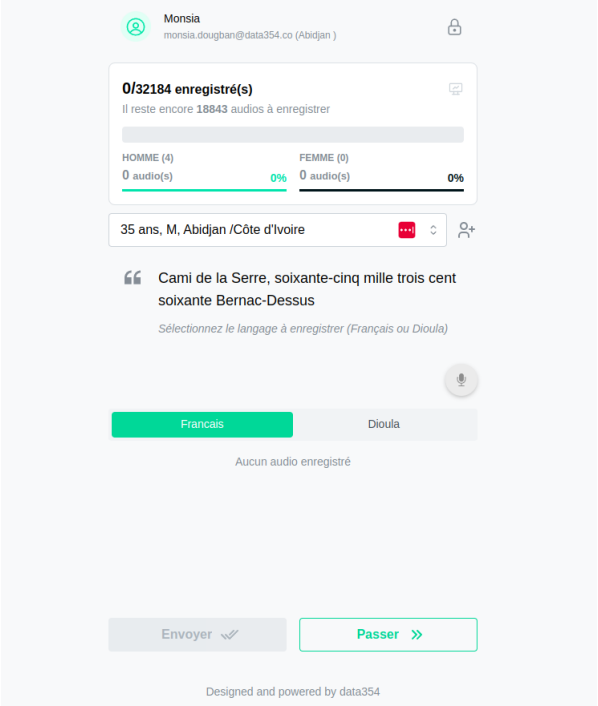

Figure 1: Screenshot of responsive web app for audio collections.

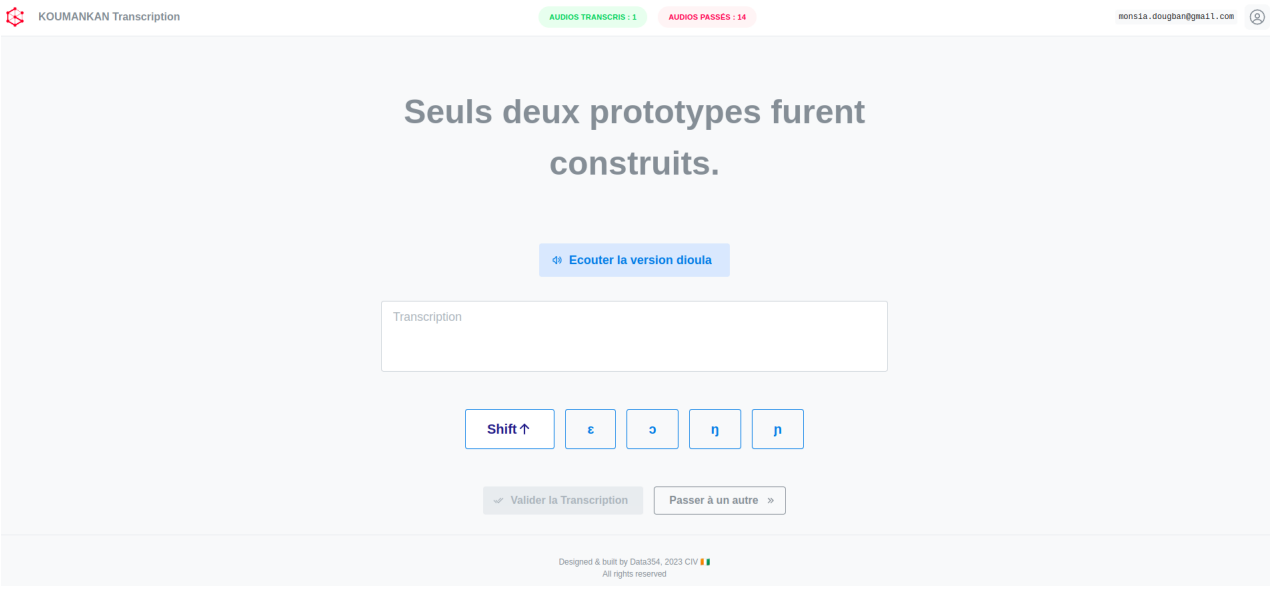

Figure 2: Screenshot of web app for audio transcriptions.