SALT-31: A Machine Translation Benchmark Dataset for 31 Ugandan Languages

Solomon Nsumba[1,2], Benjamin Akera[1], Evelyn Nafula Ouma[1], Medadi Ssentanda[2] Deo Kawalya[2], Engineer Bainomugisha[1,2], Ernest Tonny Mwebaze[1], John Quinn[1]

[1]Sunbird AI, Kampala, Uganda, [2]Makerere University, Kampala, Uganda

Correspondence: snsumba@sunbird.ai

Abstract

We present the SALT-31 benchmark dataset for evaluation of machine translation models covering 31 Ugandan languages. Unlike sentence-level evaluation sets, SALT-31 is constructed from short, scenario-driven mini-dialogues designed to preserve discourse context, pragmatics, and culturally grounded communication patterns common in everyday Ugandan settings. The dataset contains 100 English sentences organized into 20 typical communication scenarios, each represented as a five-sentence mini-sequence. It can therefore be used to evaluate both sentence-level and paragraph level machine translation, and includes nearly every language spoken in a country with high linguistic diversity. It is available at `hf.co/datasets/Sunbird/salt-31`.

1 Introduction

Evaluation datasets play a central role in advancing machine translation for low-resource languages, particularly in multilingual and linguistically diverse regions such as Africa. Prior work has highlighted that progress in African NLP is often constrained not only by model capacity, but also by the availability of representative, high-quality evaluation resources (Wilhelmina et al., 2020).

In many existing benchmarks, including widely used multilingual evaluation suites, sentences are presented in isolation. While effective for measuring lexical or syntactic adequacy, such sentence-level evaluation often fails to capture discourse phenomena such as coreference, turn-taking, pragmatic intent, and culturally grounded meaning (Bawden et al., 2018; Voita et al., 2018; Goyal et al., 2022).

This limitation is especially pronounced for Ugandan and other African languages, which are predominantly used in spoken, conversational, and community-centered contexts. Recent studies have emphasized the need for context-aware and locally grounded evaluation protocols to better reflect real-world language use in African settings (Joshi et al., 2020).

The SALT-31 Evaluation Dataset was developed to address this gap by introducing a context-aware evaluation resource grounded in realistic Ugandan communication scenarios. Building upon the Sunbird African Language Technology (SALT) initiative (Akera et al., 2022; SunbirdAI, 2023), SALT-31 emphasizes evaluation quality, linguistic diversity, and contextual coherence rather than large-scale parallel corpora. The dataset targets 31 Ugandan languages, spanning multiple language families, and additionally includes a small number of closely related regional languages. Many of these languages remain severely underrepresented in standard machine translation benchmarks. For several of them, this dataset represents the first time that any publicly available parallel text beyond bib-

Table 1: Communication scenarios in SALT-31, covering formal, informal and everyday modes of communication

ID	Scenario Domain
1	Banking transaction
2	Educational instruction
3	Conversational greetings
4	Narrative/fiction
5	Medical consultation
6	News report
7	Encyclopedic information
8	Market/shopping
9	Emergency response
10	Practical technical guide
11	Health/prenatal advice
12	Official announcement
13	Government speech
14	Agricultural extension
15	Opinion poll response
16	Public transport
17	Family conversation
18	Nutrition guidance
19	Food security
20	Civic office dialogue

Table 2: Summary of the SALT-31 Evaluation Dataset

Property	Value
Number of scenarios	20
Sentences per scenario	5
Total English sentences	100
Target languages	31
Domains	Health, market, family, school, transport, daily life
Unit of evaluation	Five-sentence mini-sequence

Table 3: Generated English text for scenario 10 (practical guide for mechanic or builder)

First, make sure the foundation trench is at least two feet deep and well compacted before laying any bricks.
Mix the cement, sand, and aggregate in a 1:2:4 ratio for strong concrete, especially for the pillars.
Use a spirit level to keep the walls straight and check alignment after every few layers of bricks.
Keep the site clean and water the curing concrete regularly for at least seven days to prevent cracks.
Finally, make sure all electrical and plumbing points are marked clearly before plastering the walls.

lical translations has been made available.

While multilingual benchmark datasets such as FLORES provide broad coverage of languages, they are typically sentence-level and therefore do not directly probe discourse consistency across turns. Furthermore, they do not capture types of communication relevant to the local context. SALT-31 complements such benchmarks by focusing on short, coherent five-sentence sequences that enable assessment of context-sensitive phenomena (e.g., role consistency, anaphora, pragmatic intent) in realistic Ugandan communication settings.

This paper describes the design principles, data generation pipeline, translation workflow, and release of the SALT-31 Evaluation Dataset. We also summarize its application in evaluating the Sunflower multilingual language model (Akera et al., 2025), which is currently deployed in a production setting.

2 Data and Methods

An overview of the SALT-31 Evaluation Dataset, including its size, scenario structure, language coverage, and release format, is summarized in Table 2.

2.1 Scenario and Sequence Generation

The English source data was generated using a structured prompt that instructed large language models (LLMs) to produce five-sentence mini-sequences for predefined scenarios. A total of 20 scenarios were defined, each yielding one five-sentence sequence, resulting in 100 English sentences. LLMs were used only to generate controlled English seed text; final sequences were manually reviewed and selected, and all translations were produced by native speakers.

An example scenario is patient-doctor communication in a medical setting, where dialogue captures symptoms, follow-up questions, and clinical reasoning. Other scenarios include everyday greetings, small talk, market negotiations, school interactions, and community discussions. An example for scenario 10, practical instructions for a builder, are shown in Table 3.

To encourage stylistic diversity and reduce model-specific artifacts, multiple LLMs were used to generate candidate sequences, including GPT-4.5, GPT-4o, DeepSeek R1, LLaMA 3.3 70B, Mistral Large, Gemini 2 Flash, and Claude Sonnet. For each scenario, outputs were reviewed and a single sequence was selected based on contextual appropriateness, cultural grounding, and linguistic naturalness for Uganda.

2.2 Data Structuring

Once finalized, the 20 mini-sequences were compiled into a structured spreadsheet, with each row corresponding to a single English sentence and metadata linking it to its scenario and sequence position. This intermediate format facilitated systematic translation and quality control.

2.3 Community-Driven Translation

Translation was carried out in collaboration with experts in mother-tongue education, early literacy development and language policy. Native speakers translated the English sentences into 31 target languages spanning three major families: Bantu (17 languages), Nilotic (11 languages), and Central Sudanic (3 languages). Representative languages from each family are shown in Table 4; the complete language list is provided in Appendix B.

Translations emphasized meaning preservation, natural phrasing, and cultural appropriateness rather than literal word-for-word mapping.

To ensure gold-standard quality, all translations underwent an independent verification process conducted by the Department of Linguistics, English Language Studies and Communication Skills. Teams of trained linguists and native speakers reviewed each translation for semantic fidelity, grammatical correctness, naturalness of expression, and

cultural appropriateness. Discrepancies were resolved through consensus review, with reference to the original English source and the intended communicative context of each scenario. This multi-stage verification process provided an additional layer of quality assurance beyond initial translation, increasing confidence in the reliability of SALT-31 as a gold-standard evaluation dataset.

3 Results and Discussion

SALT-31 evaluation covers 20 distinct communication domains (Table 1), ranging from healthcare consultations to agricultural extension contexts identified through community consultation as critical for Ugandan language use. This breadth enables assessment of MT systems across diverse registers, from formal government speeches to informal market negotiations.

Table 4 illustrates the linguistic diversity captured across these scenarios by showing a single English sentence from Scenario 10 (practical construction guide) translated into representative languages from three major families: Bantu (e.g., Luganda, Lusoga), Nilotic (e.g., Acholi, Lango), and Central Sudanic (e.g., Lugbara, Ma'di). These translations reveal systematic structural differences in word order, morphological complexity, and noun class agreement; features that pose distinct challenges for MT systems.

3.1 Baseline Model Performance

We evaluated proprietary models (GPT-4o, Gemini 2.5 Pro, Grok-3) and open-weight alternatives (Sunflower-14B/32B, DeepSeek-Chat, NLLB-1.3B) using chrF as the primary metric, which better captures morphological similarity in agglutinative languages than BLEU (Popović, 2015).

Table 5 presents average performance across all 31 languages. Sunflower models achieve the highest scores in both translation directions, with Sunflower-32B excelling at local-to-English translation (chrF=0.435) and Sunflower-14B performing best for English-to-local (chrF=0.366). Notably, these regionally specialized models outperform substantially larger general-purpose systems. GPT-4o achieves only 0.354 chrF for xx→en and 0.235 for en→xx, despite being substantially larger in scale.

This performance gap reflects fundamental challenges in massively multilingual training. Large-scale models trained on hundreds of languages must allocate limited parameter capacity across diverse distributions, often resulting in reduced performance on low-resource languages compared to focused regional approaches (Conneau et al., 2020; Arivazhagan et al., 2019). Our results suggest that regional specialization concentrating model capacity on linguistically related languages within a coherent geographic area can yield superior performance for underrepresented languages.

3.2 Performance by Language Family

Performance varies systematically across language families (Table 6). While most models achieve reasonable scores on Bantu languages (mean chrF 0.30–0.41), performance degrades for Nilotic and Central Sudanic languages. GPT-4o achieves 0.323 chrF on Bantu languages but drops to 0.139 for Nilotic and 0.094 for Central Sudanic languages. This 3-4× performance gap exposes significant inequities in current MT systems.

In contrast, Sunflower models maintain more consistent performance across families, with only modest degradation from Bantu (0.413) to Nilotic (0.312) to Central Sudanic (0.299). This consistency validates the effectiveness of training on regionally coherent language groups where structural similarities enable cross-lingual transfer even for extremely low-resource languages.

3.3 Context-Aware Evaluation Findings

The mini-sequence structure enabled qualitative analysis of discourse-level phenomena. Manual inspection of model outputs by native speakers revealed three recurring failure patterns across the mini-sequences:

Coreference errors: Models frequently failed to maintain consistent pronominal reference across sentences. In Scenario 5 (medical consultation), references to "the patient" were inconsistently translated across the five-sentence sequence, sometimes incorrectly switching gender or number.

Register inconsistency: In formal scenarios (e.g., Scenario 13, government speeches), models produced mixed registers, inappropriately combining colloquial and formal constructions within the same sequence.

Cultural misalignment: Technical terms and culturally specific concepts were often mistranslated. In Scenario 14 (agricultural advice), references to traditional farming practices were sometimes rendered with urban or formal terminology inappropriate for rural extension contexts.

Table 4: Example sentence from Scenario 10 translated across three language families, illustrating typological diversity in word order, morphology, and agreement systems

Code	Language	Translation (Scenario 10: Construction Guide)
eng	English	First, make sure the foundation trench is at least two feet deep and well compacted before laying any bricks.
Bantu Family		
lug	Luganda	Ekisooka, kakasa nti omusingi gukka waakiri fuuti bbiri era nga guggumizibbwa bulungi nga tonnassaayo mataffaali.
xog	Lusoga	okusooka, kakasa nti olukonko lw'omusingi lughanvu okuswika fuuti eibiri era nga lwidhiziibwa bukalamu nga okali kutandiika kuzimba ku matafali.
nyn	Runyankole	Ekyokubanza, reeba ngu omusingye gutimbirwe kuhisya fuuti ibiri ahansi kandi gwijwize kurungi, otakatandikire kwombeka amatafaari.
ttj	Rutooro	Ekyokubanza, rora ngu omusingi guhikire fuuti ibiri hansi, gusokiire kurungi otakataireho amatafaali.
Nilotic Family		
ach	Acholi	Me acel, nen ni tut pa bur me te ot olo to romo tyen aryo dok kitoro maber ma pud pe iketo matapwali mo iye.
laj	Lango	Me acel, nen ni bur me pandeccon tye ame tuttere romo tyelo aryo dang ocwiny aber ame pwod pe iketo birikkoro keken.
teo	Ateso	Kigeari, kowany ebe idulu aipany na ibokarit ijo bala ipuutin iare ido kibamakina kokwap kojokan eroko ijo inapakina amatapaalin adio kere.
Central Sudanic Family		
lgg	Lugbara	Okoria, I'ba kini 'bile 'bani gale fawundasoni e'dozu abi sizuri ma aliniri ma ovu fiti iri azini 'bama omi eri kililiru denga 'ba 'bani mutufali eyi vaa kuru.
mhi	Ma'di	Atijoa ru rii, kole'a nyi ba fa'undesoni a'a bu ni kolu fiti eri guru vua ure okpo oca matafali bi re ku.

Table 5: Average chrF scores across 31 languages. Specialized regional models outperform general-purpose systems despite smaller size

Model	xx→en	en→xx
Regionally Specialized		
Sunflower-32B	**0.435**	0.357
Sunflower-14B	0.419	**0.366**
General Purpose		
Gemini 2.5 Pro	0.408	0.301
GPT-4o	0.354	0.235
Grok-3	0.347	0.247
DeepSeek-Chat	0.308	0.237

Table 6: Average chrF (en→xx) by language family. General-purpose models exhibit 2-3× performance degradation for non-Bantu languages, while Sunflower maintains more consistent quality across families

Model	Bantu	Nilotic	C. Sud.
Sunflower-14B	**0.413**	**0.312**	**0.299**
Sunflower-32B	0.406	0.299	0.295
Gemini 2.5 Pro	0.369	0.230	0.173
GPT-4o	0.323	0.139	0.094
Grok-3	0.299	0.190	0.161
DeepSeek-Chat	0.295	0.172	0.148

These failures underscore SALT-31's value as a diagnostic tool. While aggregate metrics provide useful quality estimates, discourse-level evaluation reveals systematic weaknesses in handling context, a critical requirement for deploying MT systems in real-world Ugandan communication settings.

4 Conclusions

We have introduced the SALT-31 Evaluation Dataset, a context-aware machine translation benchmark covering 31 Ugandan languages across Bantu, Nilotic, and Central Sudanic language families. Unlike conventional sentence-level evaluation sets, SALT-31 structures translation tasks as five-sentence sequences drawn from 20 realistic communication scenarios that reflect authentic language use in Uganda. The option to group sentences together as mini-documents allows evaluation of models with respect to coreference resolution, register consistency, and cultural grounding phenomena that sentence-level benchmarks cannot capture. Because the dataset is focused on the contexts in which these languages are typically used, evaluation metrics help as diagnostic indicators of models' ability to handle real-world practical applications – even though the small size of the dataset means that more thorough assessment, and especially human evaluation by native speakers, would typically be needed to assess readiness for deployment in a production setting.

Evaluation of state-of-the-art MT systems on

SALT-31 reveals substantial performance gaps. While regionally specialized models like Sunflower achieve robust performance across all groups, general-purpose systems underperform despite their scale. We observed a drastic performance degradation for all models when carrying out machine translation on the lower-resourced of these languages, for which very little training data is available.

Future work will expand SALT-31 to cover code-switching patterns, dialectal variation, and speech modalities, enabling evaluation of speech-to-text systems critical for Uganda where many languages remain primarily oral. SALT-31 is released openly on Hugging Face to support reproducible evaluation and comparative benchmarking.

Limitations

SALT-31 is small (100 English sentences) and is designed for controlled, context-aware evaluation and diagnostic usage, rather than being a dataset which would support fine-grained evaluation or model training. Although scenarios were curated to reflect realistic Ugandan communication, they cannot cover all sociolinguistic registers, dialectal variation, or code-switching patterns present across the 31 languages. Future expansions will increase scenario diversity, incorporate longer contexts, and include additional evaluation protocols involving human judgments.

Acknowledgments

We acknowledge the linguists, educators, and translators from Makerere University whose expertise made this dataset possible. We also thank the Sunbird AI team for supporting dataset preparation, hosting, and evaluation.

References

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Muhanguzi, Imran Sekalala, and 1 others. 2025. Sunflower: A new approach to expanding coverage of african languages in large language models. *arXiv preprint arXiv:2510.07203*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

SunbirdAI. 2023. Salt: Sunbird african language technology. https://github.com/SunbirdAI/salt.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Nekoto Wilhelmina, Marivate Vukosi, Matsila Tshinondiwa, Fasubaa Timi, Fagbohungbe Taiwo, Akinola Solomon Oluwole, Muhammad Shamsuddeen, Kabenamualu Salomon Kabongo, Osei Salomey, Sackey Freshia, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

5 Appendices

A Prompt Used for Scenario Generation

We used the following prompt to generate English mini-sequences for all scenarios:

> You are helping design a multilingual evaluation set for machine translation for use in Uganda. For each scenario, create a short dialogue or paragraph consisting of exactly 5 sentences, capturing natural, contextually realistic communication that would make sense in a Ugandan context. Use plain, accessible language, and keep the tone appropriate to the scenario (e.g., professional for business, conversational for daily life, etc.).

B Complete Language List

Table 7 presents the complete list of 31 target languages, organized by language family.

#	Code	Language	#	Code	Language
Bantu (17 languages)					
1	cgg	Rukiga	10	nyn	Runyankole
2	gwr	Lugwere	11	nyo	Runyoro
3	kin	Kinyarwanda	12	rub	Lugungu
4	koo	Rukonjo	13	ruc	Ruruuli
5	lsm	Samia	14	rwm	Kwamba
6	lug	Luganda	15	swa	Swahili
7	myx	Lumasaba	16	tlj	Lubwisi
8	nuj	Lunyole	17	ttj	Rutooro
9	xog	Lusoga			
Nilotic (9 languages)					
18	ach	Acholi	23	kpz	Kupsabiny
19	adh	Dhopadhola	24	laj	Lango
20	alz	Alur	25	pok	Pokot
21	kdi	Kumam	26	teo	Ateso
22	kdj	Karamojong			
Central Sudanic (5 languages)					
27	bfa	Bari	30	luc	Aringa
28	keo	Kakwa	31	mhi	Ma'di
29	lgg	Lugbara			

Table 7: Complete list of 31 target languages in SALT-31, organized by language family.