KinyCOMET: Automatic Evaluation of Machine Translation Systems for Kinyarwanda–English

Prince Chris Mazimpaka[1], **Jan Nehring**[2]
Samuel Rutunda[3], **Cristina España-Bonet**[4]

[1]Mbaza Open-Source Community, Kigali, Rwanda
[2]Gesellschaft für Internationale Zusammenarbeit (GIZ), Regional AI Hub Rwanda, Kigali, Rwanda
[3]Digital Umuganda, KG 7 Ave, Kigali, Rwanda
[4]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

[1]chrismazimpaka7@gmail.com, [2]jan.nehring@giz.de,
[3]samuel@digitalumuganda.com, [4]cristinae@dfki.de

Abstract

This paper presents KinyCOMET, a new automatic evaluation metric for Kinyarwanda–English machine translation (MT). Current MT evaluation in Rwanda relies mainly on BLEU and chrF, which have been shown to correlate poorly with human judgments. To address this gap, we created a Direct Assessment (DA) dataset for Kinyarwanda-English translations and used it to fine-tune COMET models for this language pair. We evaluate two variants: KinyCOMET XLM-RoBERTa, trained from a multilingual encoder without Kinyarwanda data, and KinyCOMET Unbabel, a fine-tuned version of the Unbabel COMET model. Both models achieve strong correlations with human evaluations, with KinyCOMET Unbabel outperforming all baselines, including AfriCOMET, chrF, and BLEU. Our results show that fine-tuning pre-trained multilingual models can yield high-quality evaluators even for low-resource languages that the base model was not trained on. We release both the models and the annotated dataset publicly to foster further research on African language evaluation.

Keywords: machine translation evaluation, Kinyarwanda, COMET, African languages, low-resource NLP

1. Introduction

The African continent is home to a remarkable linguistic diversity, with estimates ranging from 1,250 to 2,100 distinct native languages (Heine and Nurse, 2000), and by some counts exceeding 3,000 (Epstein and Kole, 1998). The number of languages is a barrier to the country's economic development. Another problem is that people from rural areas often struggle to understand English, which excludes them from lucrative jobs in the tourism sector and generally limits their participation in the globalized world. Machine Translation (MT) can help to overcome these problems. Also, MT is an important building block for other technologies, such as generating training data in African languages, adapting English tooling to African languages, and more.

Rwanda is home to a thriving ecosystem around MT. A list of startups and IT companies, such as Digital Umuganda,[1] KINLP,[2] Awesomity,[3] and Artemis AI,[4] utilizes MT models. Furthermore, there is a range of open-source MT models for Kinyarwanda/English, e.g., JoeyNMT (NLLB Team, 2024) or NLLB (NLLB Team, 2024), which support many languages, including Kinyarwanda. Additionally, there is a list of open-source datasets for MT, such as those from Digital Umuganda, Mbaza NLP[5], or NLLB.

From the discussion with the community, we found out that a pain point of the Rwandan MT community is evaluation. MT practitioners in Rwanda currently use BLEU (Papineni et al., 2002) and chrF (Popović, 2015), which does not align well with human evaluations (Freitag et al., 2022). Therefore, the Rwandan NLP community evaluates their models using human annotators, which is expensive and time-consuming. The scientific community proposed more modern evaluation metrics, such as METEOR (Banerjee and Lavie, 2005), LLM-based metrics (Huang et al., 2024), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020).

Therefore, we decided to create KinyCOMET, an automatic metric based on COMET to evaluate Kinyarwanda–English machine translation models. The scientific contributions and outcomes of this paper are the following:

- Open KinyCOMET models, published on Huggingface

[1]https://digitalumuganda.com
[2]https://kinlp.org
[3]https://awesomity.rw
[4]https://artemis.rw

[5]https: //github.com/mbaza-nlp

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4881–4888
11–16 May 2026.

- Open dataset to train related models, published on Huggingface
- A comparison of KinyCOMET to the state-of-the-art metrics
- A comparison of MT models for Kinyarwanda–English translation

The final dataset[6] and the models[7] are publicly available.

2. Background

The scientific literature mentions various automatic methods to evaluate the performance of machine translation models. The first metrics were based on lexical overlap. BLEU (Papineni et al., 2002) measures a token-based overlap. Similarly, chrF (Popović, 2015) measures character-based overlap and is better suited for morphologically rich languages such as Kinyarwanda. However, both of these metrics fail if a correct translation deviates strongly from the reference.

These early approaches are known as lexical metrics because they rely on surface overlap between words or characters. In contrast, semantic metrics assess how well a translation preserves meaning rather than form, using contextual embeddings from pre-trained language models. Examples include BERTScore, COMET, and recent LLM-based methods.

Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020) is a neural machine translation evaluation metric that uses a multilingual encoder (e.g. XLM-RoBERTa (Conneau et al., 2020)) to produce sentence embeddings for source, reference, and hypothesis. Then, it predicts a quality score via a regression head trained on human-annotated quality data. One possible annotation scheme is Direct Assessment (Graham et al., 2013) (DA), in which human annotators judge how accurately and fluently a translation conveys the meaning of the source in a scale between 1 and 100. COMET includes several variants: Reference-based models (used in this work) take the source, hypothesis, and reference translation as input to predict a quality score that aligns with human judgments. Reference-free models rely only on the source and hypothesis, which is useful when references are unavailable. Ranking models are trained to order translations by quality rather than predict exact scores, while explanation models highlight the words or phrases that most influence the evaluation.

AfriCOMET (Wang et al., 2024) extends COMET to 13 African languages, but Kinyarwanda is not among these languages. An extension of COMET exists for additional languages and includes Kinyarwanda (Li et al., 2025). However, the evaluations we present in this work show that the performance of the model is not high enough for practical use.

3. Method

3.1. Data Generation

Partition	Direction	Number of Samples
train	both	3,477
train	eng2kin	1,803
train	kin2eng	1,674
valid	both	404
valid	eng2kin	199
valid	kin2eng	205
test	both	422
test	eng2kin	223
test	kin2eng	199
all	eng2kin	2,225
all	kin2eng	2,078
all	all	4,303

Table 1: Number of samples in the dataset.

We used three publicly available Kinyarwanda–English parallel corpora as source datasets: Mbaza Education, Mbaza Tourism, and the Digital Umuganda Dataset. The source datasets do not have an annotated direction, they are just parallel sentences.

29 linguistics students annotated the data as part of a university project. Following the procedure of AfriCOMET (Wang et al., 2024), they annotated DA on a scale between 1 and 100 using the open-source annotation tool.[8]

Also, a small number of the translations were faulty. E.g., either the source or target was in the wrong language, or the reference did not properly translate the source. To address this issue, we sampled 225 samples from each of the three source datasets, so 675 samples in total. Then, the students annotated DA scores between 1 and 100 for each sample. Due to a bug, each sample got annotated by many annotators (a minimum of 6, an average of 7.8, and a maximum of 29). Then, we used the best 150 samples from each of the datasets, which resulted in a dataset of 450

[6] https://huggingface.co/datasets/chrismazii/kinycomet_dataset

[7] https://huggingface.co/chrismazii/kinycomet_unbabel

[8] https://github.com/marek357/annotation-tool-frontend

samples. A manual inspection confirmed that this solved most aforementioned quality issues.

From these datasets, we produced machine translations using six systems: Claude 3.7-Sonnet (2025-02-19), OpenAI GPT-4o (2024-08-06), OpenAI GPT-4.1 (2025-04-14), Google Gemini Flash 2.0 (flash-001), and Facebook NLLB models with 1.3B and 600M parameters (NLLB Team, 2024). We used each sample twice, once for each direction. This resulted in 5400 samples, 50% for each direction kin2eng and eng2kin.

Next, the annotators annotated DA scores. Due to university procedures, 16 annotators were available for the second annotation. These annotators annotated only a subset of 4,733 samples from our 5,400 high-quality samples due to confusion about which annotators were actually available. Each sample got annotated by three annotators. To ensure high data quality, we removed samples with a standard deviation greater than 20, which resulted in the removal of 410 samples, 9.5% of the total. This resulted in 4303 samples, 2225 for the direction eng2kin and 2078 for the direction kin2eng. We split the dataset randomly into 80% train, 10% valid and 10% test partitions. Table 1 shows more detailed information on the dataset.

3.2. Dataset Examples

Table 2 shows three samples from the dataset. We randomly selected the samples, but we chose samples with shorter text to match the paper format.

3.3. Model Training and Evaluation

We fine-tuned our COMET model using two different base models. In the first case, we use XLM-RoBERTa (Conneau et al., 2020), which is a transformer model similar to RoBERTa (Liu et al., 2019) but trained on more than 100 languages. Kinyarwanda is not among them. To our knowledge, there is no open-source multilingual transformer model trained on Kinyarwanda data. We call this model-variant **KinyCOMET XLM-RoBERTa**.

In the second case, we fine-tune the Unbabel COMET model (Rei et al., 2022). Unbabel COMET is also based on XLM-RoBERTa. It was fine-tuned on the data of the WMT shared task (Freitag et al., 2022) from 2017–2020.[9] We call this model-variant **KinyCOMET Unbabel**. Both models were trained using the publicly available training scripts.[10]

Both KinyCOMET variants were trained on a Tesla T4 GPU using 16-bit mixed precision, with 3,477 training samples and 404 validation samples. They share several hyperparameters: AdamW optimizer, dropout of 0.1, layerwise decay of 0.95, and gradient clipping at 1.0. In both cases, we optimized MSE loss and selected the best checkpoint based on the highest Pearson correlation on the validation set.

The two models differ in the following settings. We used the default settings of the public Unbabel training scripts, adjusting the batch size only to match our GPU server. For the **XLM-RoBERTa** variant, we used a batch size of 8, an encoder learning rate of 1×10^{-6}, and a main learning rate of 1.5×10^{-5}. Training ran for up to 10 epochs with validation performed twice per epoch, the encoder frozen for the first 0.3 epochs, and early stopping with patience of 3.

For the **Unbabel** variant, we used a batch size of 4 with gradient accumulation over 2 steps (effective batch size of 8) and a unified learning rate of 1×10^{-5} for both encoder and task-specific layers. Training ran for up to 15 epochs with validation performed twice per epoch, and early stopping with patience of 4.

We compare our two models to several existing metrics: Unbabel (Rei et al., 2020), which is trained on 102 languages, but not on Kinyarwanda. AfriCOMET STL 1.1 (Li et al., 2025) is the version trained on the AfriCOMET data that includes Kinyarwanda. Also, we compare our models to BLEU and chrF.

Following the methodology of the authors of COMET (Rei et al., 2020), we use Kendall-Tau τ (Kendall, 1938) and Spearman ρ (Spearman, 1904) to measure the correlation between model outputs and human DAs scores. Graham et al. (2015) argues that Pearson is preferable over Kendall-Tau because it allows direct comparison of continuous human ratings, makes evaluation unit-free and more intuitive, and supports statistical significance testing.

We compare all the models above on the test partition of our dataset. Unfortunately, we cannot compare them with the SSA-COMET test set (Li et al., 2025) because, at the time of writing, the dataset is not public.

4. The Models

4.1. Evaluation of KinyCOMET

Table 3 shows the correlation of the automatic evaluation of the different metrics with human judgements, for individual directions, and averaged over both directions. Both our models, KinyCOMET Unbabel and KinyCOMET XML-RoBERTa, reach state-of-the-art results. As expected, chrF performs better than BLEU. Generally (with one exception), the COMET-based models outperform both BLEU and chrF. The low scores for both AfriCOMET variants are surprising. Unbabel, which is not trained on that

[9] https://wmt-metrics-task.github.io
[10] https://github.com/Unbabel/COMET

Source sentence	Machine Translation	Reference	Score	Direction
Hano hari kugabanuka kubuntu hafi ya bose!	Here there is a free discount for almost! everyone!	There are generous discounts for nearly everyone!	0.92	kin2eng
Umuyobozi wanjye , Godwin, yari igitangaza	My manager, Godwin, was a genius.	My guide, Godwin, was fantastic.	0.92	kin2eng
There is then a right turn about 500m further on, this is the guest house.	Hariho noneho gutaha iburyo nko muri metero 500 imbere, ni inzu y'abashyitsi.	Noneho har'ahakata iburyo muri metero 500 ukomeje imbere, aho niho hari icumbi ry'abashyitsi.	0.90	eng2kin

Table 2: Examples from the dataset

specific language pair, outperforms AfriCOMET. Notice that direction eng2kin outperforms the direction kin2eng consistently across all metrics.

4.2. Evaluation of Machine Translation Kinyarwanda–English

Table 4 shows the DA scores of the human annotators for the different MT systems used to generate our dataset. Generally, LLM-based approaches outperform NLLB models. While GPT-4.0 and GPT-4.1 have the best performance on kin2eng, Gemini-2.0 Flash achieves state-of-the-art results for eng2kin. Averaged across both directions, GPT-4o, GPT-4.1 and Gemini-2.0-Flash achieve the best results. However, the scores are very similar to each other, and all scores lie within each other's standard deviations. Especially, the LLM models have very similar results.

4.3. Distribution of Direct Assessment Scores

To understand the nature of the dataset, we compare the distribution of our DAs to those of a standard dataset used in MT evaluation, the WMT datasets (data from 2017–2022 from 41 different language pairs) (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020; Freitag et al., 2021, 2022). Also, we added the distribution of DAs that the different models predicted on the KinyCOMET datasets test partition. The WMT data uses another scale for DA as ours, so we scaled all scores individually for each dataset or model to a scale from 0-1 using Min-Max Normalization $\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$.

Figure 1 shows a boxplot of the normalized DA scores. In both datasets, the scores are centered mainly on a small range of the scale. Only KinyCOMET XML-RoBERTa and KinyCOMET Unbabel capture the distribution well.

When the majority of training data falls in such a small intervall, the models might learn to predict DA scores centered around this small area as well.

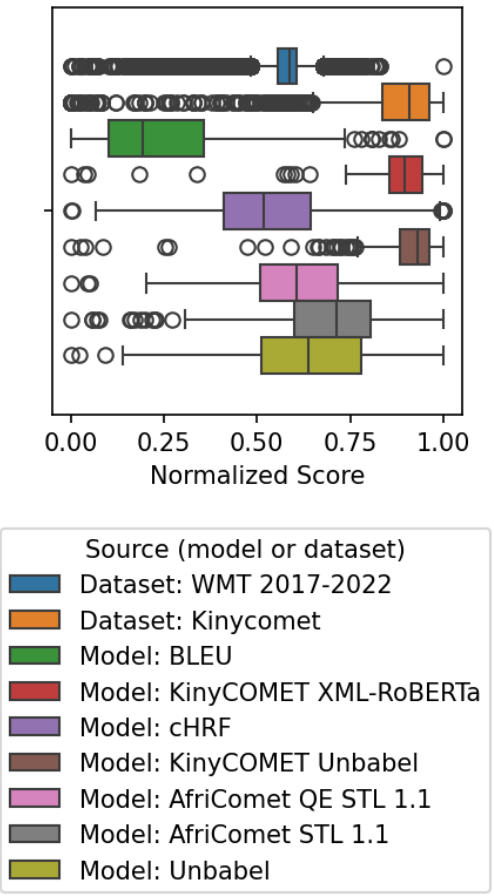

Figure 1: Distribution of DA scores in the WMT 2017-2022 data, our dataset and the model predictions on our dataset.

If the test dataset falls within the same interval, we get good results. This might also explain the low scores of AfriCOMET on our test dataset, because the AfriCOMET scores might follow a different distribution.

If we re-ran the data collection, we would include weaker MT models to capture lower DA scores also to teach the model to predict lower scores also.

Model	Direction	Pearson	Kendall's-Tau
AfriComet QE STL 1.1	both	0.34 (moderate)	0.11 (moderate)
AfriComet STL 1.1	both	0.52 (strong)	0.24 (moderate)
BLEU	both	0.30 (weak)	0.23 (moderate)
chrF	both	0.38 (moderate)	0.21 (moderate)
KinyCOMET Unbabel (ours)	both	**0.75 (strong)**	**0.42 (strong)**
KinyCOMET XML-RoBERTa (ours)	both	0.73 (strong)	0.35 (strong)
Unbabel	both	0.54 (strong)	0.39 (strong)
AfriComet QE STL 1.1	eng2kin	0.59 (strong)	0.30 (strong)
AfriComet STL 1.1	eng2kin	0.67 (strong)	0.36 (strong)
BLEU	eng2kin	0.30 (weak)	0.25 (moderate)
chrF	eng2kin	0.48 (moderate)	0.28 (strong)
KinyCOMET Unbabel (ours)	eng2kin	0.75 (strong)	**0.40 (strong)**
KinyCOMET XML-RoBERTa (ours)	eng2kin	**0.76 (strong)**	0.37 (strong)
Unbabel	eng2kin	0.52 (strong)	0.36 (strong)
AfriComet QE STL 1.1	kin2eng	0.23 (weak)	0.16 (moderate)
AfriComet STL 1.1	kin2eng	0.41 (moderate)	0.22 (moderate)
BLEU	kin2eng	0.16 (weak)	0.07 (moderate)
chrF	kin2eng	0.23 (weak)	0.11 (moderate)
KinyCOMET Unbabel (ours)	kin2eng	**0.63 (strong)**	**0.33 (strong)**
KinyCOMET XML-RoBERTa (ours)	kin2eng	0.37 (moderate)	0.21 (moderate)
Unbabel	kin2eng	0.43 (moderate)	0.26 (strong)

Table 3: Correlation of metrics with human judgements, for the individual directions and averaged over both directions.

System	kin2eng	eng2kin	both
Claude-3.7	92 (8)	86 (11)	89 (10)
GPT-4o	**93 (8)**	88 (11)	**91 (10)**
NLLB-1.3B	89 (12)	84 (16)	87 (15)
NLLB-600m	89 (12)	75 (28)	83 (22)
GPT-4.1	**93 (7)**	88 (10)	**91 (9)**
Gemini-2.0-Flash	91 (11)	**90 (9)**	**91 (10)**

Table 4: Comparison of DA for the MT systems used for both translation directions and their average (both). The table shows average DA and the standard deviation in brackets.

5. Discussion

Our model outperforms the current state of the art. Finetuning Unbabel for our specific use case proved to be more efficient than training a model from scratch. We hypothesize that KinyCOMET Unbabel benefits from transfer learning from the other languages it was originally trained on. Overall, Unbabel achieves high performance, on average better than AfriCOMET, which is surprising given that neither the base model nor the fine-tuning used Kinyarwanda data.

In general, Kinyarwanda-to-English (kin2eng) scores are higher than English-to-Kinyarwanda (eng2kin) scores. This is unexpected, as one might assume that the English language model is stronger than the Kinyarwanda model, and therefore would assess English samples more accurately. We leave it to future research to explain this phenomenon.

Our dataset also shows an unusually concentrated distribution of Direct Assessment (DA) scores, suggesting that a more diverse dataset would provide more informative evaluation results. Finally, the data clearly demonstrates that chrF and BLEU scores do not align well with human judgments.

6. Ethical Considerations

6.1. Annotator compensation

Our data annotators generated the dataset as part of a university seminar and were therefore compensated through fulfillment of their academic requirements and hands-on learning about data generation for machine translation evaluation. The annotators did not receive additional financial compensation.

6.2. Data Extractivism and Benefit Sharing

Recent discussions around data extractivism and fair benefit sharing (Okorie and Omino, 2025; Rajab et al., 2025) have criticized common licensing practices in NLP research. Typically, datasets and AI models are published under open licenses such as the Creative Commons Attribution License[11], which allow free use with the aim of promoting innovation and enabling commercial applications. However, this approach can perpetuate data extractivism: due to power asymmetries between actors in the Global North and Global South, high-resource actors from the Global North are better positioned to turn openly released datasets (from the Global South) into profitable products/technologies which can then / are then being re-introduced to the same originating communities as commercial digital services. Okorie and Omino (2025) and Rajab et al. (2025) emphasize that language datasets represent the collective knowledge of the communities that speak those languages and argue that these communities should participate in, and equitably benefit from, any downstream use or commercialization of their linguistic resources.

6.3. Licensing Framework

Because our dataset relies on linguistic resources produced by Rwandan speakers and students, equitable benefit-sharing mechanisms are particularly relevant. We intend to adopt the NOODL licensing framework (Okorie and Omino, 2025) for our dataset publication. The NOODL license distinguishes between two types of users: while users from developing countries can access the dataset for free, users from developed countries must provide a form of benefit sharing (e.g. monetary licensing fee, access to resources and infrastructure, collaboration). In order to manage this benefit sharing process, we plan to nominate a licensing custodian who can determine benefit sharing / pricing for the dataset and serve as a point of contact for Global North users seeking to license the data. Since the Kinyarwanda-speaking community consists primarily of people from Rwanda, we consider the Rwandan government to be an appropriate data custodian. While the NOODL license applies directly to the dataset, the same benefit-sharing procedures extend to our models as derivative works of the data.

6.4. Licensing Implementation Timeline

These licensing frameworks are very new, and there are no established precedents for releasing datasets under the NOODL license. In particular, engaging with the community, contacting the Rwandan government and establishing their role as data custodian will require a lengthy process. To ensure quick adoption of our dataset and models by the research community and Rwandan machine translation companies, we will initially publish them under the Creative Commons BY license. As our discussions with the Rwandan government progress, and pending their agreement to serve as data custodian, we will transition to the NOODL license. The data custodian may also choose to allow Global North users to access the data and models freely for non-commercial research purposes. Ultimately, NOODL emphasizes that benefit-sharing arrangements must be defined by the originating community (typically through an appointed data custodian). Therefore, we as the dataset creators refrain from unilaterally defining the license terms on behalf of the community.

[11] https://creativecommons.org/licenses/by/4.0

7. Conclusion

In this work, we introduced KinyCOMET, an automatic evaluation metric tailored to Kinyarwanda-English machine translation. By fine-tuning COMET models with newly collected Direct Assessment (DA) data, we demonstrated that our approach achieves strong correlations with human judgments and outperforms existing metrics, including AfriCOMET, BLEU, and chrF.

Beyond the model itself, we contribute a high-quality, publicly available dataset of human translation assessments for Kinyarwanda–English. This resource fills an important gap for African language research and provides a foundation for further work on evaluation and model development.

8. Bibliographical References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

E.L. Epstein and R. Kole. 1998. *The Language of African Literature*. Africa World Press.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Bernd Heine and Derek Nurse. 2000. *African Languages: An Introduction*. Cambridge University Press, Cambridge.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs Outperform Learned Metrics in Evaluating MT for Under-Resourced African Languages?

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:pages841–846.

Patrick Nyongabo. 2017. Kinyarwanda–English Translator. A Statistical Machine Learning Approach. Bachelor's thesis, Hendrix University, 1600 Washington Ave, Conway, AR 72032, United States.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. ArXiv:2404.02392 [cs].

Chijioke I Okorie and Melissa Omino. 2025. Addressing inequitable openness in licences for sharing african data and datasets through the nwulite obodo open data licence. *Law, Technology and Humans*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessica Ojo, Atnafu Lambebo Tonja, Wilhelmina NdapewaOnyothi Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2025. The esethu framework: Reimagining sustainable dataset governance and curation for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30763–30776, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

C. Spearman. 1904. The Proof and Measurement of Association Between two Things. *American Journal of Psychology*, 15:88–103.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

9. Language Resource References