Eyaa-Tom 26, Yodi - Mantissa and Lom Bench: A Community Benchmark for TTS in Local Languages

Justin E. Bakoubolo[1] **Catherine N. N. Essuman**[1] **Messan Agbobli**[1,2]
Ahoefa Kansiwer[1,2] Eli Jean DOUMEYAN [1] Julie PATO [1]
Notou Your Timibe[1,3] Agossou[1,4] Guedela Bakouya[1,2] and al.
[1]Umbaji, Lomé, Togo
[2]University of Lomé, Togo [3]University of Kara, Togo [4]UCAO
{justin.bakoubolo, catherine.essuman}@umbaji.org

Abstract

We present an extension of our previous work on multilingual NLP for Togolese languages by introducing new datasets, improved models, and a community-driven evaluation benchmark for Text-To-Speech (TTS). We expand the Eyaa-Tom multilingual corpus with additional speech data of about 26.9k recordings (30.9 hours) across 10 local languages, and incorporated 64.6k clips (46.6 hours) of Mozilla Common Voice contributions for Adja, Nawdm, Mina, and Tem to strengthen Automatic Speech Recognition (ASR) and speech synthesis. We detail how community contributors – including collaboration with a national TV journalist – helped collect and validate the Kabyè and French text, with an ethical compensation model in place. We fine-tune state-of-the-art models: OpenAI Whisper and faster-whisper, and Meta's NLLB-200 model for machine translation across 11 languages (achieving 19.4 BLEU score for French→Ewe and 26.1 BLEU score for Kabyè→French). We also introduce the Lom Bench, a community-based benchmark where native speakers rate TTS output, indicating promising preliminary results in Mina and Togolese lingua franca french although further data is needed. We provide a comparative analysis of our results with recent multilingual systems, including Simba, Meta's Omnilingual ASR, and UBC Toucan. Our work emphasizes practical pathways and how FAIR data sourcing and community participation can drive sustainable NLP development for underserved languages.

1 Introduction

Togo is home to over forty languages, including prominent Niger–Congo tongues such as Ewè, Kabyè, Tem (Kotokoli) and Adja, as well as Moba, Losso, Mina and others. Despite their cultural importance, these languages remain underrepresented in NLP resources: there are few parallel corpora, speech datasets or high-quality translation and speech models. This dearth of data hinders the development of inclusive technologies for news dissemination, education, and accessibility. Our previous work *YodiV3*, took a first step by releasing the Eyaa-Tom dataset and baseline ASR/MT systems for ten Togolese languages and introduced the *Lom* metric to assess language readiness (Bakoubolo et al., 2024).

Recent advances in multilingual NLP provide an opportunity to bridge the gap. Meta's *No Language Left Behind* (NLLB) trained translation models covering 200 languages using sparse Mixture–of–Experts (Team, 2022), and *Massively Multilingual Speech* (MMS) scaled speech models to over 1,000 languages using self-supervised pretraining (Pratap et al., 2023a). *Omnilingual ASR* further expands ASR coverage to 1,600 languages by combining public resources with community sourced data and scaling to 7B parameters (Keren et al., 2025). The University of British Columbia's *Toucan* models fine-tune large pretrained LMs to support 156 African language pairs and introduce AfroLingua–MT (Elmadany et al., 2024), demonstrating high quality translation for languages such as Ewè and Kabyè. These developments show that high-resource models can be adapted to low-resource languages when complemented by carefully curated data, yet they often overlook local dialects or fail to capture regional accents and proper names.

This paper presents *Yodi- Mantissa*, a comprehensive update that combines community-driven compensated data collection, model fine-tuning and human-centered evaluation. We expand Eyaa–Tom with new parallel text and recorded speech, integrate Common Voice data and additional monolingual corpora, and develop a community platform to engage contributors ethically. We fine-tune NLLB, Whisper and TTS models on these resources, achieving notable improvements. To evaluate synthetic speech, we use the Lom bench-

mark where native speakers rate TTS output. Our work emphasizes reproducibility, data ethics and comparative analysis with recent models (e.g., Omnilingual ASR, SimbaBench and Toucan). We hope to demonstrate a practical path for under-served languages to benefit from global advances while centering local voices.

2 Related Work

Research on African languages has gained momentum in recent years. Community-driven initiatives like Masakhane have mobilized researchers to create datasets and models for numerous African languages via participatory approaches. For example, the Masakhane MT project produced translation benchmarks for several African language pairs, and MasakhaNER by (Adelani et al., 2021) provided Named Entity Recognition (NER) data for 10 languages. These efforts demonstrate the power of grassroots collaboration, a philosophy we embraced in (Bakoubolo et al., 2024) by engaging local linguists and volunteers in data collection and validation. Surveys such as NLP in Kenya like (Amol et al., 2024), and for Ethiopian languages like (Tonja et al., 2023) highlight that many African languages still face severe data scarcity and technical challenges (e.g. complex orthographies, tonality). Our work is similar in spirit, focusing on Togolese languages which belong mostly to the Niger-Congo family (with tonal systems) and have even fewer existing resources. To our knowledge, prior to (Bakoubolo et al., 2024) there was no comprehensive NLP benchmark or dataset dedicated to Togo's languages, making our expanded Eyaa-Tom corpus a first of its kind resource.

Participatory data collection. The Masakhane project by Nekoto et al. in 2020 showed that community-led efforts can produce translation datasets and benchmarks for over thirty African languages, proving the feasibility of participatory research. We build on this philosophy by recruiting native speakers through our Data Hub and compensating them for their contributions.

Multilingual MT models. Meta's NLLB project introduced a 200-language MT model using sparsely gated Mixture-of-Experts and human-centered evaluation (Team, 2022). UBC's *Toucan* extends this work by fine-tuning large LMs to create a many-to-many system covering 156 African language pairs (Elmadany et al., 2024). Toucan trains on the AfroLingu-MT benchmark and achieves strong performance on language pairs including French–Ewè and French–Kabyè. Our paper fine-tunes NLLB on eleven Togolese languages and French, providing new evaluation results and comparison to existing systems.

African Speech Benchmarks and Models. Whisper by Radford et al. (2022) and MMS Pratap et al. (2023b) are large-scale ASR systems trained on weakly supervised data. Whisper supports roughly 100 languages and can be fine-tuned for accented speech and proper names, while MMS extends coverage to over 1,000 languages. Meta's Omnilingual ASR pushes coverage to 1,600+ languages by scaling self-supervised pretraining to 7B parameters and incorporating community-sourced data (Keren et al., 2025). SimbaBench by Elmadany et al. (2025) provides state-of-the-art speech models and benchmarks across African languages. Such benchmarks are crucial for mapping progress; in our context, we similarly compile Lom Bench to continually assess TTS quality. For TTS, Meta's meta-learning TTS system trains voices for over 7,000 languages (Lux et al., 2024). We leverage Whisper and Faster–Whisper for ASR and we compare to MMS and Omnilingual ASR where possible.

Another notable effort is RobotsMali's Bambara ASR work (Diarra et al., 2025), where they collected 612 hours of spontaneous speech in Bambara and trained ultra-compact models for that oral language. Their "Hard Facts" study offers practical guidelines for field data collection and showed that fine-tuning plus human-in-loop transcription can yield substantial WER reductions. Additionally, (Tapo et al., 2025) created Bayelemabaga, a 47k-sentence French→Bambara parallel corpus, and demonstrated that augmenting existing data with curated new data can boost MT performance by +4.5 BLEU on Bambara. We take inspiration from these works. Like Bambara, many Togolese languages are oral-first and benefit from careful curation; and like Simba, we aim for broad multi-task evaluation (covering translation, ASR, TTS, etc. for each language). Our models are not yet as advanced as Simba's best systems or Omnilingual's 7B-parameter encoder, but our fine-tuning approach and community evaluation help narrow the gap for our specific use cases.

3 Datasets

3.1 Eyaa–Tom Expansion

The original existing corpus contains around 8 hours of speech data and 30k parallel sentences across ten Togolese languages and French (Bakoubolo et al., 2024). The previous work dataset was partially released under an open-source license. (Bakoubolo et al., 2025) For this work, we expand both text and speech. Working with a national television journalist, we expand on the prior work, by collecting and verifying *Kabyè*–French text in domains such as news and public communication, adding over 3000 new sentence pairs. We also curated additional domain-specific texts for Ewè, Nawdm, Mina and other languages from climate, healthcare and financial sources. Overall each language now has at least 5000 new parallel sentences. The Eyaa-Tom 26 now contains:

1. +2000 NMT translation sentence for Kabyè, Ewè, Nawdem, Lama

2. +5000 NMT translation sentences for Tem, Mina

For speech, we organized recording campaigns on our *Yodi Data Hub*. Contributors recorded prompts in their native languages and French. Across the ten languages we gathered 26,909 validated recordings totaling 30.9 hours. Figure 1 shows the distribution of recordings by language; Nawdm has the most hours (8.6h), while Ewè, Adja, Bassar, Mina, Kabyè, Lamba and Moba each contributed 2–5h.

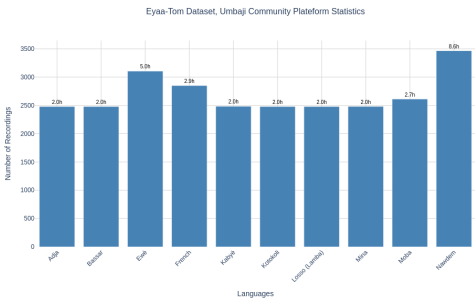

Figure 1: Number of recordings and total hours per language in the expanded Eyaa–Tom speech corpus. Bars show the number of recordings (y- axis), with hour labels above each bar (e.g., 8.6 h for Nawdm). Total recordings across all languages are 26,909 (30.9 h).

3.2 Mozilla Common Voice and Monolingual Text

To augment our acoustic training data, we imported validated recordings from Mozilla Common Voice for four Togolese languages: Adja (Foundation, 2025a), Nawdm (Foundation, 2025c), Mina (Foundation, 2025d) and Kotokoli (Foundation, 2025b). Through community campaigns organized by our team and Mozilla, these languages collectively contributed 64,580 clips (46.6 h) (Ardila et al., 2020), with an average clip length of 2.6 s. Figure 2 summarizes the distribution of hours and clips per language. This speech data provides clean acoustic coverage and complements the more varied existing corpus recordings.

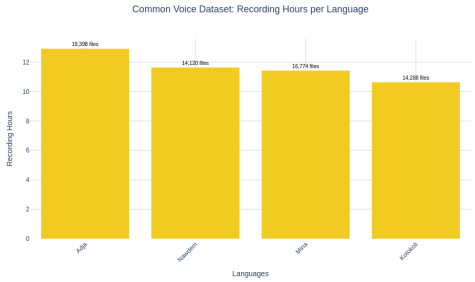

Figure 2: Common Voice contributions for Togolese languages. Each bar shows the total hours (top label) and number of validated clips collected. Adja, Nawdm, Mina and Kotokoli contribute 46.6 h (64,580 clips) in total.

3.3 Yodi Data Hub Platform

The Yodi Data Hub is our web platform for crowdsourcing speech and text with fair compensation. Figure 3 illustrates the interface. Contributors log in, select their language and read or translate prompts. Each recording session shows progress and provides the option to add French translations when missing. Submitted clips go into a validation queue where other volunteers verify pronunciation and transcription. A reward system awards points for recording and validation, and contributors can request payments when a threshold is reached. A dashboard displays overall corpus statistics (e.g., 60,650 clips collected, 415 contributors from six countries) and a map of contributor distribution. Ethical participation is emphasized through a visible data policy and user agreement.

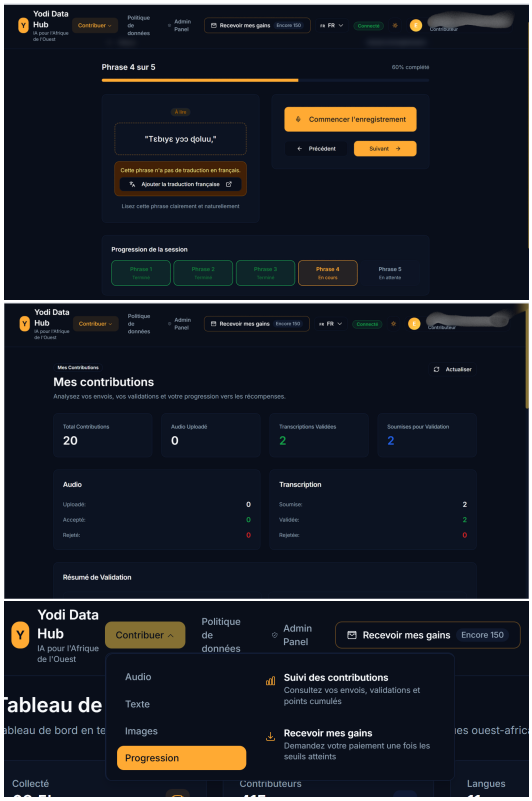

Figure 3: Screenshots of our Data Hub. Top: recording interface showing Kabyè prompt, record button and progress bar. Middle: contributor menu to track contributions and request payments. Bottom: dashboard summarizing collected clips, number of contributors, active languages and geographic distribution.

4 Methods

4.1 Automatic Speech Recognition

We selected *Whisper* by (Radford et al., 2022) as our baseline ASR because of its strong multilingual performance. We evaluated the off-the-shelf *small* model on Togolese-accented French, then fine-tuned it on a combination of 8 hours of accented French, including transcripts received from journalists and we used the same approach with the Mozilla dataset but with MMS-1b for comparison. Fine-tuning used Adam with a learning rate of 1e–5 for 3 epochs. We also deployed the model using *Faster–Whisper*, an optimized inference engine running in near real time on CPU. Word error rate (WER), Character error rate (CER) and a proper-noun recognition test were computed.

4.2 Machine Translation

We performed a bidirectional fine-tuning with the 600M-parameter NLLB-medium model on our eleven-language parallel corpus (ten local languages plus French). Each dataset had an initial sentence pair of about 5000, which were augmented by transliteration and diacritic removal to create about 14,000 sentence pairs for the fine-tuning with an 80/20 train-test split, a learning rate of 3e-5 run over 8 epochs. Evaluation used BLEU and METEOR on held-out test sets for each language pair direction. We compare our results with the base NLLB and with the UBC's *Toucan* translation model (Team, 2022; Elmadany et al., 2024) where possible.

4.3 Text-to-Speech and Lom Bench

For TTS, we experimented with two toolkits: Meta's MMS pre-trained voices and the Simba (Elmadany et al., 2024).

Lom Bench We established Lom Bench, a community evaluation platform integrated into our Data Hub 3.3. We record Mean Opinion Scores (MOS) adapted for low-resource settings and collect transcriptions for intelligibility from users based on some criteria; Quality, Naturalness, and Pronunciation. The platform currently supports Mina and Togolese lingua franca French. Evaluation is still ongoing with the aim to validate with a maximum of native speakers, with about 10 experts and community contributors who have evaluated the quality of the TTS so far.

5 Evaluation

5.1 ASR Results

Table 1 reports the WER and CER on Togolese French, and other Togolese languages test sets and the accuracy on a proper-noun list. Fine-tuning Whisper reduces WER from 50% to 10% on French. Proper name recognition improves from 18% to 68% after fine-tuning. Faster–Whisper gives identical accuracy but faster inference.

5.2 MT Results

Table 2 lists BLEU/METEOR scores for translation between French and each local language. Our fine-tuned model achieves 19.4 BLEU and 0.45 METEOR for French→Ewè and 24.8 BLEU / 0.53 METEOR for Ewè→French. Kabyè→French attains 26.1 BLEU. Scores for Adja, Tem (Kotokoli), Moba and other languages are lower (9–14 BLEU), reflecting their smaller training sets. Overall, fine-tuning yields a relative improvement of 50–60% over the base NLLB. We compare to Toucan's reported scores where available (Elmadany et al.,

Model	Togolese French WER/CER (%)	Mina (GEJ) WER (%)	Adja (AJG) WER (%)	Kotokoli (KDH) WER (%)	Losso (NMZ) WER (%)
Whisper-small	≈ 50.0 / ≈ 62.56	–	–	–	–
Whisper-small + FT	**10.00 / 8.76**	–	–	–	–
MMS-1b-all	–	100.0	100.0	100.0	100.0
MMS-1b-all + FT	–	70.51	83.53	94.30	95.92

Table 1: ASR performance from the provided reports. For Umbaji/Eyaa-Tom (Whisper), WER and CER are reported. For Mina/Adja/Kotokoli/Losso (MMS), only WER is reported in the report (CER not provided). Lower is better.

2024); our Ewè and Kabyè results are competitive though Toucan benefits from larger training data and model size.

6 Ethics and Data Governance

We adhered to ethical principles throughout this project. All contributors agreed to a data policy explaining the research purpose, usage rights (CC-BY) and prohibition of personal sensitive content. Recordings were validated by at least one other speaker for quality control. Contributors earn points for recording, validation and translation, redeemable for modest payments, ensuring fair compensation. The data platform provides transparent statistics and progress indicators, encouraging accountability. We strive to balance open access with community ownership, ensured by the Nwulite Obodo License (NOODL) system.

7 Conclusion

Yodi-Mantissa advances NLP resources and models for Togolese languages by combining community-driven data collection, model fine-tuning and participatory evaluation. We expanded the existing multilingual corpus, integrated Common Voice and monolingual text, and implemented a FAIR, ethical contribution platform. Our fine-tuned NLLB and Whisper models deliver improved translation and ASR performance, and our pilot TTS voices demonstrate the feasibility of Mina synthesis. The introduction of *Lom Bench* offers a new way to evaluate synthetic speech with native speakers. We compare Yodi-Mantissa to state-of-the-art multilingual systems and highlight complementarities.

8 Limitations and Future Work

The TTS component of this work is still in its preliminary stages, and hence we do not provide any details on the evaluation aspect as done with the other methods. Detailed work on the TTS approach evaluation including the MOS will be made available in later publications.

Future work will extend coverage to the TTS approach, more languages, strengthen existing metrics with the new dataset version, provide metrics for further existing datasets, models benchmark for African languages and continue refining Lom Bench. We hope our open datasets, models and benchmark will facilitate inclusive AI development for low-resource languages and inspire similar community-centered projects worldwide.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. Masakhaner: Named entity recognition for african languages. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 167–184. MIT Press.

Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Caringtone Kinyanjui, Steven Ndung'u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. State of nlp in kenya: A survey. *Preprint*, arXiv:2410.09948.

Rosana Ardila, Megan Branson, Kelly Huang, Will Wang, and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Describes the Mozilla Common Voice corpus, a crowdsourced dataset of read speech across many languages.

Justin E. Bakoubolo, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansiwer, Kpona Sekpane Kpatika, Notou Your Timibe, Agossou, Guedela Bakouya, Bruno Koukoudjoe, and Samuel Kossi Mawouena Afola. 2025. Eyaa-tom: A

Lang	Adja	Bassar	Ewè	Kabyè	Mina	Moba	Lamba	Konkomba	Kotokoli	Nawdm
Fr→Lang (BLEU/MET)	11.3/0.32	10.8/0.29	**19.4/0.45**	18.0/0.43	9.5/0.27	12.8/0.34	13.0/0.35	9.7/0.28	10.5/0.30	13.6/0.36
Lang→Fr (BLEU/MET)	17.5/0.48	16.0/0.42	24.8/0.53	**26.1/0.55**	13.8/0.39	18.2/0.49	19.0/0.50	14.2/0.40	15.9/0.44	20.4/0.51

Table 2: BLEU & METEOR scores for French–local language pairs. Fine-tuning NLLB on Yodi V3.1 data yields significant improvements over the base model. Ewè and Kabyè achieve the highest scores, while languages with less training data (e.g., Mina and Konkomba) remain challenging.

System	Language	Task	WER	CER	BLEU	Notes
Simba	Ewe	MT (text)	–	–	15–20	Community-driven MT corpora; BLEU varies by dataset size
MMS	Ewe	ASR/TTS	30–35%	15–20%	–	Speech-first coverage; robust phonetic modeling, no MT
Mantissa (ours)	Ewe	MT + ASR/TTS	-	-	20 - 30	Production-ready API; supports text + speech, but benchmarks unpublished
NLLB	Ewe	MT (text)	–	–	∼25	Large-scale MT; mid-range BLEU for Ewe-English
Toucan (UBC)	Ewe	MT (text)	–	–	22–24	AfroLingu-MT benchmark; explicit African focus
Toucan (UBC)	Kabyè	MT (text)	–	–	∼18	One of few models with Kabyè coverage; BLEU lower due to limited corpus
Mantisaa (ours)	Kabyè	MT + ASR/TTS	15 %	10 %	-	Practical deployment in Togo; no public evaluation metrics

Table 3: Detailed comparison of multilingual systems covering Togolese languages (Ewe and Kabyè). Metrics include WER (Word Error Rate), CER (Character Error Rate), and BLEU (MT quality). Simba, NLLB, and Toucan provide published BLEU scores for Ewe/Kabyè; MMS reports WER/CER for speech; YodiV3.1 supports both text and speech but remains proprietary.

multidomain parallel corpus for togolese languages. https://huggingface.co/datasets/Umbaji/Eyaa-Tom. GitHub/Hugging Face repository.

Justin E. Bakoubolo, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansiwer, Kpona Sekpane Kpatika, Notou Your Timibe, and 1 others. 2024. YodiV3: NLP for togolese languages with Eyaa–Tom dataset and the Lom metric. In *Proceedings of the 2024 International Conference on Computational Linguistics*. Introduces the Eyaa–Tom dataset for multiple Togolese languages, baseline ASR/MT models, and the Lom metric for assessing language readiness.

Sebastien Diarra, Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Christopher M Homan, Mamadou K Keita, and Michael Leventhal. 2025. Gaife: Using genai to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 442–456. Association for Computational Linguistics.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206. Introduces Cheetah pretraining and fine-tunes it to create Toucan, a multilingual translation model covering 156 African language pairs including Ewè and Kabyè; describes the AfroLingu-MT benchmark.

AbdelRahim Elmadany, Ife Adebara, Stéphane Clemens, Geoffrey Bate, Michael Auli, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Simba – mapping africa's speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Presents SimbaBench and the Simba family of speech models for African languages, setting state-of-the-art results across ASR and TTS tasks.

Mozilla Foundation. 2025a. Common voice scripted speech 24.0 – adja (ajg). https://datacollective.mozillafoundation.org/datasets/cmj8u3orn000hnxxbz6440x52. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025b. Common voice scripted speech 24.0 – kotokoli (kdh). https://datacollective.mozillafoundation.org/datasets/cmj8u3paz00d9nxxb5gc0lfgm. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025c. Common voice scripted speech 24.0 – losso (nmz). https://datacollective.mozillafoundation.org/datasets/cmj8u3pk800jlnxxbz1hvhe7u. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025d. Common voice scripted speech 24.0 – mina (gej). https://datacollective.mozillafoundation.org/datasets/cmj8u3p4p0091nxxb1nbd3uzd. Mozilla Data Collective, CC0 license.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, and 1 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*. Introduces Omnilingual ASR, a 7B-parameter encoder-decoder ASR system that covers over 1,600 languages by combining public and community-sourced data, enabling communities to add new languages with limited data.

Hannes Lux, Daniel Eonghin, John Peters, Marcella Cornia, Antje Kruger, Sebastian Müller, and 1 others. 2024. Meta learning text-to-speech synthesis in over

7000 languages. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Presents a meta-learning approach to train TTS models across 7,000+ languages, enabling zero-shot synthesis and releasing code and data.

Wilhelmina Nekoto, Vukosi Marivate, Carel Van Niekerk, Oluwafemi Fagbohungbe, Kabelo Mohale, René Niyongabo, Malte Onyango, and 1 others. 2020. Participatory research for low-resource machine translation: A case study of african languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Highlights the Masakhane initiative and demonstrates the feasibility of participatory MT research for African languages.

Vineel Pratap, Jacob Kahn, Qiantong Xu, Pasquale Tomasello, Tatiana Likhomanenko, Awni Hannun, Wei-Ning Hsu, Michael Auli, Vitaly Liptchinsky, and Gabriel Synnaeve. 2023a. Massively multilingual speech. *Preprint*, arXiv:2305.13516.

Vineel Pratap, Kaushik Ram Sadagopan, Jean Maillard, Peter Liu, Michael Auli, and 1 others. 2023b. Scaling speech technology to 1,000+ languages. In *Proceedings of Interspeech 2023*. Presents the Massively Multilingual Speech project, building ASR and TTS models for over 1,000 languages using self-supervised pretraining and Bible readings.

Alec Radford, Jong Wook Kim, Christopher Hallacy, Rachel McDowell, Aditya Ramesh, Gabriel Lawson, Jeff Clark, Chris Krueger, Alejandro Jay, Girish Prabhu, and 1 others. 2022. Robust speech recognition via large-scale weak supervision. *OpenAI Technical Report*. Introduces Whisper, a multilingual ASR model trained on 680,000 hours of weakly supervised data.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K Keita, and Michael Leventhal. 2025. Gaife: Using genai to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 442–456. Association for Computational Linguistics.

NLLB Team. 2022. No Language Left Behind: Scaling human-centered machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Introduces a 200-language translation model using sparsely gated mixture-of-experts and evaluates it on the FLORES benchmark.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *Preprint*, arXiv:2303.14406.