**Using Large Language Models and Transformers to Conserve Endangered Languages: The Case of Endangered Languages in Ethiopia**

Solomon T Tessema

**Author Note**

This research was conducted independently to explore the application of large language models and transformers for endangered language preservation. The author welcomes inquiries and collaborations from individuals or organizations interested in this area of research.

Correspondence regarding this article should be addressed to Solomon T. Tessema. Email: Solomon.tessema@outlook.com

**Using Large Language Models and Transformers to Conserve Endangered Languages: The Case of Endangered Languages in Ethiopia**

**Solomon T. Tessema**

**Introduction**

Language is an essential part of human culture and identity. Unfortunately, many of the world’s languages are endangered, with thousands facing extinction by the end of the 21st century (UNESCO, 2021). Ethiopia, a linguistically diverse country with over 80 languages, faces significant challenges in preserving its linguistic heritage. Many Ethiopian languages lack written resources, documentation, or educational tools, which exacerbates their vulnerability. Advances in artificial intelligence (AI), specifically large language models (LLMs) and transformer-based architectures, provide a novel opportunity to address this issue.

This research proposes the use of LLMs and transformers to conserve endangered languages in Ethiopia by developing tools and methodologies tailored to low-resource linguistic contexts. The study aims to bridge the gap between technology and linguistic preservation, offering actionable insights for revitalizing endangered languages globally.

**Problem Statement**

Endangered languages often lack the written documentation and linguistic resources required to sustain their usage. In Ethiopia, many languages are primarily oral, and existing NLP tools disproportionately favor resource-rich languages like English. Without intervention, these languages face extinction, taking with them invaluable cultural, historical, and social knowledge.

Transformers, a state-of-the-art approach in natural language processing (NLP), have demonstrated remarkable success in handling resource-intensive languages. However, their application to low-resource or endangered languages remains underexplored. This research addresses the critical question: **How can LLMs and transformers be adapted and applied to conserve and revitalize endangered Ethiopian languages effectively, despite minimal data availability?**

**Objectives**

1. **Primary Objective**:  
   To develop methodologies leveraging LLMs and transformers to support the conservation of endangered Ethiopian languages through data generation, analysis, and language modeling.
2. **Secondary Objectives**:
   * To design tools, such as translators and educational resources, for endangered Ethiopian languages.
   * To explore data augmentation and transfer learning techniques for low-resource languages.
   * To involve communities in co-designing culturally appropriate and useful AI-based solutions.

**Literature Review**

Recent advancements in NLP have been driven by transformer-based architectures like BERT, GPT, and T5 (Vaswani et al., 2017). While these models excel in tasks requiring large amounts of text data, their utility for low-resource languages remains limited (Conneau et al., 2020). Research on multilingual models, such as mBERT and XLM-R, has shown potential for cross-lingual transfer, but challenges persist when working with highly underrepresented languages (Ruder et al., 2021).

Efforts to preserve endangered languages have traditionally relied on manual documentation or community-driven initiatives (Hinton et al., 2018). However, such approaches struggle to scale. LLMs offer a scalable, technology-driven alternative, provided they are adapted to the unique constraints of low-resource languages. This study builds on the intersection of NLP and linguistics, aiming to create tools that preserve Ethiopian linguistic diversity.

**Methodology**

**Research Design**

This study will employ a mixed-methods approach, combining quantitative analysis of language data with qualitative insights from community engagement. The research design consists of the following phases:

1. **Data Collection**:
   * Work with native speakers and linguists to gather textual, audio, and video data for selected endangered Ethiopian languages.
   * Augment data using techniques such as back-translation, synthetic text generation, and cross-lingual transfer.
2. **Model Development**:
   * Fine-tune pre-trained multilingual transformers (e.g., mBERT, XLM-R, GPT-3) on the collected data.
   * Apply data augmentation techniques and low-resource adaptations such as meta-learning and few-shot learning.
3. **Tool Development**:
   * Develop applications like machine translation, speech recognition, and educational tools for endangered languages.
   * Design user-friendly interfaces for use by educators, linguists, and community members.
4. **Community Validation**:
   * Engage with native speakers and cultural stakeholders to validate the tools and ensure cultural appropriateness.
5. **Evaluation**:
   * Evaluate the models’ performance using metrics like BLEU for translation tasks, WER for speech recognition, and F1 score for classification tasks.

**Technologies and Tools**

1. **Data Collection and Annotation**:
   * Tools: ELAN (for linguistic annotation), Praat (for phonetic analysis), and manual transcription systems.
2. **Natural Language Processing Models**:
   * Pre-trained Models: mBERT, XLM-R, and GPT-3.
   * Frameworks: Hugging Face Transformers, PyTorch, TensorFlow.
3. **Data Augmentation**:
   * Techniques: Back-translation, cross-lingual transfer, and noise injection.
   * Tools: FastText for embedding generation and augmentation scripts.
4. **Applications and Deployment**:
   * Platforms: Flask or Streamlit for web applications, mobile frameworks for app development.

**Expected Outcomes**

1. **Preservation of Linguistic Data**:  
   A digitized repository of endangered Ethiopian languages, including annotated text and audio corpora.
2. **Functional Tools**:  
   Tools for automatic translation, speech-to-text conversion, and educational content generation in selected languages.
3. **Scalable Methodologies**:  
   Generalizable techniques for adapting LLMs to other low-resource or endangered languages.
4. **Community Empowerment**:  
   Increased engagement of local communities in language preservation efforts through accessible AI tools.

**Ethical Considerations**

1. **Community Consent**:  
   All data collection and tool development will be conducted with the informed consent of native speakers and stakeholders.
2. **Data Ownership**:  
   Linguistic data will remain the property of the communities from which it originates, with access granted through mutually agreed terms.
3. **Bias and Fairness**:  
   Efforts will be made to minimize biases in AI models and ensure inclusivity across dialects and cultural nuances.

**Conclusion**

This research addresses a pressing need to preserve endangered Ethiopian languages using cutting-edge AI technologies. By leveraging transformers and LLMs, the study will provide scalable, culturally sensitive tools for language conservation. Its interdisciplinary nature bridges linguistics, AI, and social sciences, contributing to both academic knowledge and real-world impact.

**References**

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451.

Hinton, L., Huss, L., & Roche, G. (Eds.). (2018). *The Routledge Handbook of Language Revitalization*. Routledge.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2021). Transfer learning in natural language processing. *The Oxford Handbook of Computational Linguistics 2nd Edition*, 704-721.

UNESCO. (2021). *Atlas of the World’s Languages in Danger*. Retrieved from <https://www.unesco.org>.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.