CURATING YORUBA SPEECH RESOURCES: DATA NOTE AND ROADMAP

ABSTRACT

This data note documents the creation of a consolidated registry for Yoruba speech resources sourced from Mozilla Common Voice, OpenSLR, and Coqui's dataset index. We summarise available metadata, highlight demographic and technical coverage gaps, and propose a phased roadmap for improving the breadth and quality of Yoruba speech corpora. The registry is distributed as `data/yoruba_speech_registry.csv` and is intended for publication on Zenodo alongside this note.

1. BACKGROUND

Yoruba (ISO 639-1: `yo`) is a Niger-Congo language spoken by more than 45 million people across Nigeria, Benin, Togo, and the Yoruba diaspora. Despite its large speaker base, Yoruba remains under-served in speech technology resources relative to high-resource languages. Existing corpora are fragmented: Mozilla Common Voice offers community-contributed speech, OpenSLR host a curated studio-quality dataset from Google, and community indexes such as Coqui's STT dataset list currently lack Yoruba coverage. This fragmentation hinders researchers and product teams seeking representative audio covering dialects, age groups, and use cases.

2. METHODS

We harvested metadata from publicly accessible dataset cards and indexes without downloading any audio content. For Common Voice, we extracted Yoruba statistics from the Corpus 23.0 JSON metadata file. For OpenSLR we analysed catalogue pages and auxiliary TSV files listing speaker segments. We also scanned the Coqui STT dataset index to confirm the absence of Yoruba entries. All intermediate sources are archived in the `data/` directory and referenced in the registry. Processing scripts are provided in `scripts/` to ensure reproducibility.

3. EXISTING YORUBA SPEECH CORPORA

3.1 MOZILLA COMMON VOICE

Corpus 23.0 (released 17 September 2025) reports 4,852 Yoruba clips contributed by 132 users, totalling 8.14 hours with 5.76 validated hours. Gender metadata shows a slight female majority (37%) with 30% male and 33% unspecified. Age annotations are concentrated among contributors in their twenties (61%) with minimal representation from older speakers. Sentence domain tagging is effectively absent for Yoruba because all clips are reported under an unspecified domain.

3.2 OPENSLR SLR86

OpenSLR's SLR86 release contains high-quality Yoruba read speech recorded by volunteers under studio conditions. The catalogue provides separate metadata for male (1,691 clips) and female (1,892 clips) speakers along with annotation guidelines describing disfluency tags such as `[breath]` or `[external]`. The corpus is licensed under Creative Commons Attribution-ShareAlike 4.0 International. Duration metrics are not published, but the original Interspeech 2020 paper estimates approximately 36 hours of recorded speech; this figure requires verification from raw audio and is therefore excluded from the registry.

3.3 COQUI STT DATASET INDEX

The latest revision of the Coqui dataset index does not list Yoruba resources. This absence confirms that downstream tooling depending on the index will not surface Yoruba data, reinforcing the need for an openly maintained registry.

4. GAP ANALYSIS

Table 1 (embedded in the registry CSV) summarises gaps:

- **Demographics**: Common Voice contributions are dominated by younger adults, and more than a third of speakers decline to share gender information. No dialect metadata is available, limiting coverage of major regional varieties (Ibadan, Lagos, Ekiti, Ijebu, etc.). OpenSLR SLR86 provides binary gender splits only and omits age information.
- **Recording context**: Common Voice clips are crowd-sourced with varying audio quality, while SLR86 captures clean studio speech. There is no publicly catalogued spontaneous or conversational Yoruba speech. Background noise annotations exist only for SLR86.
- **Text normalisation and orthography**: Yoruba orthography relies on tone marks and underdots. Common Voice sentences sometimes omit diacritics, creating ambiguity. SLR86 follows strict orthography but lacks documentation on normalisation pipelines. A harmonised text-normalisation checklist is absent across corpora.
- **Accessibility of metadata**: Common Voice JSON metadata does not explicitly state the dataset license, requiring external confirmation. SLR86's about page lacks duration summaries. Coqui's index omission means discovery is limited to manual searching.
- **Roadmap dependencies**: Releasing additional Yoruba corpora will require targeted community outreach, transcription tooling that supports tonal orthography, and alignment across data hosts to expose consistent metadata.

5. ROADMAP

PHASE 1: METADATA CONSOLIDATION (0–3 MONTHS)

1. Publish the registry and this note on Zenodo with a DOI, enabling citation and downstream reuse.
2. File an issue with Mozilla Common Voice requesting that license metadata be embedded in per-language JSON statistics.
3. Submit a pull request to Coqui's dataset index adding Yoruba resources with direct links and quality notes.

PHASE 2: DATA ENRICHMENT (3–9 MONTHS)

1. Partner with Yoruba language communities and universities to diversify speaker demographics, prioritising older adults and underrepresented regions.
2. Extend Common Voice sentence collection to include domain-specific prompts (health, finance, transportation) and ensure diacritics are retained during validation.
3. Annotate SLR86 clips with estimated durations and background condition labels derived from metadata analysis to improve downstream filtering.

PHASE 3: NEW CORPUS DEVELOPMENT (9–18 MONTHS)

1. Launch a spontaneous speech collection pilot capturing dialogues and code-switching scenarios common in Lagos and diaspora communities.
2. Develop open-source Yoruba text-normalisation rules and lexicons, integrating them into transcription toolchains.
3. Establish an annual quality audit across hosts (Common Voice, OpenSLR, future partners) to monitor demographic balance, transcription accuracy, and licensing clarity.

6. ZENODO DEPOSITION PLAN

The Zenodo record will include the registry CSV, this PDF note, and ancillary metadata files (e.g., `scripts/build_registry.py`, source metadata snapshots in `data/`). Suggested keywords: `Yoruba`, `speech`, `dataset registry`, `Common Voice`, `OpenSLR`. Contributors should supply ORCID identifiers and select the "Open Access" embargo. Funding acknowledgement: "Self-initiated project to consolidate Yoruba speech resources."

7. CONCLUSION

Curating a single registry for Yoruba speech resources reveals significant coverage gaps across demographics, dialects, and metadata transparency. The recommended roadmap prioritises low-cost interventions—metadata consolidation, community outreach, and documentation improvements—that enable researchers and developers to build inclusive Yoruba speech technologies. Publishing the registry on Zenodo with a DOI ensures that future updates can be versioned and cited, anchoring

long-term collaboration across the Yoruba language technology ecosystem.

REFERENCES

1. Mozilla Common Voice. "Corpus 23.0 Yoruba statistics." `cv-corpus-23.0-2025-09-05.json` (2025).
2. OpenSLR. "Crowdsourced high-quality Yoruba speech data set." SLR86 catalogue (2020).
3. Coqui STT. "Dataset index" (accessed 2025).