Open but Unvetted: The Ethics of African Language Data

Ernst A.P. van Gassen
Arktos AI Labs
The Netherlands
evg.gassen@gmail.com

Abstract

Creative Commons (CC) licenses are prevalent in African natural language processing (NLP) corpus releases, but their compatibility implications are rarely examined systematically. CC-BY-SA and CC-BY-NC cannot be combined in a single published dataset; a NoDerivs (ND) clause prohibits redistribution of tokenised or annotated derivatives. This paper presents an empirical audit of license provenance across more than twenty corpus families used in African NLP, applying established compatibility rules to three case-study languages: Kituba/Munukutuba, Zarma, and Moore. Four failure modes are documented with primary-source evidence: outright prohibition (JW300, removed from OPUS after a legal audit confirmed a Terms of Service violation); composite license misrepresentation (WAXAL, whose CC-BY 4.0 claim is contradicted by its HuggingFace dataset card); a ND restriction not reflected in the CC-BY label (Tanzil); and data persistence failure (the Congolese Radio Corpus, where 402 of 405 source URLs are no longer accessible). A due diligence checklist and a survey of legally compliant enrichment opportunities conclude the paper. We argue that lawful data use is an ethical baseline: for African language communities with limited institutional recourse, license violations are not only legal risks but ethical failures that compound existing power asymmetries.

Keywords: dataset licensing, license compatibility, data provenance, Creative Commons, African NLP, low-resource languages, reproducibility, corpus auditing

1. Introduction

NLP researchers are typically not trained in law. For high-resource languages, this is often not a practical problem. Many widely used corpora have been informally vetted through decades of use. For low-resource African languages, these conditions often do not hold.

Since 2019, parallel corpora, named entity recognition (NER) datasets, sentiment benchmarks, and speech resources have been released for dozens of African languages. Much of this output has not received systematic license review. The consequences are beginning to surface. JW300 (Agić and Vulić, 2019) was a parallel corpus covering 300+ languages, including many with no alternative source. It was built in violation of the Jehovah's Witnesses website Terms of Service, which prohibit text and data mining. A legal audit by the Centre for Intellectual Property and Information Technology (CIPIT) in Nairobi confirmed this (Centre for Intellectual Property and Information Technology Law (CIPIT), 2020). OPUS subsequently removed the corpus. Datasets, models, and benchmarks that incorporated JW300 now carry a contaminated provenance chain.

JW300 is not an isolated case. GEITje, a Dutch-language model, was taken offline after copyright enforcement action by Stichting BREIN (Rijgersberg, 2023; nu.nl / Tweakers, 2024; RTL Nieuws, 2024; Tweakers, 2024; GoingDutch.ai, 2024). For high-resource languages, such incidents are disruptive but recoverable. For African languages, losing a single corpus may mean losing the only usable source for that language. Common Corpus (Langlais et al., 2025), a roughly two-trillion-token corpus curated for open licensing, illustrates the baseline: in our audit of its training table, 15 native Sub-Saharan languages account for about 91 rows combined, compared with 18,485 for English (Section 5.1).

The stakes are higher for low-resource African languages than in many other NLP settings, for reasons that compound: when a source is lost, there may be no usable substitute; annotation investment is sunk; a single license conflict can eliminate a substantial share of the available data; and problematic sources can propagate through multilingual benchmark releases.

This paper's contribution lies in application rather than framework. License compatibility logic and machine-readable rights expression have been standard tools in the library and open source communities since the early 2000s; their limited adoption in African NLP practice is a discipline-specific gap, and documenting this gap is the paper's empirical contribution. We also advance a normative argument: lawful data use is an ethical baseline, not a separate concern from the ethics of NLP research. A corpus that violates its source's license is not only a legal risk but also an ethical failure toward the communities whose language it encodes. For African language communities with limited institutional recourse, such failures may compound existing power asymmetries. We propose that data provenance declarations become standard practice in NLP ethics statements, analogous to Institutional Review Board (IRB) approval statements in human subjects research.

Three languages anchor the case studies. **Kituba** (`ktu/mkw`), the vehicular language of southern Congo-Brazzaville, has 5–8 million speakers but is absent from FLORES-200. **Zarma** (`dje`), spoken by 5 million people in Niger, has MT and NER resources but lacks representation in major benchmarks; OCHA lists it as a priority communication language. **Moore** (`mos`), with FLORES-200

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 128–139
12 May 2026.

coverage and active research groups in Burkina Faso, serves as the upper bound for a relatively well-studied low-resource language.

2. Related Work

Machine-readable rights metadata and license interoperability. The challenges of license compatibility and machine-readable rights expression predate NLP's engagement with open data by two decades. Open source license incompatibility, the inability to combine code under the GNU General Public License (GPL) with other copyleft licenses, was recognised as an engineering and legal problem by the late 1990s. It led to practical responses such as the SPDX (Software Package Data Exchange) identifier standard and the Open Source Initiative (OSI) license approval process (Rosen, 2004; Linux Foundation, 2010). The Open Digital Rights Language (ODRL), a W3C standard first published in 2001, proposed a formal ontology for expressing permissions, prohibitions, and obligations over digital content in machine-actionable form (Iannella and Villata, 2012). Creative Commons addressed machine-readability directly. The CC Rights Expression Language (ccREL, 2008) embeds license metadata using RDFa, enabling automated tools to parse the license of a webpage without human interpretation (Abelson et al., 2008). Library and information science communities engaged these issues early. They debated machine-actionable rights statements and digital repository interoperability over the same period (Coyle, 2004). These instruments have existed for over twenty years. Their limited application in African NLP corpora is therefore not due to novelty. The tools and concepts are established. The gap is discipline-specific, and documenting it is the central contribution of this paper.

Legal scholarship on open licensing. Creative Commons licensing has attracted sustained scholarly critique. Katz (2006) identifies two structural problems: variant proliferation creates user confusion, and ShareAlike terms create compatibility deadlocks that prevent the legal distribution of derivatives. Boyle (2003) provides a theoretical framing, arguing that restrictive intellectual property (IP) licensing constitutes a second enclosure movement.

Data provenance and license audits in NLP. Gebru et al. (2021) and Bender and Friedman (2018) propose structured documentation standards for datasets and NLP corpora. Both argue that provenance and licensing should be treated as first-class metadata. Dodge et al. (2021) applies this approach to large webtext corpora, identifying machine-generated text and benchmark contamination that documentation would likely have revealed. Kreutzer et al. (2022) audits 205 language-specific corpora across five multilingual web-crawl datasets and finds that at least 15 contain no usable text, while many use ambiguous language codes. The challenges of web-mined corpora for large language models (LLMs) pre-training are surveyed in Perełkiewicz and Poświata (2024).

The closest prior work to this paper is the Data Provenance Initiative (Longpre et al., 2023, 2024). It audits 1,800+ text datasets used to LLMs, reporting license omission rates above 70% and error rates above 50%. The initiative operates at the level of general-purpose LLM training data and does not focus on African languages, construct a license compatibility matrix, or analyse the failure modes documented here. Mahari and Longpre (2024) extends this line of work into legal analysis, arguing that provenance documentation affects fair use claims for fine-tuning data.

Legal analysis of AI training data. Henderson et al. (2023) analyses the four fair use factors under United States copyright law as applied to foundation model training, concluding that fair use is plausible but not guaranteed. Lee et al. (2023) maps copyright questions across the full generative AI supply chain. Jernite et al. (2022) proposes a multi-stakeholder data governance framework that addresses licensing at each stage of the data lifecycle. None of these works applies this framework to low-resource African languages.

African NLP data licensing. Nekoto et al. (2020) is the founding Masakhane paper and one of the first African NLP works to address data ownership and licensing governance explicitly. It brought the JW300 licensing issue to community attention and motivated the subsequent CIPIT legal audit (Centre for Intellectual Property and Information Technology Law (CIPIT), 2020). Adelani et al. (2021, 2022) document licensing decisions for MasakhaNER releases. Okerie and Marivate (2024) surveys the African NLP community on copyright barriers, finding that the JW300 withdrawal created downstream disruption for projects with no alternative sources. Omino (2025) proposes the Nwulite Obodo Open Data License (NOODL), a tiered community license designed for African language datasets.

Tiedemann (2020) demonstrates the value of Tatoeba for low-resource machine translation (MT) benchmarking. Here, we focus on the African-language subset and the licensing constraints governing which sources can be legally combined. The compatibility matrix in Section 4 is the practical output of this analysis.

3. License Taxonomy

I define six license tiers for African NLP text corpora, ordered from least to most restrictive. For non-specialist readers: **NC** (Non-Commercial) means the resource may not be used for revenue-generating purposes as defined by the license. What constitutes commercial use is context-dependent and jurisdiction-sensitive, but publishing an annotated dataset via a paid service or commercial API is a clear case. **ND** (NoDerivs) means

the license prohibits *sharing* modified, adapted, tokenised, or otherwise derived versions. Private use may still be permitted under applicable law (e.g. fair use or text and data mining (TDM) exceptions), but annotated datasets derived from ND sources cannot be legally distributed under the license terms.

Tier	Description	Risk
T1	CC0 / public domain / government text (where applicable). No restrictions.	None
T2	CC-BY / MIT / Apache 2.0 (permissive licenses). Attribution required; no share-alike, no NC, no ND.	Low
T3	CC-BY-SA. Share-alike propagates to all published derivatives.	Medium
T4a	CC-BY-NC. Non-commercial restriction (NC): the resource may not be used for revenue-generating purposes. Incompatible with T3.	High
T4b	CC-BY-ND or CC-BY with an undisclosed ND clause. NoDerivs (ND): the license prohibits distributing modified or adapted versions; annotated datasets derived from T4b sources cannot be legally distributed.	High
T5	Terms of Service violation, permission denied, or copyright holder prohibition.	**Prohibited**

Table 1: License tier taxonomy. T4a and T4b are separated because their restrictions operate differently and are mutually incompatible with T3.

The key practical distinction is between T3 (share-alike propagates but derivatives are permitted) and T4b (derivatives are not permitted). Many practitioners conflate these, treating all non-T2 sources as merely requiring a more restrictive output license. This is incorrect: a T4b source cannot be incorporated into any published annotation dataset, regardless of the output license chosen.

Database rights and web-mined corpora. The tier taxonomy captures copyright license compatibility, but not database rights. In EU jurisdictions, corpus collections may also be protected by *sui generis* database rights, independent of copyright in the underlying text. Licenses such as the Open Database License (ODbL) govern extraction and reuse of the database as a whole, but do not necessarily clear rights in the underlying content (Open Data Commons, 2007). This distinction is particularly relevant for web-mined corpora, where the dataset license may apply to the collection while the underlying text remains copyrighted. For annotation datasets, which redistribute text, this creates an additional layer of legal risk not captured by the tier classification.

4. License Compatibility Matrix

Table 2 shows the legally valid output license when two corpus sources are combined. "×" denotes an incompatible combination, where no single license satisfies both sources' requirements.

	T1	T2	T3	T4a	T4b	T5
T1	T1+	T2	T3	T4a	×	×
T2	T2	T2	T3	T4a	×	×
T3	T3	T3	T3	×	×	×
T4a	T4a	T4a	×	T4a	×	×
T4b	×	×	×	×	×	×
T5	×	×	×	×	×	×

Table 2: License compatibility matrix. Each cell shows the required output license when combining a row-source and column-source in a published dataset. × = incompatible combination; no valid output license exists. T1+ = any compatible license acceptable. T4b and T5 are incompatible with all other tiers.

A note on provenance quality independent of license tier. The compatibility matrix captures output license requirements, not the trustworthiness of the collection process. Two datasets can carry the same license while having very different provenance. ParaCrawl (Bañón et al., 2020) is a web-scale parallel corpus co-financed by the EU Connecting Europe Facility, with the University of Edinburgh as lead institution, and carries CC0. ParaCrawl explicitly states that it does not own the underlying text; CC0 applies to the packaging and database rights only. Its institutional context provides a degree of accountability that informal web scrapes do not. JW300 also presented as open-access but was built in violation of platform Terms of Service. The license tier alone does not distinguish these cases; provenance must be assessed separately. The matrix therefore addresses compatibility, not provenance; both must be evaluated in practice.

4.1. Dataset License versus Redistribution Rights

Several widely used corpora are web-mined, meaning the dataset license reflects packaging or database rights rather than rights in the underlying text (Perełkiewicz and Poświata, 2024). CCMatrix (Schwenk et al., 2021) is mined from Common Crawl snapshots and does not state a text license on OPUS or in its paper. NLLB mined bitext (NLLB Team et al., 2022) uses Common Crawl Web Extracted Text (WET) files as a primary source; the ODC-BY license on NLLB bitext governs database rights, not the underlying text. WURA (Oladipo et al., 2023) is built by auditing mC4 (itself derived from Common Crawl) and additional focused crawls. For these corpora, the dataset-level license does not clear the underlying text for redistribution or relicensing.

A use-case distinction is practically important. Model *training* on mC4-derived text may be defensible under fair use or EU TDM exceptions, depending on jurisdiction. Publishing an *annotated dataset*

derived from the same text constitutes redistribution of copyrighted content. The dataset's Apache 2.0 or CC packaging label does not change this; it applies to the collection, not the underlying text. This distinction is not captured by the compatibility matrix. For the primary output of the African NLP community, published annotation datasets, redistribution risk therefore applies to mC4-derived sources regardless of their stated license. Rights-cleared sources (UDHR, TICO-19, FLEURS, SMOL, original speech recordings) avoid this risk. Practitioners who use WURA or Leipzig as annotation seeds for published named entity recognition (NER) or part-of-speech (POS) datasets are redistributing copyrighted web text without explicit permission from the original rights holders.

One indicator of provenance quality is institutional context. Projects with identifiable institutional backing, named investigators, and public ethics disclosures tend to provide more accountability than anonymous uploads. This is a signal of lower risk, not a guarantee. Institutional affiliation does not substitute for verification of the actual collection method.

ParaCrawl (Bañón et al., 2020) illustrates both sides of this distinction. It is an EU-funded project with named academic leads and documented collection methods, but it is still web-mined. The underlying text originates from websites across multiple jurisdictions, and the CC0 label applies to the dataset as released rather than implying rights over the individual texts. Institutional context therefore reduces uncertainty but does not resolve underlying rights questions. For African languages, ParaCrawl's bonus releases include English–Swahili (132,517 sentence pairs, CC0) and English–Somali (14,879 sentence pairs, CC0).

Three results from this matrix are practically important for African NLP:

(1) T3 × T4a = incompatible. Wikipedia (CC-BY-SA, T3) and the 27Group Feriji Zarma corpus (CC-BY-NC, T4a) cannot be combined in a single published dataset under a valid license. A practitioner who annotates Wikipedia sentences alongside Feriji sentences and publishes the result would face incompatible licensing requirements. Wikipedia's share-alike requirement implies CC-BY-SA output, while Feriji's non-commercial restriction requires CC-BY-NC. No single license satisfies both.

(2) T4b × anything = blocked. Any corpus with a ND clause, including Tanzil, a widely used Quran translation corpus (`tanzil.net`), cannot be used to create a published annotation dataset under the license terms. The annotation constitutes a derivative work. This is not a question of the output license; distributing an annotated dataset derived from such sources is not permitted under the license, regardless of jurisdiction-specific exceptions that may apply to private use.

(3) T3 propagates; T4a similarly. A T2 source combined with a T3 source produces a T3 output. A T2 or T3 source combined with a T4a source produces a T4a output, as the non-commercial restriction is inherited. Practitioners who use Wikipedia as an annotation seed must therefore release under CC-BY-SA 4.0. MasakhaNER 2.0 (Adelani et al., 2022) uses Wikipedia text in its annotation pipeline. The HuggingFace dataset card lists CC BY-NC 4.0 for the dataset release, while the source-text licensing is heterogeneous. Practitioners should verify the specific version they use. The key point is that license decisions must be made *before* annotation begins. Choosing CC-BY-SA may limit certain downstream commercial uses.

5. African NLP Corpus Survey

Table A1 surveys corpus families used in African NLP with their tier assignments (see Appendix A). An asterisk (*) marks web-mined corpora where the dataset license covers packaging or database rights rather than the underlying text.

5.1. Common Corpus: African Language Representation

We streamed the full Common Corpus training split (Langlais et al., 2025) and filtered by the `language` field. All 91 native Sub-Saharan rows carry CC-BY-SA licenses; the `subset` and `url` fields are `null` throughout. Text content identifies the source as Wikipedia 2023 via MediaWiki markup (e.g. `{{infobox tanàna` in Malagasy rows).

The 16 rows labelled "Various open data" (Lingala 7, Kabyle 5, Wolof 3, Hausa 1) appear to be language-identification errors on French archival documents; none contains usable African-language text. Common Corpus is therefore not an independent African-language source: it largely repackages the same Wikipedia dumps audited in Table A1, with less detailed provenance metadata. Researchers should not count both as separate entries.

The ratio is approximately 200:1 (English 18,485 rows; all 15 native Sub-Saharan languages combined, 91 rows); Afrikaans alone, with 11 rows, exceeds the native total. This imbalance also reflects a structural feedback loop. Platforms such as YouTube generate CC-licensed ASR transcripts only for languages with an existing seed model. Swahili is currently the only Sub-Saharan language with YouTube ASR support. It accumulates more CC text with each upload, while the same process does not occur for Lingala, Kikongo, or Tshiluba. A critical mass of labelled speech data is therefore a prerequisite, not merely a goal.

5.2. Applying the Compatibility Matrix to Case-Study Languages

Table 3 shows which source combinations are legally valid for the three case-study languages and the output license each combination requires.

For Moore, the license landscape is relatively clean. The main sources (MT560, FLORES-200, WURA, MooreFRCollections) are all T2 or T3 and compatible. One caveat applies: the 125,695-row Moore sentiment dataset (`michsethowusu/mossi-sentiments-corpus`) assigns labels via English back-

Combination	Valid?	Output license
Kituba (ktu/mkw)		
Leipzig mkw + SMOL ktu	Yes	T2 (CC-BY)
Leipzig mkw + kgwiki (CC-BY-SA)	Yes	T3 (CC-BY-SA)
Leipzig mkw + Mozilla TTS mkw (NOODL)	**No**	NOODL-1.0
Zarma (dje)		
MT560 dje + 27Group NER	Yes	T2 (CC-BY 4.0)
MT560 dje + 27Group noisy GEC	Yes	T3 (CC-BY-SA 4.0)
MT560 dje + 27Group Feriji	Yes	T4a (CC-BY-NC 4.0)
Wikipedia + Feriji	N/A	No Wikipedia
Feriji + 27Group GEC (T3)	**No**	T4a $\times$ T3 = $\times$
Moore (mos)		
MT560 mos + FLORES-200 mos	Yes	T3 (CC-BY-SA 4.0)
MT560 mos + MooreFRCollections	Yes	T2 (CC-BY 4.0)
MT560 mos + MossiSentiments	Yes	T2 (MIT)
FLORES-200 + WURA mos	Yes	T3 (CC-BY-SA 4.0)

Table 3: Compatibility analysis for case-study language combinations. "No" entries indicate legally invalid combinations under the license terms.

translation through DistilBERT. In the NLP annotation literature, *silver* labels are automatically generated, while *gold* labels are human-annotated in the target language. This dataset should not be used as a gold standard for sentiment benchmarking without human verification of a stratified sample.

For Zarma, the combination of Feriji (T4a) with the 27Group noisy GEC corpus (T3) is incompatible. Both corpora are published by the same research group. A practitioner who used both in a single annotation pipeline would face incompatible licensing requirements and could not release the resulting dataset under a valid license.

6. Four Failure Modes

The four cases below are not intended as a random sample of poor practice. They represent four structurally distinct ways in which license compliance can fail in African NLP, each with different causes and implications. **(a) Rights violation at source** (JW300): the corpus was built in breach of the source's Terms of Service; no downstream license choice can remedy a prohibited collection. **(b) Label misrepresentation** (WAXAL): a composite dataset was published with a uniform license claim that contradicts the per-provider terms; practitioners acting in good faith on the stated label may unknowingly violate those terms. **(c) Hidden restriction** (Tanzil): a NoDerivs clause was present on the license page but absent from the CC label visible to aggregators; standard tier-based interpretation fails when the label is incomplete. **(d) Infrastructure failure** (Congolese Radio Corpus): the corpus existed as a set of third-party platform URLs; when those URLs became unavailable, the resource could no longer be verified. These four modes share a common structure: in each case, an implicit legal assumption was made that would not have survived explicit examination.

6.1. Prohibition: JW300

JW300 (Agić and Vulić, 2019) was a parallel corpus covering 300+ languages, built from the Jehovah's Witnesses website `jw.org`. It was widely used in African NLP from 2019 onward due to coverage of languages with little or no alternative parallel text.

The legal issue is clear. The `jw.org` Terms of Service (ToS) prohibit text and data mining. A legal audit by CIPIT Nairobi confirmed this (Centre for Intellectual Property and Information Technology Law (CIPIT), 2020). OPUS removed the corpus following Masakhane's formal request for permission, which was denied (Walled Culture, 2020). This corresponds to Tier 5 in the taxonomy above: prohibited regardless of how the corpus was obtained.

The risk for current practitioners is indirect. LLMs, cross-lingual embeddings, and benchmark systems trained before 2021 may incorporate JW300-derived representations. Derivative datasets built from such models may therefore carry uncertain provenance. African NLP papers should include an explicit statement: *"This dataset does not include JW300-derived text or derivatives thereof."* This is analogous to ethics approval statements in human subjects research: a reproducible declaration that reviewers can verify.

6.2. Composite License Misrepresentation: WAXAL

WAXAL (Diack et al., 2026) is a 2026 speech dataset covering 19 African languages for automatic speech recognition (ASR) and 16 for text-to-speech (TTS). The associated arXiv paper claims that the collection is released under a uniform CC-BY 4.0 license. The HuggingFace dataset card (Google, 2026) contradicts this, listing per-provider licenses. Per-provider attribution can be traced from WAXAL's supplementary tables:

- **CC-BY 4.0 (T2):** University of Ghana contributions only: Akan, Ewe, Dagbani, Dagaare, Ikposo (ASR); Fante, Twi (TTS).
- **CC-BY-SA 4.0 (T3):** All other contributions: Makerere University (Acholi, Luganda, Masaaba, Nyankole, Soga), Digital Umuganda (Fula, Lingala, Shona, Malagasy, Amharic, Oromo, Sidama, Tigrinya, Wolaytta), Media Trust (Fula, Igbo, Hausa, Yoruba, Nigerian Pidgin), Loud and Clear (Kikuyu, Luganda, Luo, Swahili), AIMS Senegal (Bambara, Pular, Wolof).

The dataset therefore combines contributions under different licenses rather than a single uniform license.

The composite misrepresentation creates a concrete legal failure. A practitioner who reads the WAXAL arXiv abstract, downloads the Lingala subset, annotates a NER dataset from its transcripts, and publishes under CC-BY 4.0 would violate the CC-BY-SA 4.0 share-alike requirement of the Digital Umuganda contribution. This occurs despite acting in good faith on the stated license. Lingala, Hausa, Igbo, Yoruba, and all Makerere-sourced languages in WAXAL require CC-BY-SA 4.0 output for any published derivative.

A noteworthy asymmetry exists. Digital Umuganda's standalone AfriVoice dataset, which covers the same Lingala recordings, is released under CC-BY 4.0. The same speech data carries different terms depending on which dataset packaging it is accessed through. Practitioners cannot resolve this without per-provider provenance tracing that the arXiv paper does not facilitate.

Composite dataset papers should include a per-language provenance and license table as a required metadata artifact.

6.3. Hidden NoDerivs Restriction: Tanzil

Tanzil (`tanzil.net`) provides Quran translations in approximately 40 languages, including several African languages (Hausa, Swahili, Somali, Amharic, partial Yoruba). Its stated license is CC-BY 3.0. In the NLP literature, CC-BY is typically treated as Tier 2: permissive, derivatives allowed, attribution required.

The Tanzil license page explicitly states: *"You are not allowed to modify this text in any way"* (Tanzil Project, 2010). This is a NoDerivs restriction (Tier 4b). It is not disclosed in the CC-BY label. A practitioner who tokenises Tanzil text, aligns it to a parallel target, and publishes the result as a training dataset has violated this restriction. The derivative prohibition applies regardless of the output license chosen.

The ND clause reflects the religious status of the Quran in Islam. Tanzil's policy holds that Quranic text may not be altered, in order to preserve the integrity of a text considered holy in Islamic tradition (Tanzil Project, 2010). This restriction is therefore not incidental but reflects a normative constraint on modification. As a result, even technical transformations such as tokenisation or alignment may fall under the prohibition on distributing modified versions.

An annotation dataset derived from Tanzil text cannot legally be published under any open license. The NoDerivs clause prohibits the derivative work entirely. To the extent that Tanzil-derived text has been incorporated into African NLP pipelines for languages with Quran translation coverage, those pipelines carry this undisclosed legal risk.

No modern African-language Quran translation is available in a clearly public-domain or CC-BY (without ND) machine-readable format. The classical English translations of Sale (1734, Project Gutenberg #7440), Rodwell (1861, #3434), and Palmer (1880, Wikisource) are public domain. These English public domain translations provide no African-language text and are of no direct utility for practitioners building African-language NLP resources.

6.4. Data Persistence Failure: The Congolese Radio Corpus

The CRC (Wheatley et al., 2020) for Lingala was published with a claim of hundreds of hours of broadcast audio sourced from YouTube. An audit conducted in February 2026 found that **402 of 405 YouTube IDs referenced in the CRC are no longer accessible**, returning 404 errors due to video removal or channel deletion. The reproducible portion of the corpus is approximately 14.4 hours of elicited LRSC speech and Radio Okapi broadcasts.

This is not a criticism of the original authors. It highlights a structural limitation: **corpora that depend on third-party platform URLs are often non-persistent.** A published corpus that cannot be reproduced by a subsequent researcher may not function as a stable scientific resource. The CRC is not an isolated case. Common Voice removes recordings when contributors withdraw consent. HuggingFace datasets are occasionally removed by their owners. YouTube channels are deleted routinely.

Corpora distributed via repositories with persistent identifiers, such as Zenodo DOIs, OpenSLR stable IDs, or LDC catalogue numbers, have remained reproducible over time. We recommend that African NLP publication venues adopt a data availability standard requiring either (a) a persistent DOI-backed deposit for all corpus resources, or (b) an explicit statement of which components are platform-dependent and may become unavailable.

A related issue is the lack of provenance documentation in community HuggingFace uploads. Several large parallel corpora for African languages carry labels such as "MT560/OPUS-derived" with no source URLs, translation pipeline documentation, or quality filter parameters. For example: `michsethowusu/english-luba-kasai_sentence-pairs_mt560` (292,000 rows, CC-BY 4.0), `michsethowusu/english-congo-swahili_sentence-pairs_mt560` (272,000 rows, CC-BY 4.0), and `michsethowusu/english-zarma_sentence-pairs_mt560` (60,000 rows, CC-BY 4.0) fall into this category. These datasets cannot be audited for license provenance. A practitioner cannot verify whether T5 sources were included in the pipeline, making them legally ambiguous despite carrying permissive license labels.

Data persistence and digital sovereignty. The CRC failure is not only a technical problem; it also raises questions of data sovereignty. The platforms implicated in African NLP data loss, including YouTube, HuggingFace, GitHub, and OPUS, are maintained by US or European organisations with limited direct accountability to African language

communities. When a corpus becomes unavailable on these platforms, there may be no institution with the mandate or capacity to recover it. This suggests a role for African-controlled digital infrastructure for language data. Initiatives such as SADILAR (South African Centre for Digital Language Resources) and the ISLRN persistent identifier system point toward a model in which African language resources are deposited in regionally managed archives with persistent identifiers (Nekoto et al., 2020; Omino, 2025). The CRC case illustrates a practical consequence of this dependency: a published corpus may become an unverifiable resource.

7. Enrichment Opportunities Within the Open-License Landscape

The foregoing analysis may appear pessimistic. The legal constraints are significant, several documented corpora have licensing issues, and authentic open-license text for under-resourced African languages is limited. A more productive reading is that clearly identifying these constraints makes enrichment tractable.

Transcribing untranscribed speech. WAXAL includes speech subsets for which transcripts are not released. The University of Ghana subsets (Akan, Ewe, Dagbani, Dagaare, Ikposo ASR; Fante, Twi TTS) carry CC-BY 4.0 licensing. Transcribing these recordings with community annotators and releasing them under CC-BY 4.0 would produce new, derivative-safe text corpora without requiring additional data collection.

Annotation of existing T2/T3 seeds. For each case-study language, T2 seed text exists and can be annotated for named NER, POS, or sentiment. The key legal decision is whether to include Wikipedia (T3, requiring CC-BY-SA 4.0 output) or restrict annotation to T2 sources (permitting CC-BY 4.0 output). This decision must be made before annotation begins, as it affects downstream commercial usability. A further distinction applies to web-mined T2 sources such as WURA and Leipzig: their packaging license does not clear the underlying text for redistribution. Publishing an annotated dataset whose seed sentences come from WURA constitutes redistribution of mC4-derived content. Rights-cleared T2 sources (FLEURS, SMOL, TICO-19) do not carry this risk and are preferable as annotation seeds when coverage is sufficient.

For Kituba, the Leipzig Corpora Collection (Goldhahn et al., 2012) `mkw_community_2017` entry (143,476 sentences, CC-BY, T2) is, to the author's knowledge, the largest available Kituba text corpus and has not been used in published NLP work. Combined with the SMOL `gatitos__en_ktu` pairs (863 sentences, CC-BY 4.0, T2), it provides an NER annotation seed with known provenance that permits CC-BY 4.0 output without share-alike propagation.

For Zarma, the MT560 parallel corpus (60,515 sentences, CC-BY 4.0, T2) provides a suitable annotation seed for CC-BY 4.0 output. The 27Group noisy GEC corpus (508,869 sentences, T3) is also available but requires CC-BY-SA 4.0 output and is incompatible with Feriji (T4a).

Parallel and bridged resources. For zero-pivot African–African pairs, the UDHR (T1, public domain) provides the same 30 articles across 570+ language editions in sentence-aligned form on OPUS. Any two editions can be paired directly without an English or French intermediary. NTREX-128 (Federmann et al., 2022) provides 1,997 professionally translated news sentences in 24 African languages under CC-BY-SA 4.0; the shared source enables direct pairing of any two languages. For languages outside these resources, bridge construction via FLORES-200 (T3, CC-BY-SA) or TICO-19 (T1, CC0) is possible where both languages have segments aligned to the same pivot. English, French, Arabic, and Portuguese cover the main regional pivot groups. Global Voices (OPUS, CC-BY 3.0) provides human-translated Swahili (≈20K pairs) and Amharic (≈1K). None of these resources substitutes for large training corpora, but all are legally clean and currently available.

The kgwiki discovery. A finding with direct enrichment implications is that the Kongo Wikipedia (`kgwiki`), labeled and indexed as *Kongo*, is written in Kituba/Munukutuba. This is supported by article content inspection and the `Svngoku` dataset card, which explicitly invites speakers of Munukutuba, Kituba, and Kikongo ya Leta to contribute. As a result, 1,200+ articles of usable Kituba text (CC-BY-SA 4.0, T3) may have been overlooked by practitioners searching under the standard ISO codes (`ktu`, `mkw`). The same mislabeling appears in FLORES-200's `kon_Latn` entry. Researchers building Kongo NLP systems may therefore have trained on Kituba data, while researchers building Kituba systems may have missed this resource. Resolving this ISO code confusion, which involves at least five codes (`kon`/`kg`, `ktu`, `mkw`, `kwy`), is a prerequisite for systematic enrichment.

Toward African-controlled data infrastructure. The enrichment opportunities identified above depend on existing open-license resources. They do not address the structural issue that African language data is predominantly hosted, licensed, and controlled by non-African institutions. A complementary approach is gated access: corpus holders deposit resources under standard CC license terms but control who can access them.

This model is already used at scale by HuggingFace gated datasets and PhysioNet. Its value for African language communities is twofold. First, legal clarity is maintained without introducing new license instruments, as standard CC terms continue to govern data use. Second, access requests create researcher-to-researcher contact: corpus holders can see who is using their data and for what purpose, supporting community coordination.

Africa Arxiv (africarxiv.org) demonstrates that African-controlled academic infrastructure can be community-maintained without large institutional backing. Zenodo restricted deposits with ISLRN identifiers could provide a practical implementation, combining persistent identification with controlled access. The remaining challenge is governance and sustained funding, which is organisational rather than technical.

The checklist in Section 7 and this infrastructure model operate at different levels. The checklist addresses what individual researchers can do with existing resources. The gated access model addresses what the community may need to build to avoid future losses such as the CRC and JW300 cases.

8. A Legal Due Diligence Checklist

A practical due diligence procedure. The following five-step procedure provides a minimal workflow for license-compliant dataset construction.

Step 1: Inventory sources. Consult Wikipedia, Leipzig, UDHR, Tatoeba, FLORES-200, FLEURS, WAXAL, WURA, TICO-19, Common Voice, and OPUS. Avoid: CCMatrix (no license), TED2020 (T4b), JW300 (T5), and Tanzil (T4b).

Step 2: Assign tiers. Use Table 1. Verify licenses against original sources: Tanzil is T4b despite its CC-BY label; WAXAL subsets are T3 despite a CC-BY 4.0 claim. Distinguish model training (TDM exceptions) from dataset redistribution (higher risk).

Step 3: Apply compatibility. Use Table 2. Resolve $\times$ by: (a) dropping one source; (b) adopting the most restrictive license; or (c) requesting a waiver. Dropping ShareAlike favours commercial use; dropping Non-Commercial protects community rights.

Step 4: Verify and Archive. Cross-reference ISO codes (e.g. kgwiki is Kituba, not Kongo). Record license versions and checksums. Deposit snapshots in persistent archives (Zenodo/SADILAR) and include a provenance declaration in the paper's ethics statement.

9. Discussion

None of the errors documented here were wilful; each reflects a legal assumption that NLP practice has not consistently made explicit. The compatibility matrix (Table 2) requires no legal expertise: it is a lookup table. Tier assignments require a one-time provenance check. The checklist requires discipline.

ShareAlike as protection. Wikipedia's CC-BY-SA requirement is often framed as a constraint: share-alike propagates and limits commercial use. This framing merits scrutiny. ShareAlike prevents a well-resourced actor from taking community data, adding proprietary value, and closing the derivative. For African communities producing resources for languages with limited commercial incentive, SA may be the appropriate choice precisely because it prevents that scenario. CC-BY maximises ecosystem adoption, while CC-BY-SA protects resources from enclosure. Both are legitimate goals. The checklist in Section 6 provides the vocabulary for this intentional decision.

Omino (2025) extends this logic. NOODL is a community-designed instrument with guardrails for Global South data communities, reflecting recognition that standard CC licenses may not address community sovereignty. As a 2025 proposal, it remains a direction rather than a tested instrument.

Ethics of Lawful Use. The NLP community has established ethics statements for human subjects and consent, but not yet for data provenance. For low-resource African languages, where a single corpus loss is unrecoverable, documentation is an ethical necessity. A corpus that violates its source's license is both a legal risk and an ethical failure toward the communities it encodes, reinforcing existing power asymmetries.

We propose a standard data provenance declaration in ethics statements, analogous to IRB approval:

> *All corpus sources were reviewed for provenance. No ToS-violating or prohibited sources are included. Tier assignments are documented in the supplementary material.*

This requires only one-time provenance checking per source. Peer review can help establish this as a community norm.

Database Rights. One dimension this paper does not resolve is the *sui generis* database right in EU jurisdictions. A corpus holder may hold rights over a collection independently of copyright in its text. The Open Database License (ODbL) addresses this right explicitly; standard CC licenses do not (Open Data Commons, 2007). Whether a text corpus constitutes a "database" under this framework remains an open legal question with consequences for those redistributing corpora assembled by European institutions. This underexplored dimension warrants further attention.

10. Conclusion

The four case studies show a common pattern: implicit legal assumptions that would not survive examination. JW300 was used because it appeared open; Tanzil because its label indicated CC-BY; WAXAL because per-provider terms were not traced; and CRC because URL persistence was not verified. All were avoidable.

Concrete outcomes include: the incompatibility of Feriji (T4a) and GEC (T3) within a single dataset; the discovery of the Leipzig `mkw` entry as the largest open Kituba corpus; and the mislabeling of Kituba text in the Kongo Wikipedia (`kgwiki`) and FLORES-200. Data logging and persistent archiving should become standard publication practices for African NLP work.

11. LRE Map

This paper does not introduce new language resources; it audits the license provenance of existing ones. No new LRE Map entries are created. All resources cited here are existing catalogued resources; their identifiers (ISLRN, HuggingFace dataset IDs, OPUS corpus IDs, or GitHub repositories) are referenced in the bibliography. The LRE Map URL is `http://lremap.elra.info`.

12. Ethical Considerations

This paper audits the license provenance of existing resources; no new datasets, models, or personal data were collected. The analysis draws on published reports, license page text, and dataset cards retrieved in January–February 2026. None of the corpora identified as legally problematic (JW300, Tanzil, TED2020) were used to produce any outputs. Licenses may change after the retrieval date; practitioners should verify them independently. This paper does not constitute legal advice. Future annotation work on the languages surveyed should follow community consent protocols as outlined by Nekoto et al. (2020).

13. Limitations

Jurisdiction-specific law (e.g., EU TDM exceptions under the DSM Directive) may affect practical conclusions; redistribution of modified corpora would remain restricted. Tier assignments for MT560/HuggingFace datasets are provisional due to undocumented provenance. This audit does not demonstrate a downstream NLP application; future work building annotated corpora for the three case-study languages would test the practical value of the checklist. The tier assignments in Table A1 are also provided as a structured data file in the supplementary material to support reuse and citation.

14. References

Hal Abelson, Ben Adida, Mike Linksvayer, and Nathan Yergler. 2008. ccREL: The Creative Commons rights expression language. Technical report, Creative Commons.

David Ifeoluwa Adelani, Jade Abbott, et al. 2021. MasakhaNER: Named entity recognition for African languages. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1116–1131. MIT Press.

David Ifeoluwa Adelani, Graham Carr, et al. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508. Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 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. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Marcin Junczys-Dowmunt, Samuel Ma, Prashant Mathur, Paul Paul, Johann Roturier, and Rico Sennrich. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. volume 6, pages 587–604. MIT Press.

James Boyle. 2003. The second enclosure movement and the construction of the public domain. *Law and Contemporary Problems*, 66(1/2):33–74.

Centre for Intellectual Property and Information Technology Law (CIPIT). 2020. Masakhane projects' use of the JW300 dataset for natural language processing: Copyright issues, contract overrides and cross-border implications.

Karen Coyle. 2004. Rights expression languages: A report for the library of congress.

Thierno Diack et al. 2026. WAXAL: A large-scale multilingual speech dataset for African languages. *arXiv preprint arXiv:2602.02734*.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Systematic Biases in MT Research*.

Timnit Gebru et al. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

GoingDutch.ai. 2024. GEITje takedown. `https://goingdutch.ai/nl/posts/geitje-takedown/`. Accessed February 2026.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 759–765. European Language Resources Association (ELRA).

Google. 2026. WAXAL: A large-scale multilingual African language speech corpus – dataset card. https://huggingface.co/datasets/google/WaxalNLP. Accessed February 2026. Dataset card specifies per-provider licenses: University of Ghana contributions are CC-BY 4.0; Makerere University, Digital Umuganda, Media Trust, and Loud and Clear contributions are CC-BY-SA 4.0. Contradicts the uniform CC-BY 4.0 claim in the arXiv paper.

GoURMET Consortium. 2020. GoURMET: Generalisation of underrepresented languages with modern transformers and evaluation of robustness. EU Horizon 2020 Project 825299; lead institution: University of Sheffield; CC0 parallel corpora available via OPUS.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research*, 24.

Renato Iannella and Serena Villata. 2012. ODRL version 2.0 core model. W3C community group report, W3C.

Yacine Jernite, Huu Nguyen, Stella Biderman, et al. 2022. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222.

Zachary Katz. 2006. Pitfalls of open licensing: An analysis of creative commons licensing. *IDEA: The Intellectual Property Law Review*, 46(3).

Julia Kreutzer et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, et al. 2025. Common corpus: The largest collection of ethical data for LLM pre-training. *arXiv preprint arXiv:2506.01732*. Approximately two trillion tokens; multilingual, with 53% of tokens from non-Western-country sources; African languages listed as a future expansion target (Swahili, Wolof, Bambara); web-mined components carry provenance questions analogous to those in WURA and CCMatrix.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. 2023. Talkin' 'bout AI generation: Copyright and the generative-AI supply chain. *Journal of the Copyright Society of the USA*.

Linux Foundation. 2010. SPDX specification 1.0. https://spdx.org/specifications. Software Package Data Exchange standard; current version maintained at spdx.org.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing and attribution in AI. *arXiv preprint arXiv:2310.16787*.

Shayne Longpre, Robert Mahari, Anthony Chen, et al. 2024. The data provenance initiative: A large scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6.

Robert Mahari and Shayne Longpre. 2024. Discit ergo est: Training data provenance and fair use. *Network Law Review*. Winter 2024. Also available at SSRN 4795277.

Wilhelmina Nekoto et al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

NLLB Team, Marta R Costa-jussà, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*. FLORES-200 benchmark included; 200 languages, CC-BY-SA 4.0.

nu.nl / Tweakers. 2024. Ontwikkelaar haalt Nederlands AI-taalmodel offline na verzoek Stichting BREIN. https://www.nu.nl/tweakers/6343889/ontwikkelaar-haalt-nederlands-ai-taalmodel-offline-na-verzoek-stichting-brein.html. Accessed February 2026.

Chijioke Okerie and Vukosi Marivate. 2024. How African NLP experts are navigating the challenges of copyright, innovation, and access. Carnegie Endowment for International Peace.

Ifeoluwa Adeyemi Oladipo, Abdulmumin Idris, Aremu Anuoluwapo, et al. 2023. Better quality pretraining data and T5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168. Association for Computational Linguistics.

Melissa Omino. 2025. The nwulite obodo open data license (NOODL): Licensing African datasets to support research and AI in the global south. Conference on Copyright and the Public Interest in Africa and the Global South, Johannesburg. CIPIT, Strathmore University.

Open Data Commons. 2007. Open Database License (ODbL) v1.0. https://opendatacommons.org/licenses/odbl/1-0/. Addresses *sui generis* database rights independently of copyright in database contents.

Michał Perełkiewicz and Rafał Poświata. 2024. A review of the challenges with massive web-mined corpora used in large language models pre-training. *arXiv preprint arXiv:2407.07630*. ICAISC 2024. Surveys noise, duplication, bias, and legal issues in web-mined LLM pre-training data.

Edwin Rijgersberg. 2023. GEITje: A large open dutch language model. GitHub repository; model subsequently removed from HuggingFace at the request of Stichting BREIN due to copyright concerns over Dutch GigaCorpus training data. Demonstrates that training data provenance problems are not limited to low-resource languages; a Dutch (high-resource) language model was taken down following copyright enforcement action by a national rights management foundation.

Lawrence Rosen. 2004. *Open Source Licensing: Software Freedom and Intellectual Property Law*. Prentice Hall, Upper Saddle River, NJ.

RTL Nieuws. 2024. Illegale dataset van zinnen uit Nederlandse films en boeken offline. https://www.rtl.nl/nieuws/tech/artikel/5465687/illegale-dataset-van-zinnen-uit-nederlandse-films-en-boeken-offline. Accessed February 2026.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the WEB. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 932–944. Association for Computational Linguistics.

Tanzil Project. 2010. Text license — tanzil documents. Accessed February 2026. States: “You are not allowed to modify this text in any way.”.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182.

Tweakers. 2024. BREIN haalt illegale Nederlandstalige dataset voor trainen AI-modellen offline. https://tweakers.net/nieuws/225340/brein-haalt-illegale-nederlandstalige-dataset-voor-trainen-ai-modellen-offline.html. Accessed February 2026.

Walled Culture. 2020. A “blatant no” from a copyright holder stops vital linguistic research work in Africa.

15. Language Resource References

Wheatley, Julian and others. 2020. *Congolese Radio Corpus (CRC) for Lingala*. **Data persistence failure.** Originally claimed hundreds of hours of YouTube broadcast audio. Audit conducted February 2026 found 402 of 405 YouTube IDs dead (404 errors). Reproducible content: approximately 8.3 hours elicited LRSC speech (IPA-transcribed) + approximately 6.1 hours Radio Okapi broadcast audio = approximately 14.4 hours total.

A. African NLP Corpus Survey

Corpus	Tier	License	African coverage	Notes
UDHR	T1	Public domain	570+ languages	**Directly parallel**: same 30 articles across all editions; zero-pivot A–B pairs; available sentence-aligned via OPUS.
Wikidata labels	T1	CC0	Many languages	Structured data; not running text.
TICO-19	T1	CC0	amh, hau, kin, lin, lug, orm, som, swc, swh, tir, zul (16 total)	3,071 professionally translated segments; COVID domain; CC0 (not CC-BY as sometimes cited); the only OPUS resource with swc and lin as distinct codes.
GoURMET (GoURMET Consortium, 2020)	T1	CC0	amh, hau, ibo, swa, tir, yor	EU Horizon-funded; University of Sheffield lead; institutional provenance equivalent to ParaCrawl.
NLLB bitext (NLLB Team et al., 2022)	T2*	ODC-BY 1.0	35+ African languages incl. mos, lin, bam, fuv	Web-mined; ODC-BY governs database rights, not rights in the underlying text; contains Common Crawl-derived content; covers Moore and Bambara not in other T1/T2 sources.
FLEURS transcripts	T2	CC-BY 4.0	20 sub-Saharan languages	Text transcripts of speech; high quality.
WURA (Oladipo et al., 2023)	T2*	Apache 2.0	16 African + eng/fra/por/arz	Built on mC4 (Common Crawl-derived) plus focused crawls; also includes English, French, Portuguese, and Arabic (Egyptian); Apache 2.0 applies to packaging only; underlying text is copyrighted web content; redistribution as published annotation seed carries higher legal risk than model training use.
MT560/HuggingFace	T2	CC-BY 4.0	Many; varies	Provenance often undocumented; treat as T2 with caution.
SMOL/gatitos	T2	CC-BY 4.0	Selected pairs incl. ktu	Source sentences selected from Common Crawl; professional translations; small.
Common Voice	T2	CC0 (recordings)	kin, swa, hau, yor, others	Speech only; text prompts vary.
Wikipedia	T3	CC-BY-SA 4.0	40+ language editions	High entity density; propagates SA.
FLORES-200	T3	CC-BY-SA 4.0	30 African languages	1,012 sentences/language; propagates SA.
Leipzig Corpora	T2*	CC-BY	250+ languages	Downloadable sentence corpora are CC-BY; the NC restriction applies to online portal tools only; corpora are built via web crawling.
African Storybook	T2/T4a	CC-BY or CC-BY-NC	60+ languages incl. dje	Per-story license; NC stories incompatible with T3 if mixed.
BibleTTS	T3	CC-BY-SA 4.0	aka, twi, ewe, hau, lin, yor	Speech with text alignment; SA propagates.
ParaCrawl	T1*	CC0	swa (132,517), som (14,879)	EU CEF-funded; university-reviewed; CC0 covers the packaging only; ParaCrawl explicitly does not own the underlying text.
NTREX-128	T3	CC-BY-SA 4.0	24 African languages	1,997 professionally translated news sentences; same source enables direct African–African pairing; evaluation scale only.
Global Voices (OPUS)	T2	CC-BY 3.0	swa (≈20K), amh (≈1K)	Human-translated citizen journalism; no ShareAlike; only two African languages with meaningful coverage in OPUS.
eBible.org	T1–T3	Variable per translation	1,000+ languages	Must verify per translation; some T1, some T3, some T5.
Tanzil	T4b	CC-BY 3.0 + NoDerivs	hau, swa, som, amh, yor (partial)	ND clause explicit on license page; mislabeled as CC-BY on aggregator sites; distributing annotated derivatives is not permitted under the license.
27Group Feriji	T4a	CC-BY-NC 4.0	dje (Zarma)	Incompatible with T3 (Wikipedia).
Mozilla TTS mkw	T4b	NOODL-1.0	mkw (Kituba)	Community-protective instrument designed for African language data; restricts redistribution and AI derivatives without permission in order to protect community sovereignty over the resource. T4b here reflects compatibility behaviour only; NOODL is not a CC instrument. As a 2025 proposal it has not yet undergone legal validation; treat as direction rather than tested instrument (Omino, 2025).
NaijaSenti	T2	CC-BY 4.0	hau, ibo, yor, pcm	Twitter-sourced; follow platform ToS for redistribution.
bible-uedin	T1	CC0	dje, hau, swa, amh, yor, others	Both OPUS and the source repository assert CC0. Practitioners should verify that the licensors hold the rights to apply CC0 to each translation before relying on this label. Covers Zarma (`dje`).
TED2020	T4b	CC-BY-NC-ND 4.0	swa, others	Both NC and ND; derivative annotation datasets prohibited regardless of output license. Frequently cited without noting T4b status.
CCMatrix	*Unknown*	*Unstated*	Many (automatically mined)	Mined from Common Crawl; no stated license on OPUS or in paper; derived datasets carry an irresolvable provenance gap.
JW300	T5	ToS violation	300+ languages	Prohibited; OPUS removed; provenance contamination risk.

Table A1: License tier assignments for major African NLP corpus families. Asterisk (*) marks web-mined corpora where the dataset license applies to packaging or database rights rather than the underlying text. Italic entries in the Tier column represent cases where the stated license differs from the effective license after review.