Small, License-Clean, and Balanced LID Eval with Low-Resource & Constructed Languages

Anonymous ACL submission

Abstract

Language identification looks solved until you step off the happy path: scarce data, mixed scripts, romanization, and conlangs. We build a compact, license-clean evaluation set that deliberately mixes high-resource anchors (e.g., English, Portuguese), low/mid-resource languages (e.g., Amharic, Dzongkha, Yoruba), and constructed languages (e.g., Esperanto, Interlingua, Ido, Interlingue, Lojban, Toki Pona, Lingua Franca Nova, Volapük, Kotava). To stay redistribution-safe, Wikipedia is pointer-only; we materialize locally from pinned dumps. We round it out with Tatoeba (short user sentences) and UDHR translations (legal register), then run deduplication, script tagging, length bucketing, and a stratified split. The goal is pragmatic: a small, reproducible pack that exposes failure modes on scarce data and non-Latin scripts without licensing surprises.[1]

1 Introduction

LID works great on the usual suspects and genres, but breaks in the corners we actually care about: truly low-resource languages, romanized text, and constructed languages that overlap vocabulary with Indo-European families. Rather than chasing scale, we aim for a *small, reproducible, license-clean* evaluation pack: (i) easy to re-run and compare across models, (ii) diverse in scripts and domains, and (iii) explicit about scarcity so differences actually show up.

2 Method (overview)

Our pipeline is intentionally simple and scripted end-to-end:

1. **Source discovery & licensing.** Choose sources we can redistribute (or reference via pointers) and record terms in-repo.
2. **Collection.** Wikipedia (pointer-only), Tatoeba (CC-BY), UDHR translations (public document; CC0 packaging on HF).
3. **Normalization.** NFC, whitespace cleanup, light HTML stripping for Wikipedia.
4. **Balancing.** Tight per-language caps (≈ 200), with ultra-low languages kept at "all available".
5. **P1-04/05 (next).** Dedup (exact + near-dup) and script/romanization tagging.
6. **Bucketing & split.** Length buckets, then a stratified dev/test split.

3 Data Collection

Our goal was to assemble a small, *license-clean* and *balanced* evaluation set for language identification (LID) that stresses both high-resource anchors and genuinely low-resource or constructed languages. We combine three complementary sources—Wikipedia (encyclopedic prose), Tatoeba (short, user-contributed sentences), and UDHR translations (legal register)—and capture every step in code and manifests for reproducibility.

3.1 Language set and acceptance criteria

Languages were selected by four criteria: (i) availability from at least one permissive source we can redistribute or reproduce, (ii) diversity of writing systems (Latin, Arabic, Ethiopic, Georgian, Lao, Khmer, Myanmar, Tibetan, CJK), (iii) a balanced per-language target size (nominally $\sim$200 items), and (iv) inclusion of low-resource and constructed languages to stress LID.

Sets used.

- **Anchors (high-resource):** en, pt, es, tr, ja.
- **Constructed:** eo, ia, io, ie, lfn, vo, avk, jbo, tok.

[1] Code/manifests: https://github.com/oteomoura/klingon-lid-experiment

- **Low/mid-resource:** am (Amharic), ka (Georgian), ur (Urdu), lo (Lao), km (Khmer), my (Burmese), dz (Dzongkha), yo (Yorùbá); plus ultra-low *kek* (Q'eqchi') and *fuf* (Pular/Fulfulde).

When upstream sources could not reach the nominal cap (e.g., `kek`, `fuf`), we retained *all* available items and treat them explicitly as scarcity cases.

3.2 Licensing and storage policy

Wikipedia. Because Wikipedia text is CC BY-SA, we do not commit text. We pin a specific dump (20231101.*xx*) and commit *pointer manifests* (page/revision IDs), together with code that locally materializes the text for evaluation. This keeps the repository non-SA while preserving exact reproducibility from the pinned dump.

Tatoeba. Sentences are CC BY 2.0 FR. We map project codes to ISO–639–3, select per-language subsets, and commit the derived JSONL with attribution in paper/README.

UDHR. The UDHR text is public; we use the *UDHR-LID* packaging distributed on Hugging Face (CC0 for the dataset wrapper). We fetch by language code and commit the resulting JSONL with provider information.

3.3 Source ingestion and normalization

Wikipedia (pointer-only). From `wikimedia/wikipedia` Parquet shards for dump 20231101.*xx*, we uniformly sample up to 120 articles per language and store only pointers (page + revision). Locally materialized text is normalized and kept out of version control.

Tatoeba. We stream the master `sentences.csv`, map our project codes (2-letter or 3-letter) to ISO–639–3 (as used by Tatoeba), and select up to 120 sentences per target language with light character-length filtering (short lines preserved to avoid style bias).

UDHR. We fetch up to 120 segments per language via the UDHR-LID HF packaging. In practice, useful coverage for the low/mid set required a permissive minimum-length threshold; we settled on a small cutoff (effectively admitting short provisions) and rely on later balancing in the split.

Text normalization (all sources). We apply NFC normalization, collapse internal whitespace, and drop residual markup. Each JSON object records `text`, `source` (wikipedia/tatoeba/udhr-lid), `domain` (encyclopedic/sentences/legal), `lang` (project code), `code` (ISO–639–3 where applicable), licensing/provider fields, and a minimal

4 Experiments

We will report: (i) coverage per language and source, (ii) dedup rates, (iii) script distribution, and (iv) baseline LID accuracy by tier (anchors, mid, ultra-low, conlangs). Given the design, we expect most mistakes around romanization, visually similar scripts, and conlangs overlapping with Romance/Germanic vocab.

5 Results

We will include per-language accuracy, confusion patterns, and ablations on script/romanization handling.

Limitations

Wikipedia changes over time; we pin dump IDs to reduce drift but do not freeze upstream. Domains differ by source (encyclopedic vs. legal vs. short sentences), which we embrace to probe robustness but it is still a bias. For `kek` and `fuf`, upstream sources cap the absolute amount of text; we keep all available and analyze them as scarcity cases.

Ethics Statement

We use publicly available data with permissive terms (or pointer-only for ShareAlike sources). We avoid redistributing SA text directly and document licenses in-repo. Scripts include non-Latin writing systems; we take care to avoid romanization "over-correction" during normalization.

Acknowledgments

Thanks to Wikimedia, Tatoeba, and the maintainers of the UDHR-LID packaging for making this work feasibly reproducible.

References

CIS LMU. 2024. Udhr-lid: Universal declaration of human rights (lid) dataset. https://huggingface.co/datasets/cis-lmu/udhr-lid. CC0-1.0 packaging; accessed 2025-08-11.

Tatoeba Project. 2025. Tatoeba: Example sentences in many languages. https://tatoeba.org/. CC-BY 2.0 FR; accessed 2025-08-11.

Unicode Consortium. 2016. The universal declaration of human rights (udhr) in unicode. https://www.unicode.org/udhr/. XML assemblies of UDHR translations; accessed 2025-08-11.

Wikimedia Foundation. 2023. Wikipedia dataset (parquet) on hugging face. https://huggingface.co/datasets/wikimedia/wikipedia. Dump 20231101; pointer-only in our repo; accessed 2025-08-11.

A Reproducibility & Code

Repo and manifests: https://github.com/oteomoura/klingon-lid-experiment. Rerun instructions are encoded as `make` targets (pointer-only for Wikipedia; local JSONL for Tatoeba/UDHR).