Tonative: Community-Driven Extension of African Datasets Through Human-AI Collaboration

Sharon Ibejih and Cynthia Amol | (sharonibejih, amol)@tonative.org

Motivation

Community-driven initiatives produce high-quality data but face significant scalability challenges, while synthetic data generation offers scale but introduces validity issues.

Sustainable, and extensible datasets are needed to capture language evolution over time, yet current approaches each have their own critical limitations that prevent them from adequately addressing the resource gap for African languages.

Introduction

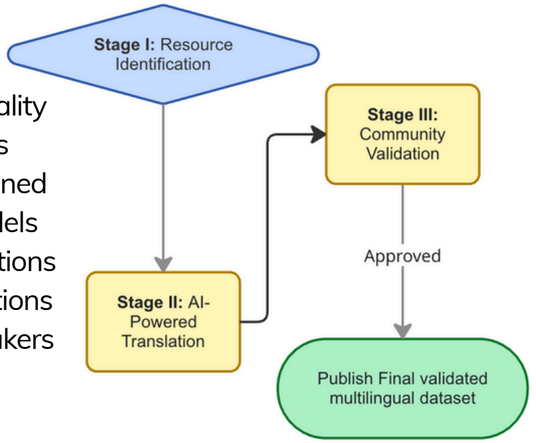
While recent community-led efforts have improved representation for major African languages, significant gaps remain. Two competing approaches currently exist in dataset creation: community-driven curation and synthetic data generation, both presenting their own limitations.

This paper presents **Tonative** (to-native), a hybrid method that combines human-AI collaboration to reduce manual validation efforts and maintaining quality control.

Our goal is to extend existing multilingual datasets into more African languages.

Methodology

- Identify high-quality existing datasets
- Leverage pretrained multilingual models for initial translations
- Validate translations with native speakers and linguists



Result

Dataset	Source Language	Target Language	Target Language ISO	Target Language Family	Original Sample	Human Validated Sample
KKD	Swahili	English	eng	-	29230	29230
XNLI	English	Igbo	ibo	Niger-Congo	7500	5025
XNLI	English	Hausa	hau	Afro-Asiatic	7500	5150
XNLI	English	Luo	luo	Nilo-Saharan	7500	4276
XNLI	English	Kinyarwanda	kin	Niger-Congo	7500	3579
XNLI	English	Nigerian Pidgin	pcm	English-based Creole	7500	3137
XNLI	English	Kikuyu	kik	Niger-Congo	7500	2397
XNLI	English	Yoruba	yor	Niger-Congo	7500	1241

Limitations

- Methodology is still being refined for optimal effectiveness.
- Requires available volunteer validators for target languages
- Dataset scalability is still dependent on community engagement levels

Future Work

- Continuous improvement of methodology
- Expansion to more African languages and datasets
- Sustainable incentive structures for validators

Tonative WiML Women in Machine Learning NEURAL INFORMATION PROCESSING SYSTEMS