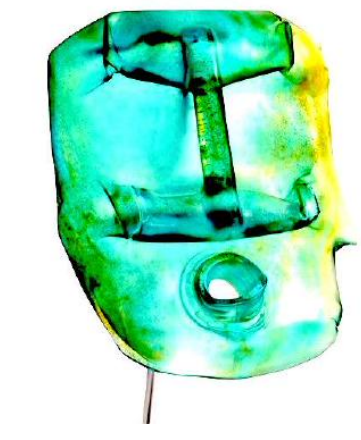


Towards Neural Machine Translation for Edoid Languages

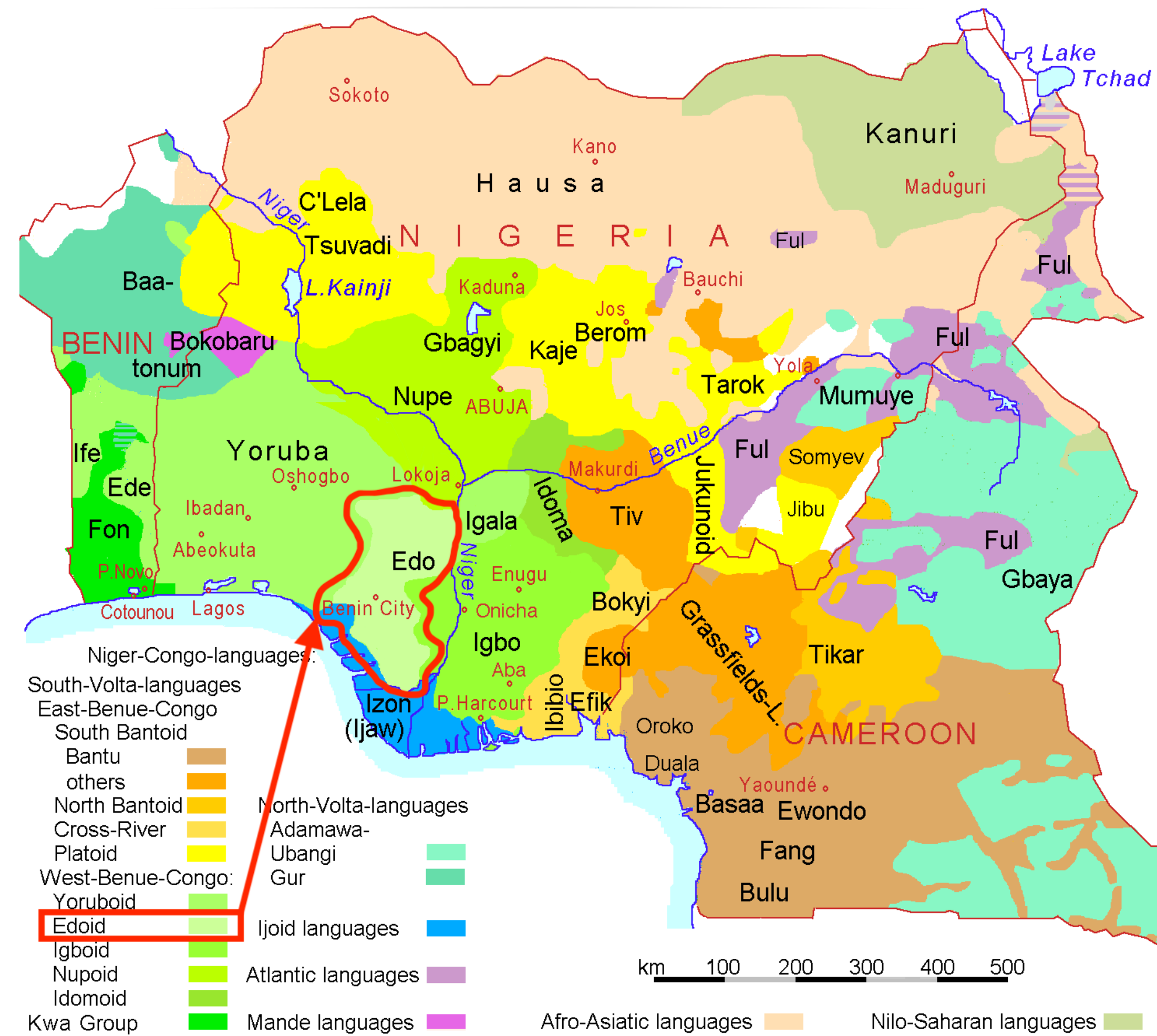
Iroro Fred Ọ̀nọ̀mẹ̀ Orife

Niger-Volta Language Technologies Institute



Overview

- Many minority Nigerian languages have resigned their previous prestige and purpose
- \exists ~30 Edoid languages with ~5M speakers
- Edoid language technology (LT) is non-existent



Readily accessible translation apps can \rightarrow

- Advance language literacy, documentation & preservation
- Facilitate good governance, national development
- Expedite economic & social & technological empowerment

This Work

We trained and evaluated baseline Neural Machine Translation (NMT) models for four widely spoken Edoid languages listed below with their classification

Language	Branch	# Speakers
Èdó	North-Central	~1.6 \rightarrow 2.3M
Ésán	North-Central	~630k
Urhobo	Southwestern	~1 \rightarrow 1.5M
Isoko	Southwestern	~660k

Dataset: JW300 dataset is a large-scale, parallel corpus comprising more than 300 languages of which 101 are African. JW300 text is drawn from the Watchtower and Awake! religious magazines by Jehovah's Witnesses (JW).

Models:

- Transformer models trained with the open-source, Python 3 machine translation toolkit **JoeyNMT**
- Training hardware was the free-tier configuration on Google Colaboratory, a single core Xeon CPU instance and a Tesla K80 GPU.

Experimental Setup

We trained baseline Transformer models for each language using:

- Byte Pair Encoding (BPE) subword tokenization
- Word-level tokenization.

For BPE, **4000 BPE tokens** were used based on ablation study by Martinus et al. for South African languages

Results

Per-language BLEU scores by BPE or word-level tokenization

Language	BPE		Word		Tokens	Sentences
	dev	test	dev	test		
Èdó	7.92	12.49	5.99	8.24	229,307	10,188
Ésán	4.94	6.25	3.39	5.30	87,025	4,128
Urhobo	15.91	28.82	11.80	22.39	519,981	25,610
Isoko	32.58	38.05	32.38	38.91	4,824,998	214,546

Analysis

- Urhobo and Isoko translation quality was **adequate** when reviewed by L1 speakers, correlating with higher BLEU scores.
- BPE tokenization \rightarrow **+37%** for Èdó and Ésán
- BPE tokenization \rightarrow **+32%** for Urhobo but none for Isoko.
- An **Isoko-sized** training set may required to achieve a satisfactory performance for Èdó and Ésán.
- A full ablation study with different (subword) tokenization approaches is needed to discover a more optimal representation.
- A **human evaluation study** and extensive error analysis will be crucial to better understand the linguistic features where NMT models under-perform.

References

- [Awobuluyi, 2016] Awobuluyi, d. (2016). Why We Should Develop Nigerian Languages. *Issues in Contemporary African Linguistics: A Festschrift for Oladele Awobuluyi*, 11:347.
- [Elugbe, 1989] Elugbe, B. O. (1989). *Comparative Edoid: phonology and lexicon*. University of Port Harcourt Press.
- [Martinus and Abbott, 2019] Martinus, L. and Abbott, J. Z. (2019). A Focus on Neural Machine Translation for African Languages. *CoRR*, abs/1906.05685.
- [Odoje, 2013] Odoje, C. (2013). Language Inequality: Machine Translation as the Bridging Bridge for African languages.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.



Paper (arXiv)
2003.10704



Code (Github)
Niger-Volta-LTI/edoid-nmt