YodiV3: NLP for Togolese Languages with Eyaa-Tom Dataset and the Lom Metric

Justin E. Bakoubolo[1], **Catherine Nana Nyaah Essuman**[1], **Messan Agbobli, PhD**[1,2],
Ahoefa Kansiwer, PhD[1,2], Kpona Sekpane Kpatika[1], Notou Your Timibe, PhD[1,3],
Agossou, PhD[1,4], Guedela Bakouya, PhD[1,2],
Bruno Koukoudjoe, PhD[1,2], Samuel Kossi Mawouena Afola[1] et al.[1]
[1]Umbaji
[2]University of Lomé
[3]University of Kara
[4]UCAO-UUT

Abstract

Most of the 40+ languages spoken in Togo are severely under-represented in Natural Language Processing (NLP) resources. We present **YodiV3**, a comprehensive approach to developing NLP for ten Togolese languages (plus two major lingua francas) covering machine translation, speech recognition, text-to-speech, and language identification. We introduce **Eyaa-Tom**, a new multi-domain parallel corpus (religious, healthcare, financial, etc.) for these languages. We also propose the **Lom metric**, a scoring framework to quantify the AI-readiness of each language in terms of available resources. Our experiments demonstrate that leveraging large pretrained models (e.g.NLLB for translation, MMS for speech) along with YodiV3 leads to significant improvements in low-resource translation and speech tasks. This work highlights the impact of integrating diverse data sources and pretrained models to bootstrap NLP for underserved languages, and outlines future steps for expanding coverage and capability.

1 Introduction

Togo is home to dozens of languages, including *Ewè*, *Kabyè*, *Tem* (Kotokoli), and many others spoken by millions collectively. However, most of these languages lack the data and tools needed for modern NLP applications. The scarcity of machine translation (MT) systems, speech technologies, and even basic linguistic resources (e.g.digital dictionaries) hinders information access and technology inclusion for the related communities. Recent advances in multilingual NLP have started to include a few Togolese languages—for instance, Facebook AI's *No Language Left Behind* (NLLB) project released MT models for Ewè and Kabyè [Team et al., 2022], and their *Massively Multilingual Speech* (MMS) initiative produced speech recognition and synthesis models covering those languages [Pratap et al., 2023]. Yet, these models often struggle on domain-specific content and other local languages not covered in training.

In this work, we address the above gaps by developing an end-to-end NLP pipeline for 10 key Togolese languages. First, we assembled a new dataset called **Eyaa-tom**[1] comprising parallel text (and audio) in multiple domains such as religious texts, healthcare, financial operations. Using this data, we train **YodiV3**, a multilingual model which supports translation, as well as speech recognition (ASR) and text-to-speech (TTS) components for selected languages.

Additionally, we introduce the **Lom metric** ("lom" meaning "score" in Nawdm) to quantify the state of language technology readiness for each language. The Lom metric aggregates the availability of core resources like a Bible or liturgical text, a dictionary, MT, ASR, TTS, language identification, and OCR models. This provides a quick overview of which languages are more digitally equipped and which need more attention.

We report experiments showing that YodiV3 improves translation quality and speech in low-resource settings by leveraging domain-specific data and as well as large fine-tunned models. We also present the Lom scores for the ten languages, revealing significant disparities: e.g., Ewè and Kabyè lead with much higher scores, whereas languages like Mina need more resources. Our results underscore the importance of targeted data collection and the integration of existing models to support "the last mile" languages. We conclude with our plans to incorporate more datasets and extend coverage to additional Togolese languages, further bridging the NLP divide.

- We present the first multi-domain NLP dataset for 10 Togolese languages, including

[1]*Eyaa-tom* means "People words'" in Kabyè.

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 143–149

20k+ parallel audio-text segments and additional annotated resources.

- We develop and evaluate baseline models for ASR, TTS, NMT, and language identification (LID) on these languages, demonstrating the feasibility of NLP with minimal resources.
- We introduce the *Lom* metric, which consolidates various resource indicators into a single score for each language, revealing disparities and guiding future work.

2 Related Work

Research on NLP for low-resource languages has gained momentum in recent years. Projects like Masakhane have leveraged participatory approaches to create translation datasets and models for numerous African languages [Nekoto et al., 2020]. For speech, the Mozilla Common Voice project released crowdsourced speech corpora for languages such as Swahili, Luganda, and Kabyle [Ardila et al., 2020], providing a foundation for ASR in some African languages. However, many languages of West Africa remain underrepresented in these initiatives.

Closer to our focus, Tonja et al. surveyed NLP for Ethiopian languages, highlighting the challenges of limited data and orthographic complexities. Our work is similar in spirit, but targets languages of Togo, which have distinct linguistic characteristics (many are Niger-Congo languages with tonal systems) and even fewer existing resources. To our knowledge, no comprehensive NLP dataset or benchmarks existed for the Togolese languages prior to this work. Also, a comprehensive examination of the current state of Natural Language Processing (NLP) in Kenya is presented in the paper titled "State of NLP in Kenya: A Survey" by Cynthia Jayne Amol et al. This survey delves into ongoing efforts in dataset creation, machine translation, sentiment analysis, and speech recognition for Kenyan languages such as Kiswahili, Dholuo, Kikuyu, and Luhya. Despite these advancements, the authors highlight that the development of NLP in Kenya remains constrained by limited resources and tools, leading to the under-representation of most Kenyan languages in digital spaces. The paper critically evaluates available datasets and existing NLP models, emphasizing the need for large-scale language models and better digital representation of Kenyan languages. Additionally, it analyzes key NLP applications tailored to local linguistic needs and explores the governance, policies, and regulations shaping the future of AI and NLP in Kenya, proposing a strategic roadmap to guide future research and development efforts.

In the speech domain, recent advances in self-supervised learning have shown promise for low-resource ASR; for example, wav2vec 2.0 pretraining [Baevski et al., 2020] can drastically reduce the data needed to train speech recognizers. We capitalize on such advances in our ASR models. For TTS, while classic autoregressive architectures like Tacotron 2 [Shen et al., 2018] produce high-quality speech, they can be impractical with limited data and compute. Non-autoregressive models such as Glow-TTS [Kim et al., 2020] offer faster and more data-efficient synthesis, which we explore in our setting. Our work ties these threads together by building a full pipeline (ASR $\rightarrow$ NMT $\rightarrow$ TTS, with LID) for multiple truly low-resource languages.

African Language Identification and Models. One foundational effort for African NLP is language identification (LID). Adebara et al. introduced **AfroLID**, a neural LID toolkit covering 517 African languages, which significantly outperforms previous LID tools on many African languages. Building on such resources, [Adebara et al., 2023] developed **SERENGETI**, a massively multilingual language model for 517 African languages. These works, led by the UBC NLP group, demonstrate the feasibility of broad-coverage models for African languages, including those of Togo. However, LID and language models alone do not directly provide translation or speech technology, which are our focus.

Masakhane and African NLP Initiatives. The Masakhane research community has spearheaded collaborative NLP projects for African languages. For example, the **Masakhane MT** project mobilized researchers to create machine translation datasets and baselines for numerous African languages [Orife et al., 2020]. Similarly, **MasakhaNER** provided high-quality named entity recognition data for ten African languages [Adelani et al., 2021] including Ewè. Our work is inspired by these community-driven efforts, and we extend the spirit of Masakhane to Togo by focusing on local languages and tasks (MT, ASR, TTS) that have immediate real-world application (e.g. healthcare information delivery).

Multilingual Translation and Speech by Big Tech. NLLB (*No Language Left Behind*) by Meta AI released MT models for 200+ languages, including Ewè and Kabyè, achieving unprecedented coverage [Team et al., 2022]. This demonstrated that low-resource languages can be handled within a single massive model given sufficient training data and compute. Meanwhile, Meta's **MMS** project (*Massively Multilingual Speech*) scaled speech technology (ASR, TTS, and spoken LID) to over 1,000 languages [Pratap et al., 2023] including the majority of the languages mentioned in this work. MMS included ASR/TTS models for Ewè and Kabyè, which we leverage as starting points. Our work differs in that, we build a new architecture based on the transformer architecture and incorporate some new neural quantization layers (to reduce costs) and adapt these large models on our curated Togolese datasets, focusing on specific domains (like religious or financial speech) where out-of-the-box NLLB/MMS performance may be suboptimal. We also address languages not covered by NLLB/MMS (e.g. Adja), using a combination of data augmentation and smaller neural models.

3 Dataset Creation: Eyaa-Tom

To enable training and evaluation of NLP models for Togolese languages, we built the **Eyaa-Tom** dataset. Eyaa-Tom consists of parallel text (and audio) in 10 local languages of Togo, with translations to French and English. The languages covered are: Ewè, Kabyè, Adja,Tem (Kotokoli), Moba, Lamba(Togo), Konkomba, Mina (Gen), Bassar, Nawdm. While some of them seem to be related, some dialects have evolved and tend to be now considered as languages, (i.e. Mina has its own alphabet and syntax despite the strong relationships with Ewè). The dataset present a clear separation between the dialects and languages with the intent of improving quality of service and further achieve research.

An overview of the dataset contents for each language can be seen in Table 1.

As shown in Table 1, each language has at least 2,000 parallel language pair sentences from religious texts with another language. These were obtained from publicly available translations. Many of these languages also had audio recordings collected (via the community contributions platform for specific and service phrases).We manually aligned a portion of this audio with the text to use for ASR, Speech translation, and TTS training namely. In addition to the religious domain, we collected parallel corpora in other domains for a subset of languages. For example, we are working with the community to translate financial and healthcare services sentences. Furthermore, we constructed a named entity list of over 1,500 Togolese personal names and locations, across several languages to support NER tasks.

The dataset was created through a combination of methods:

Community Contributions : A significant portion of the data was gathered via the *Umbaji Community Contribution Platform*—an online platform developed by the Umbaji community specifically to collect datasets for African languages. This platform enabled volunteers and native speakers to contribute text and audio in their local languages, ensuring a wide and authentic representation of linguistic data.

Field Research : Another major component of the dataset is collecting through fieldwork conducted by our linguists. They visit rural areas and work closely with local communities, including traditional chieftaincies, to gather texts, oral histories, poems, and other culturally significant materials in local languages. This approach ensures the inclusion of diverse linguistic features and contexts that might not be available in written form.

Collaboration with Mozilla Common Voice : We collaborated with Mozilla Common Voice, contributing over 2,000 validated voice samples for at least four of the languages in our dataset. This collaboration helped in expanding the spoken data component and aligning it with global standards for open-source language datasets.[Mozilla, 2025a][Mozilla, 2025d][Mozilla, 2025b][Mozilla, 2025c]

In this process, community contributors were actively engaged, with informed consent obtained prior to participation, and incentives provided to encourage contributions. Additionally, linguists were fairly compensated for their expertise, ensuring high-quality linguistic data. To foster inclusivity, we prioritized gender representation by intentionally recruiting a significant number of women, reinforcing our commitment to equitable data collection practices.

Overall, Eyaa-Tom provides a unique blend of

Language	Min. Religious (sentences or utterances)	Min. Other (utterances)	Min Total (utterances)
Ewè	2,000	2,961	4,961
Kabyè	2,000	2,316	4,316
Tem	2,000	2,316	4,316
Moba	2,000	2,483	2,483
Lamba(Togo)	2,000	2,316	4,316
Adja	2,000	2,316	4,316
Mina (Gen)	2,000	2,316	4,316
Bassar	2,000	2,316	4,316
Nawdm	2,000	3,410	5410
Konkomba	2,000	2,316	4,316

Table 1: Eyaa-Tom dataset statistics: number of parallel sentence pairs by domain for each Togolese language. "Religious" denotes primarily scripture and liturgical texts (often with corresponding audio). "Other" includes secular domains (healthcare, finance, public service) and additional named-entity lists.

domain-specific data tailored to real-world use cases in Togo. While modest in size compared to high-resource benchmarks, it is the first to offer such comprehensive parallel and spoken data across numerous Togolese languages. Data quality is ensured through community review and consistent orthography.

Integration with Hugging Face : Portions of the dataset are also hosted on Hugging Face, making it easily accessible to the broader machine learning and NLP research community.[Umbaji, 2025]

4 Model: YodiV3

We developed **YodiV3**, a multi-faceted model architecture that addresses both text and speech tasks for the ten languages. YodiV3 consists of several components:

Machine Translation (MT). YodiV3 includes an encoder-decoder neural translation model that can translate between each Togolese language and French/English.

Automatic Speech Recognition (ASR). We explore two approaches: (1) a standard *auto-regressive* Transformer model that generates translations one token at a time, and (2) a *non-auto-regressive* (NAR) model aimed at faster inference and which is less compute intensive. YodiV3's ASR component is currently capable of recognizing speech in at least the two main languages (Ewè, Kabyè) with reasonable accuracy (as shown in Section 6), and provides baseline models for the others that can be improved with more data.

Text-to-Speech (TTS). Similarly we built a TTS system namely for Ewè (since Ewè has more data for training). Additionally, we developed a voice cloning approach for "Togolese-accented" French and English: essentially, we fine-tuned an English/French TTS model on a small set of recordings from Togolese speakers, so that the synthesized French/English maintains the accent characteristics. This is useful for public service announcements where code-switching occurs. YodiV3's TTS module can thus speak in Ewè, Kabyè, and Togolese accented French/English. Extending TTS to the other languages is future work, likely requiring significantly more recording efforts.

Deployment The TTS and ASR models are constantly deployed and maintained through the community's Whatsapp AI chatbot and community contributions interface.

5 The Lom Metric

To quantify the state of NLP support for each language, we propose the **Lom metric**. This metric aggregates the presence of various foundational resources and technologies for a given language. We consider eight factors: (1) availability of a major **WOG** corpus (*Word of God*, i.e., a significant religious text such as the Bible), (2) a digital **Dictionary**/lexicon, (3) an **NMT** system (Neural Machine Translation), (4) an **ASR** system, (5) a **TTS** system, (6) a **Speech LID** model (SLID), (7) a **Text LID** model (TLID), and (8) an **OCR** system

for printed text. For each language, we assign 1 point if the resource is available (even in prototype form), or 0 if not. The Lom score (0–8) is the sum of points. Table 2 presents the status for the ten languages in our study.

From Table 2, we can see Ewè and Kabyè have the maximum Lom score (7/8), reflecting that they have a Bible, a published dictionary, and we have developed or leveraged MT, ASR, TTS, and LID, for it.

The Lom metric is a purely qualitative metric useful for guiding resource allocation: languages with very low scores need basic resource creation (data collection, orthography standardization), while those with mid-level scores might benefit from targeted projects (e.g., developing a TTS for Mina, or an ASR for Adja). It also provides an easy way to communicate to stakeholders or funders on how a language is positioned in terms of digital readiness.

6 Experiments and Preliminary Results

We conducted experiments to evaluate the performance of YodiV3 on translation and speech tasks, using the Eyaa-Tom data. Rather than exhaustively tuning the models, we focus on highlighting key results that demonstrate the effectiveness of our approach.

Machine Translation Quality. Improved performance on tasks such as NER for Togolese names as compared to all the models tested.

Speech-to-Text & Text-to-Speech Evaluation. Auto-regressive model show increased accuracy on many more tokens inputs, but overall, models incorporating neural non-auto-regressive quantization needs less compute but tend to be less precise for the initial tests.

Finally, our experiments reaffirm insights from prior work: multi-domain data is vital for performance. For instance, when evaluating Ewè→French translation specifically on health-related sentences, the model trained with our health subset achieved 30+ BLEU score, whereas a model trained only on religious text fell below 15 BLEU score on the same health test, demonstrating the importance of in-domain data. This aligns with observations by Team et al. [2022] that low-resource MT models benefit greatly from any domain-specific data available. Similarly, our use of pretrained models mirrors the success of Pratap et al. [2023] in showing that massive multilingual pretraining can jump-start speech technology for languages that lack sufficient data.

7 Conclusion and Future Work

We presented YodiV3 and the Eyaa-Tom dataset as steps toward inclusive NLP for Togo's languages. Our experiments show that combining carefully curated data with large pretrained models can yield workable translation and speech systems even for extremely low-resource languages. We also introduced the Lom metric, which revealed how unevenly resources are distributed across languages, providing a road-map for future resource development.

In the future, we plan to integrate additional existing datasets and models to further improve and expand YodiV3. This includes incorporating new releases from projects like Masakhane (e.g., any Togo-specific NLP datasets) or updates from the NLLB/MMS teams. We aim to extend the Eyaa-Tom corpus to more languages of Togo (such as Akebu, Ikposso, and others) to eventually cover all major language groups in the country.

Additionally, we will explore semi-supervised and active learning techniques to make the most of limited data, and continue to refine the Lom metric (possibly weighting the categories by importance or difficulty).

The AR model is based on a Transformers architecture similar to mBART. The NAR model uses a conditional masked language model (e.g., Levenshtein Transformer) which we train from scratch on our data. Both models are trained on the Eyaa-Tom parallel text. We found that fine-tuning the pretrained NLLB model greatly stabilizes training for the low-resource languages and yields higher translation quality.

These would allow us to publish our final work and compare it to exiting models and work.

Another important future direction is deployment: we intend to provide an API to work with local organizations to deploy YodiV3's translation and TTS capabilities in real-world settings (e.g., rural clinics or community radio). Such deployment will provide feedback to guide further research (for instance, identifying which errors are most critical to fix). We also foresee expanding our evaluation to include human evaluation with native speakers for translation quality and user acceptance of TTS and ASR. Future work would

Lang	WOG	Dict.	NMT	ASR	TTS	SLID	TLID	OCR	Lom
Ewè	Y	Y	Y	Y	Y	Y	Y	N	7
Kabyè	Y	Y	Y	Y	Y	Y	Y	N	7
Tem	Y	Y	Y	Y	Y	Y	Y	N	7
Adja	Y	Y	N	N	N	N	Y	N	3
Moba	Y	Y	Y	Y	Y	Y	Y	N	7
Lamba(Togo)	Y	-	N	N	N	N	Y	N	-
Konkomba	Y	Y	Y	Y	Y	Y	Y	N	7
Mina	Y	Y	N	N	N	N	Y	N	3
Bassar	Y	Y	Y	Y	Y	Y	Y	N	7
Nawdm	Y	Y	Y	Y	Y	Y	Y	N	7
Ifè	Y	Y	Y	Y	Y	Y	Y	N	7

Table 2: Lom metric evaluation for Togolese languages as of 2024. "Y" indicates the resource/technology is available (at least in experimental form), "N" indicates not yet available. WOG = presence of a significant religious text corpus; Dict. = digital dictionary or word list; NMT = machine translation; ASR = speech recognition; TTS = speech synthesis; SLID = spoken language identification; TLID = text language identification; OCR = optical character recognition. The final Lom score is out of 1.

Model	Translation		SNR/Classification		Speech Synthesis	
	Ewè	Kabyè	Ewè	Kabyè	Ewè	Kabyè
V1,zindi	—	—	0.97	—	—	—-
V2, B^2	—	—	0.1	0.33	—	—
V3, T	0.90	0.88	—	—	—	—
V3,ASR	—	—	0.50	0.5		—
V3,TTS	—	—	—	—	0.88	0.87

Table 3: Performance of multiple models across tasks and languages on the Eyaa-Tom dataset. This table is the history of Yodi and its performance. For each task it features accuracy. Translation is measured from and to french. It also features improvements done since the last publication. SNER stands for Spoken Name Entity Recognition. YodiV1,zindi is the final V1 version, not train on Eyaa-Tom however, developed owing to a competition on zindi and presents the best performances

also include a scientific comparison between cognate languages in Togo and similarities between them. By iteratively improving data, models, and evaluation metrics, we hope to steadily raise the Lom scores for all Togolese languages, ensuring none are left behind in the NLP revolution.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages. In *Proceedings of LREC 2022*, pages 2540–2547, Marseille, France.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. Serengeti: Massively multilingual language models for africa. *arXiv preprint*.

D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, S. Mayhew, I. A. Azime, S. H. Muhammad, C. C. Emezue, J. Nakatumba-Nabende, P. Ogayo, A. Diallo, A. Akinfaderin, T. Marengereke, and 2 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

R. Ardila, M. Branson, and K. Davis et al. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of LREC 2020*.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS 2020*.

J. Kim, S. Kim, J. Kong, and S. Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *NeurIPS 2020*.

Mozilla. 2025a. Common voice dataset for ajg.

Mozilla. 2025b. Common voice dataset for gej.

Mozilla. 2025c. Common voice dataset for kdh.

Mozilla. 2025d. Common voice dataset for nmz.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

I. Orife, D. I. Adelani, J. O. Alabi, K. Amponsah-Kaakyire, I. A. Azime, T. S. Bateesa, H. Buzaaba, C. Chukwuneke, A. Diallo, B. F. P. Dossou, R. Eiselen, C. C. Emezue, A. Faye, D. Gebreyohannes, T. R. Gwadabe, M. A. Hedderich, . O. Ishola, M. Katusiime, D. Klakow, and 13 others. 2020. Masakhane – machine translation for africa. *arXiv preprint*.

V. Pratap, E. Adjaye, T. D. Nguyen, A. Babu, T. Likhomanenko, P. Andrews, C. Fuegen, and R. Collobert. 2023. Mms: Scaling speech technology to 1,000+ languages. In *Proceedings of EMNLP 2023*.

J. Shen, R. Pang, and R. Weiss et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of ICASSP 2018*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

A. L. Tonja, T. D. Belay, and I. A. Azime et al. 2023. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. In *Proceedings of RAIL 2023*.

Umbaji. 2025. Eyaa-tom dataset.