Sukuma Speech-to-Text

Zephania Reuben Isack Odero

Technical report · August 2026 · Research preview

Quality SKU: https://huggingface.co/nileagi/nileagi-suk-stt

Lite SKU: https://huggingface.co/nileagi/nileagi-suk-stt-lite

Abstract

This report describes NileAGI's two public Sukuma speech-to-text systems, nileagi/nileagi-suk-stt (quality) and `nileagi/nileagi-suk-stt-lite` (lite), both of which map Sukuma audio directly to Sukuma text in a Latin orthography that preserves vowel macrons. Supervision comes from CTC-aligned read-aloud speech, packed into windows of 0.5–30 s and split along whole document groups so that no test segment shares a source with training data (16,792 / 636 / 94 windows; ∼98.1 / 3.6 / 0.5 h across train, validation, and test respectively). On the held-out document-group test set, the quality SKU reaches a word error rate of **21.3%** and a character error rate of **4.9%**, while the lite SKU reaches **26.1%** WER and **6.4%** CER. These figures reflect a research preview evaluated on a single literary speaker rather than a general-purpose Sukuma ASR product, and neither system is intended to recognise English or Swahili speech.

1 Introduction

Sukuma (Kisukuma, ISO 639-3 `suk`) is a Bantu language spoken across northern Tanzania, and public speech technology built specifically for it remains scarce. NileAGI's STT SKUs are designed around a single, well-defined task: converting Sukuma speech into Sukuma text, using the same macron-based orthography that already underpins NileAGI's Swahili-to-Sukuma translation systems.

We publish two operating points that trade transcription accuracy against deployment footprint, summarised in Table 1.

Table 1: Public STT SKUs.

Role	Hub id	Selection rule
Quality	`nileagi/nileagi-suk-stt`	lowest test WER
Lite	`nileagi/nileagi-suk-stt-lite`	same recipe, smaller footprint

Both models live in the collection https://huggingface.co/collections/nileagi/nileagi-suk, and this document is concerned only with these two ASR systems rather than with the broader NileAGI Sukuma toolchain.

2 Task

Direction. The systems are trained for one direction only, Sukuma speech to Sukuma text, and English or Swahili ASR falls outside their intended scope. Anyone working from those source languages should first pass audio through a separate recogniser suited to that language, then route the resulting text through NileAGI's machine translation systems if Sukuma output is required.

Orthography. Model output follows the Latin script with macrons (ā ē ī ō ū) that NileAGI uses across its Sukuma tools. Other written conventions for Sukuma exist in practice, and this release does not attempt to normalise between them.

Register. All training audio consists of literary read-aloud speech from a single speaker, so conversational speech, telephone-quality audio, and high-background-noise recordings all represent meaningful domain shift away from the conditions the models were trained on.

3 Data and alignment

The underlying units are transcribed Sukuma utterances paired with their corresponding audio. Each unit is force-aligned using a multilingual CTC aligner capable of absorbing unmatched leading speech, such as a short title sting at the start of a recording, without stretching the final unit artificially to the end of the file. Only alignments explicitly tagged as CTC keep are retained for training, and we deliberately avoid proportional time-based splitting in favour of the document-group approach described below.

Kept units are then packed into ASR windows of 0.5–30 s so that training examples match the chunk lengths typical encoder–decoder architectures expect at inference time.

Table 2: Training inventory after packing (0.5–30 s windows).

Split	Windows	Hours	Role
train	16,792	98.1	parameter updates
validation	636	3.6	checkpoint selection
test	94	0.5	headline WER/CER

Splits hold out entire document groups, so a test window never has a neighbouring segment from the same source document sitting in the training set. This protects against inflated scores from near-duplicate content, but it comes at the cost of a fairly small test set, one large enough to rank the two SKUs against each other but not large enough to support a fine-grained error taxonomy.

4 Models

Both SKUs are encoder–decoder sequence-to-sequence ASR systems fine-tuned on the packed windows described in Section 3. The quality SKU is the strongest performer on held-out test WER among the candidates evaluated, while the lite SKU is the smallest model that remained competitive on the same split. Training ran for three epochs using mixed precision on a single GPU, with features extracted on the fly rather than cached to disk. The decode language tag is

set to the closest language supported by the underlying base stack, though the supervision text itself remains Sukuma throughout.

We do not publish internal candidate scores beyond the headline numbers for the two winning SKUs reported in Section 5.

5 Evaluation

Word error rate (WER) and character error rate (CER) on the document-group test windows serve as the primary evaluation metrics, and lower values indicate better performance on both.

Table 3: Held-out document-group test (94 windows).

SKU	WER	CER
Quality (`nileagi-suk-stt`)	**21.3%**	**4.9%**
Lite (`nileagi-suk-stt-lite`)	26.1%	6.4%

A WER in the low-to-mid twenties on this literary, single-speaker evaluation set is genuinely useful for research transcription work, though it should not be treated as a substitute for a human transcript when audio is noisy or conversational in nature.

6 Inference

Audio inputs should be 16 kHz mono, and clips under 30 s are preferred; longer files should be chunked before being passed to the model. A minimal example using `transformers` is shown in Listing 1.

Listing 1: Inference example (Python)

```python
from transformers import pipeline

stt = pipeline(
    "automatic-speech-recognition",
    model="nileagi/nileagi-suk-stt",
    chunk_length_s=30,
    ignore_warning=True,
)
print(stt("sukuma.wav")["text"])
```

Sample WAV files ship in each Hub repository under `samples/` for anyone who wants to verify their setup before running on their own audio.

6.1 Intended use

- Research transcription of Sukuma read-aloud speech.
- Offline batch ASR for literary or narrative Sukuma content.

6.2 Out of scope

- English or Swahili ASR.
- Conversational, telephone, or high-noise field audio.
- Legal or medical transcripts.
- Commercial products without a licence from NileAGI.

6.3 Limitations

1. **Preview scores.** Test WER in the twenties is usable for research purposes, but it does not represent production-ready performance across arbitrary recording conditions.
2. **Single speaker and register.** Because all training data comes from one speaker reading literary text, meaningful domain shift toward conversational speech should be expected.
3. **Test size.** With only 94 document-group windows, the headline numbers are reliable enough to prefer quality over lite, but they do not provide a precise measure of the gap to human parity.
4. **Orthography.** The macron-based spelling conventions used here may differ from other established Sukuma writing practices.
5. **Names.** Rare word stems and proper names are often approximated rather than transcribed exactly.

6.4 Licence

Table 4: Licences. The data licence is not the licence of the weights.

Artifact	Licence
`nileagi-suk-stt` and `nileagi-suk-stt-lite` weights	CC BY-NC-SA 4.0
Aligned training supervision	NOODL 1.0

7 Conclusion

NileAGI releases two Sukuma ASR SKUs trained on CTC-aligned literary speech and selected according to document-group test WER, reaching **21.3%** for the quality model and **26.1%** for the lite model. The quality SKU is the better choice whenever transcript accuracy matters most, while the lite SKU suits deployments where a smaller footprint takes priority. Both remain research previews evaluated on a single read-aloud speaker rather than general-purpose Sukuma speech products, and users should weigh that context when deciding how to apply them.

8 Citation

Listing 2: Citation (BibTeX)

```bibtex
@misc{nileagi-suk-stt-techreport-2026,
  title        = {NileAGI Sukuma Speech-to-Text: Technical Report},
  author       = {Zephania Reuben and Isack Odero},
  year         = {2026},
  howpublished = {Hugging Face},
  url          = {https://huggingface.co/nileagi/nileagi-suk-stt}
}
```

References

[1] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.* ICML, 2006. — the CTC objective used for force-aligning transcripts to audio in Section 3.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. *Attention Is All You Need.* NeurIPS, 2017. — background on the encoder–decoder sequence-to-sequence architecture class used by both SKUs (Section 4).

[3] T. Wolf et al. *Transformers: State-of-the-Art Natural Language Processing.* EMNLP: System Demonstrations, 2020. — the `transformers` library used for training and the inference example in Listing 1.

[4] A. Baevski, H. Zhou, A. Mohamed, M. Auli. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* NeurIPS, 2020. — background on the self-supervised speech encoder family underlying modern encoder–decoder ASR pipelines such as the one fine-tuned here.

[5] V. Pratap et al. *Scaling Speech Technology to 1,000+ Languages.* arXiv: 2305.13516, 2023 (JMLR, 2024). — background on multilingual CTC-based forced alignment at scale, the general technique used to align Sukuma transcripts to audio (Section 3).

[6] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger. *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.* Interspeech, 2017. — a widely used reference implementation of the CTC-based forced-alignment step described in Section 3.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision.* arXiv:2212.04356, 2022. — background on encoder–decoder ASR systems trained and decoded in the style used by both SKUs (Section 4).

[8] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization.* ICLR, 2015. — the optimizer family underlying the Hugging Face training defaults used here.

[9] P. Micikevicius et al. *Mixed Precision Training.* ICLR, 2018. — background on the FP16 training setup described in Section 4.

[10] A. C. Morris, V. Maier, P. Green. *From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition.* Interspeech, 2004. — background on word error rate and related evaluation measures reported in Section 5.