Enhancing Access to Education PSAs through Kiswahili, Luo and Maasai Translation

Shalyn Muita, Arnold Bophine, Jemimah Bochaberi, Chanteel Kimathi, Melvin Lebo
United States International - University

INTRODUCTION

- **Problem**: In a world of global information, vital educational messages and PSAs often struggle to reach every member of a community. Language barriers can leave children, parents, and community leaders disconnected from critical knowledge.
- **Our Mission**: This project is a step toward changing that. It's an act of cultural preservation and empowerment, using technology to ensure we are creating a bridge, one word at a time, so that every voice is heard and every message is understood

PRIOR WORK

- MarianMT, MBart, and NLLB have advanced multilingual translation but mainly for high-resource or general-domain content.
- Kiswahili is moderately supported; Luo remains underrepresented in MT research.
- Prior studies show that fine-tuning on domain-specific data (e.g., health, legal) improves translation quality, but few target public service announcements (PSAs).
- Most existing work overlooks localized civic communication needs in Kenya.

METHODOLOGY

Data Collection & Preprocessing

- Scraped and curated 1,100+ education PSAs.
- Cleaned data by standardizing formats, fixing dates, validating URLs, and filtering unsupported languages.
- Flagged anomalies (e.g., short descriptions, invalid topics) for manual review.

Initial Translation

- Used a Google Translate script to auto-translate English PSAs into Swahili.
- Translations were manually cleaned and aligned for quality.

Model Fine-Tuning (Attempt 1)

- Fine-tuned MarianMT (opus-mt-en-sw) using English–Swahili pairs.
- Trained on 90% of data; 10% held out for testing.
- Achieved BLEU score: 59.71 after one epoch.

Advanced Translation Pipeline

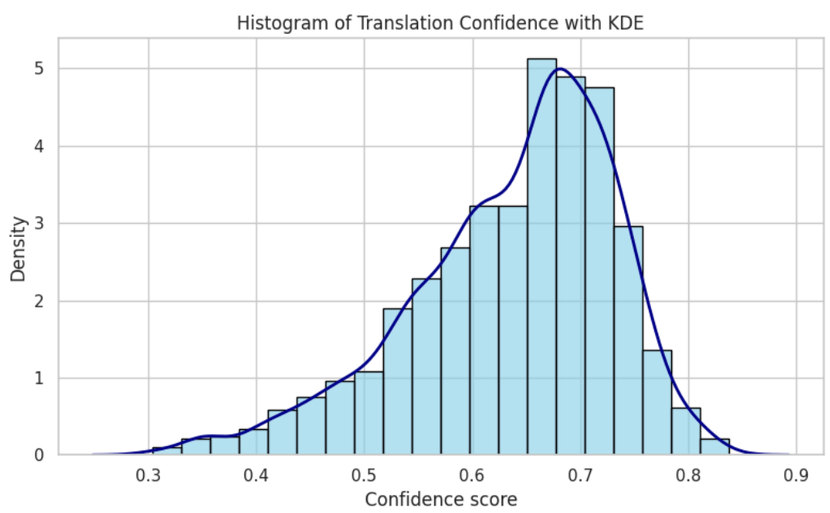
- Implemented a two-stage pipeline using NLLB-200 for English → Swahili → Luo.
- Confidence scores calculated using token-level probabilities.
- Forced language tags ensured accurate output.

The MAA translation uses a custom dictionary for word-for-word translation on small section of dataset (civic), evaluated by BLEU, perplexity, and coherence.

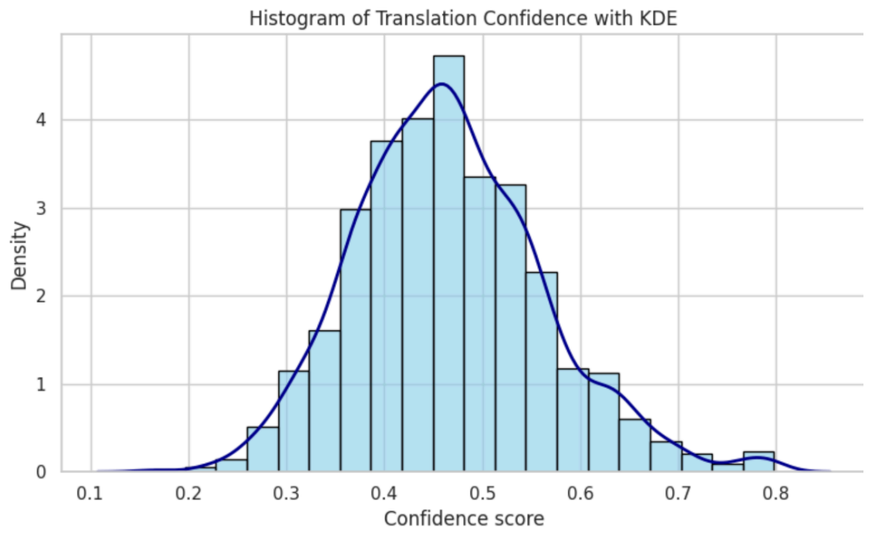
Quality & Interpretability

- Visualized translation confidence scores using KDE plots.
- Extracted examples with highest/lowest confidence.
- Final dataset saved with translations and confidence scores for deployment.

ENG TO KISW RESULTS



KISW TO LUO RESULTS



	Swahili Translations	Luo Translations	Luo Confidences
0	Wazoeze walimu!	Tieg jopuonj!	0.797792
1	Wasaidie wanafunzi!	Kony jopuonjre!	0.796322
2	Wasaidie wanafunzi!	Kony jopuonjre!	0.796322

INTERPRETATION

ENG → KISW Results

- Confidence scores mostly between 0.6–0.75, peaking near 0.7.
- Translations are generally high-confidence and consistent.

KISW → LUO Results

- Confidence scores mostly around 0.5, with more variation.
- Overall lower average confidence, but some translations (e.g. “Kony jopuonjre!”) score high (~0.796).

Conclusion:

ENG→KISW model is more reliable. KISW→LUO needs improvement but shows potential.

ENG TO MAA RESULTS

```
Sample original English sentence: I go to school
Sample model-based Maasai translation: anaa lumo sukulu
Sample model-based back-translated English: I go school

--- Quantitative Metrics ---
Average BLEU score (English -> Maasai -> English): 0.1628
Average Coherence score: 0.1636
Average Perplexity score (lower is better): 9.7570
```

CHALLENGES & LIMITATIONS

- **Limited Translations:** Human-translated Swahili data was sparse, limiting the richness of fine-tuning.
- **Translation Ambiguity:** The Luo translation model exhibited lower average confidence, possibly due to less available training data or more linguistic complexity.
- **Domain-Specific Terms:** Technical education terms (e.g., “CBC,” “KUCCPS”) were inconsistently translated without custom glossaries.

FUTURE WORK

- Enhance translation quality by integrating expert reviews for low-confidence outputs, fine-tuning models with domain-specific content for low-resource languages, applying LDA for thematic analysis, exploring quality estimation models, and building interactive dashboards for real-time insights.
- These strategies collectively strengthen translation reliability, contextual relevance, and stakeholder engagement in low-resource language settings.