Yankari: Monolingual Yoruba Dataset

Maro Akpobi
African Center For Language Preservation
maro@acflp.org

October 14, 2024

Abstract

This paper presents Yankari, a large-scale monolingual dataset for the Yoruba language, aimed at addressing the critical gap in Natural Language Processing (NLP) resources for this important West African language. Despite being spoken by over 30 million people, Yoruba has been severely underrepresented in NLP research and applications. We detail our methodology for creating this dataset, which includes careful source selection, automated quality control, and rigorous data cleaning processes. The Yankari dataset comprises 51,407 documents from 13 diverse sources, totaling over 30 million tokens. Our approach focuses on ethical data collection practices, avoiding problematic sources and addressing issues prevalent in existing datasets. We provide thorough automated evaluations of the dataset, demonstrating its quality compared to existing resources. The Yankari dataset represents a significant advancement in Yoruba language resources, providing a foundation for developing more accurate NLP models, supporting comparative linguistic studies, and contributing to the digital accessibility of the Yoruba language.

1 Introduction

Natural Language Processing (NLP) has made tremendous strides in recent years, yet these advancements have primarily benefited high-resource languages, leaving many African languages, including Yoruba, underrepresented in NLP research and applications. This paper introduces Yankari, a large-scale, high-quality monolingual dataset for Yoruba, a language spoken by over 30 million people in West Africa. Despite its significant speaker population, Yoruba has long suffered from a lack of comprehensive, ethically-sourced language resources suitable for modern NLP tasks. Yankari addresses this critical gap by providing a carefully curated corpus of 51,407 documents from 13 diverse sources, totaling over 30 million tokens. Our methodology prioritizes ethical data collection, rigorous quality control, and the preservation of linguistic authenticity. By avoiding problematic sources such as religious texts and machine-translated content, Yankari offers a more balanced and representative sample of contemporary Yoruba language use.

This paper details our data collection and processing pipeline, discusses the challenges encountered in creating resources for low-resource languages, and provides a thorough analysis of the dataset's composition and potential biases. We also offer insights into the ethical considerations surrounding the creation and use of such resources. Through Yankari, we aim to facilitate the development of more accurate and culturally appropriate NLP models for Yoruba, contribute to the preservation of linguistic diversity in the digital age, and provide a replicable approach for creating high-quality datasets for other low-resource languages. To our knowledge, Yankari represents the first large-scale, non-religious domain monolingual resource created specifically for Yoruba. This work not only provides a valuable asset for Yoruba NLP but also offers a replicable approach for developing similar datasets for other low-resource languages.

2 Related Works

This section examines existing resources for Yoruba NLP, highlighting their limitations and the need for a comprehensive, ethically-sourced Yoruba dataset.

2.1 Monolingual Yoruba Datasets

2.1.1 Yorùbá Text C3

Alabi et al. (2020) introduced Yorùbá Text C3, compiled from various web sources including the Bible, JW300 (Agić and Vulić, 2019), books, news articles, and Wikipedia. While broad in scope, this dataset is heavily skewed towards religious content, particularly Christianity. This bias significantly limits its utility for NLG tasks requiring balanced and diverse text. Moreover, the inclusion of JW300 data raises serious ethical and legal concerns. Hutchinson (2024) points out that the Jehovah's Witnesses have explicitly prohibited the use of their data in NLP research, making the continued use of JW300 not just ethically questionable but potentially illegal.

2.1.2 MENYO-20k

Adelani et al. (2021) developed MENYO-20k, a multi-domain English-Yoruba corpus primarily for machine translation tasks. While it offers more diverse content, its relatively small size of 20,100 sentences and focus on translation rather than monolingual text generation limit its applicability for large-scale NLG tasks.

2.2 Multilingual Datasets Including Yoruba

2.2.1 Wura Dataset

The Wura dataset, developed by Oladipo et al. (2023), is a multilingual dataset containing approximately 68,000 Yoruba documents. It integrates content from JW300 and Wikipedia, inheriting similar biases and ethical issues as Yorùbá Text C3. Our manual inspection of the Wura dataset revealed critical quality issues not previously reported, including formatting errors and inappropriate content.

2.2.2 Large-Scale Web-Crawled Corpora

Multilingual datasets such as mC4, OSCAR, and the Afriberta-Corpus also include Yoruba content. However, these datasets often suffer from noise, poor formatting, and limited source diversity. Ogueji et al. (2021) used the Afriberta-Corpus, which primarily sources data from the BBC News and Common Crawl, resulting in a dataset lacking the domain diversity necessary for robust NLG.

Nguyen et al. (2023) introduced CulturaX, covering 167 languages. However, its Yoruba subset showed an alarmingly high duplication rate of 24.48% and contained machine-translated pages, false positives in language detection, and a significant amount of religious text.

2.3 Large-Scale Multilingual Efforts for African Languages

The Cheetah project by Adebara et al. (2024) focuses on natural language generation for 517 African languages, including Yoruba. This ambitious project utilizes existing corpora from various sources, including OSCAR (Ortiz Suárez et al., 2020), CC-100 (Conneau et al., 2020), Afriberta-Corpus (Ogueji et al., 2021), and mC4 (Xue et al., 2021).

While Cheetah represents a significant step towards improving NLP capabilities for African languages, it faces challenges common to large-scale multilingual efforts. These include potential issues with data quality, bias towards certain domains or text types, and the inclusion of machine-translated content.

2.4 Ethical Considerations in Using Religious Texts

The ethical implications of using religious texts in NLP, particularly for low-resource languages, are profound and often overlooked. Hutchinson (2024) challenges the NLP community's casual approach to sacred texts, arguing that the prevalence of Christian texts in datasets for low-resource languages reflects a legacy of colonialism and missionary work. This creates an ethical dilemma where NLP technologies risk becoming unwitting agents of cultural imperialism and religious proselytism.

2.5 The Need for Yankari

Given these severe limitations and ethical concerns in existing resources, there is an urgent need for a high-quality, diverse, and extensive monolingual Yoruba dataset that does not rely on restricted or problematic sources. This is where **Yankari** comes in, directly addressing the gaps and ethical issues left by previous datasets.

Yankari aims to provide a large-scale, ethically sourced corpus that represents a wide range of Yoruba language use. By avoiding the pitfalls of previous datasets, such as over-reliance on religious texts or machine-translated content, Yankari seeks to offer a more balanced and authentic representation of the Yoruba language. This approach aligns with recent calls in the NLP community for more thoughtful and ethical dataset creation, particularly for low-resource languages

3 Methodology

This section details our approach to creating the Yankari dataset, including data collection, processing, quality assurance steps, and corpus analysis.

3.1 Data Collection

Our data collection process focused on gathering content from diverse, high-quality sources to ensure a representative sample of contemporary Yoruba language use. We carefully selected 13 sources, including news outlets, blogs, educational websites, and Wikipedia. Table 1 provides an overview of these sources and their contributions to the dataset.

3.2 Analysis of Existing Datasets

To inform our data collection and curation process, we conducted a detailed analysis of the Wura dataset (Oladipo et al., 2023). Our investigation revealed several critical issues:

- High repetition: 18.01% of the dataset contains the word 'asteroidi' (asteroid), indicating a significant bias towards astronomical content.

- Duplication: After removing duplicates and cleaning, only 17,103 unique entries remained, representing just 45% of the original dataset.

- Quality issues: We found formatting errors, inappropriate content, and entries in non-Yoruba languages.

These findings highlight the pressing need for more stringent data curation practices in low-resource language datasets. The recent study by Hernandez et al. (2022)on scaling laws and the interpretability of learning from repeated data offers valuable insights that influenced our methodology. Their extensive research reveals that data repetition can severely hinder model performance, particularly by disrupting the balance between memorization and generalization. Additionally, repeated data can obstruct the development of "induction heads," which are vital for in-context learning in large language models. Most importantly, their work underscores the pivotal role that high-quality training data plays in the effectiveness of language models. These insights significantly shaped our careful approach to developing Yankari, underscoring the critical need for rigorous data curation, quality control, and the preservation of diversity in our dataset.

Source	Docs	Type	Domain
yo.wikipedia.org	16,809	Encyclopedia	General
alaroye.org	10,535	News	Current Affairs
www.bbc.com	8,252	News	Current Affairs
www.awikonko.com.ng	5,438	Blog	Culture
yoruba.von.gov.ng	2,542	News	Current Affairs
sportsinyoruba.wordpress.com	2,328	Blog	Sports
www.asejere.net	2,079	Blog	Entertainment
asa.ooduarere.com	1,744	Blog	Culture
radionigeriaibadan.gov.ng	824	News	Current Affairs
iroyinowuro.com.ng	603	News	Current Affairs
oroyoruba.blogspot.com	139	Blog	Culture
yo.globalvoices.org	81	News	Current Affairs
edeyorubarewa.com	33	Blog	Fashion
Total	51,407		

Table 1: Data sources for the Yankari dataset

These sources were chosen for their reliability and the authenticity of their Yoruba content. We deliberately avoided certain large-scale datasets:

- Web-crawled datasets like mC4 and OSCAR, due to concerns about quality and authenticity, particularly the issue of translationese.
- Religious text datasets like JW300, to avoid potential biases and ethical issues associated with religious texts.

3.3 Data Processing Pipeline

Our data processing pipeline consisted of several key steps:

3.3.1 HTML Parsing and Text Extraction

We implemented a robust HTML parsing system using BeautifulSoup to extract clean text from web pages. This process involved:

- Removing all script and style elements
- Extracting text from relevant HTML tags (e.g., <p>, <h1>, <h2>, etc.)
- Preserving the document structure by maintaining paragraph boundaries

3.3.2 Encoding Normalization

All text was converted to UTF-8 encoding to ensure consistency across the dataset.

3.3.3 Deduplication

We employed a two-step deduplication process:

1. Exact matching at the document level to remove duplicate web pages
2. Near-duplicate detection at the paragraph level using MinHash and Locality-Sensitive Hashing (LSH) techniques

3.3.4 Content Filtering

We implemented several filtering steps to improve data quality:

- Removal of very short texts (less than 50 characters)
- Filtering out texts with a high ratio of non-Yoruba characters
- Removal of lines which don't end in a terminal punctuation mark
- Discarding documents with fewer than 5 sentences
- Removal of documents containing Lorem ipsum placeholder text

3.4 Corpus Statistics

Our final Yankari dataset consists of:

- Total number of documents: 51,407
- Total number of tokens: 30,438,702
- Average tokens per document: 592.11

3.5 Domain Analysis

The distribution of web domains in our corpus reflects the diverse sources we targeted. The top 5 domains by number of documents are:

1. yo.wikipedia.org: 32.70%
2. alaroye.org: 20.49%
3. www.bbc.com: 16.05%
4. www.awikonko.com.ng: 10.58%

5. yoruba.von.gov.ng: 4.94%

This distribution ensures a balance between encyclopedic content, news, and cultural discussions, providing a comprehensive representation of written Yoruba across various domains.

3.6 Ethical Considerations and Excluded Content

Throughout our data collection and processing, we prioritized ethical considerations:

- We explicitly removed data from restricted sources.
- We filtered out machine-translated content to maintain linguistic authenticity.
- We removed inappropriate or offensive material to ensure the dataset's suitability for a wide range of applications.
- We respected copyright and intellectual property rights by only including publicly accessible content and providing attribution through source URLs.

In line with our commitment to transparency, we acknowledge that our content filtering process may have introduced certain biases:

- The removal of very short texts may have disproportionately affected certain types of content, such as social media posts or headlines.
- Our focus on standard Yoruba may have led to the underrepresentation of regional dialects or colloquial expressions.
- The exclusion of content with non-Yoruba characters might have removed some culturally relevant content that includes code-switching or borrowings from other languages.

3.7 Output Format

The final dataset is stored in JSONL format, with each line containing a separate JSON document with the following fields:

- **text**: The main content of the document in Yoruba.
- **url**: The original URL from which the content was sourced.
- **source**: A code indicating the source of the document (e.g., "ACFLP").

Here's a sample entry from our dataset:

```
{
  "text": "O ma se o! Ijamba oko ofurufu gba emi eeyan marun-un: Ko din ni
            eeyan marun-un ti won tun ti ta teru nipa lonii, ojo Isegun, ojo
            karun, osu keta, odun 2024 nigba ti won so pe oko baluu kekere
            kan ja lule lori ise...",
  "url": "https://www.awikonko.com.ng/2024/03/o-ma-se-o-ijamba-oko-ofurufu-gba
      -emi.html",
  "source": "ACFLP"
}
```

Note: The sample text above has been transliterated to remove diacritical marks for display purposes. The actual dataset contains the full Yoruba text with proper diacritical marks.

This format allows for easy processing and analysis of the dataset while maintaining crucial metadata about each document's origin.

3.8 Quality Assurance

To ensure the highest quality of our dataset:

- We involved native Yoruba speakers in the data cleaning and validation process.

- We conducted regular spot checks throughout the data processing pipeline.

- We performed a final manual review of a randomly selected subset of the data to verify its quality and authenticity.

3.9 Limitations and Potential Biases

We acknowledge the following limitations and potential biases in our dataset:

- Internet Bias: Our data collection method inherently favors content that is available online, which may not fully represent the language use of Yoruba speakers with limited internet access.

- Written Language Bias: The dataset primarily captures written Yoruba, which may differ from spoken varieties of the language.

- Source Bias: Despite our efforts to include diverse sources, certain domains (e.g., Wikipedia and news sites) are overrepresented due to their abundance of content and ease of collection.

- Temporal Bias: The dataset may overrepresent contemporary language use, potentially underrepresenting historical or traditional forms of Yoruba.

- Standardization Bias: Our filtering process may have inadvertently favored standardized Yoruba, potentially underrepresenting regional variations or colloquial forms of the language.

These limitations highlight areas for future work and expansion of the Yankari dataset.

4 Limitations and Ethical Considerations

While the Yankari dataset represents a significant contribution to Yoruba language resources for NLP, it is important to acknowledge its limitations and the ethical considerations that arise from its creation and potential use. This section outlines key areas of concern that should be considered by researchers and practitioners working with this dataset.

4.1 Representation Bias

The dataset primarily captures written Yoruba from internet sources, which may not fully represent the language's full range of use. This limitation has several implications:

- Spoken Language: The dataset does not include samples of spoken Yoruba, which may differ significantly from written forms in terms of vocabulary, syntax, and idiomatic expressions.

- Informal Variants: Internet sources tend to favor more formal language use, potentially underrepresenting colloquial or informal variants of Yoruba that are common in everyday communication.

- Demographic Skew: Internet access and content creation are not uniformly distributed across all Yoruba-speaking demographics, which may lead to overrepresentation of certain socioeconomic or age groups in the dataset.

4.2 Automated Processing Limitations

Our reliance on automated collection and cleaning processes, while enabling the creation of a large-scale dataset, introduces potential issues:

- Error Propagation: Automated systems may consistently make certain types of errors that could have been identified and corrected through manual review by native speakers.

- Contextual Nuances: Automated processes may miss subtle contextual cues or cultural references that human reviewers would recognize, potentially leading to misinterpretations or loss of important linguistic information.

- Quality Variance: The quality of processing may vary across different types of content or sources, potentially introducing inconsistencies in the dataset.

4.3 Diacritization Challenges

Our approach to diacritization normalization may not fully capture the nuances of Yoruba tonal patterns:

- Tonal Ambiguity: Incorrect or missing diacritical marks could lead to ambiguity in word meanings, as Yoruba is a tonal language where tone can distinguish between otherwise identical words.
- Standardization Issues: The lack of a universally adopted standard for Yoruba orthography, particularly in online content, may result in inconsistencies in diacritic usage across the dataset.

4.4 Ethical Implications

The large-scale nature of our data collection raises several ethical concerns:

- Privacy: While we have made efforts to remove personal information, the scale of the dataset means there is a risk of unintended inclusion of private or sensitive information.
- Cultural Appropriation: There is a risk of the dataset being used in ways that appropriate Yoruba cultural expressions without proper understanding or respect for their origins and significance.

Acknowledging these limitations and ethical considerations is crucial for the responsible use and further development of the Yankari dataset. We encourage users of this dataset to be mindful of these issues and to work towards mitigating them in their applications and future research.

5   Conclusion

The Yankari dataset represents a significant step forward in addressing the resource gap for Yoruba in Natural Language Processing. By providing a large-scale, high-quality, and ethically sourced corpus, we have laid a foundation for advancing NLP research and applications in this important West African language. Our rigorous methodology, which prioritizes data quality, diversity, and ethical considerations, sets a new standard for the development of language resources for low-resource languages.

The creation of Yankari has highlighted several critical challenges in developing NLP resources for languages like Yoruba. These include the scarcity of diverse, high-quality online content, the complexities of automated processing for languages with limited existing NLP tools, and the ethical considerations surrounding data collection and potential misuse. By transparently discussing these challenges and our approaches to addressing them, we hope to contribute to the broader conversation on responsible AI development for diverse languages and cultures.

References

Adebara, I., Elmadany, A., and Abdul-Mageed, M. (2024). Cheetah: Natural language generation for 517 african languages.

Adelani, D. et al. (2021). Menyo-20k: A multi-domain english - yoruba corpus for machine translation and domain adaptation. In *Proceedings of the Fourth Workshop on Technologies for MT of Low Resource Languages*.

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Alabi, J. et al. (2020). Massive vs. curated embeddings for low-resourced languages: The case of yorùbá and twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C., Amodei, D., Joseph, N., Kaplan, J., and McCandlish, S. (2022). Scaling laws and interpretability of learning from repeated data.

Hutchinson, B. (2024). Modeling the sacred: Considerations when using religious texts in natural language processing. *arXiv preprint arXiv:2404.14740*.

Nguyen, D. M. et al. (2023). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Ogueji, K. et al. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. *arXiv preprint arXiv:2011.03823*.

Oladipo, A. et al. (2023). Wura: A multilingual dataset of african languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Ortiz Suárez, P. J., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.