Correcting FLORES Evaluation Dataset for Four African Languages

Idris Abdulmumin[1*+], **Sthembiso Mkhwanazi**[2], **Mahlatse S. Mbooi**[2], **Shamsuddeen Hassan Muhammad**[3*+], **Ibrahim Said Ahmad**[4*+], **Neo Putini**[5], **Miehleketo Mathebula**[1], **Matimba Shingange**[1], **Tajuddeen Gwadabe**[*+], **Vukosi Marivate**[1,6]

[1]Data Science for Social Impact, University of Pretoria, [2]Council for Scientific and Industrial Research, South Africa, [3]Imperial College, London, [4]Northeastern University, [5]University of KwaZulu-Natal, [6]Lelapa AI, [*]HausaNLP, [+]MasakhaneNLP

correspondence: idris.abdulmumin@up.ac.za

Abstract

This paper describes the corrections made to the FLORES evaluation (dev and devtest) dataset for four African languages, namely Hausa, Northern Sotho (Sepedi), Xitsonga, and isiZulu. The original dataset, though groundbreaking in its coverage of low-resource languages, exhibited various inconsistencies and inaccuracies in the reviewed languages that could potentially hinder the integrity of the evaluation of downstream tasks in natural language processing (NLP), especially machine translation. Through a meticulous review process by native speakers, several corrections were identified and implemented, improving the overall quality and reliability of the dataset. For each language, we provide a concise summary of the errors encountered and corrected and also present some statistical analysis that measures the difference between the existing and corrected datasets. We believe that our corrections improve the linguistic accuracy and reliability of the data and, thereby, contribute to a more effective evaluation of NLP tasks involving the four African languages. Finally, we recommend that future translation efforts, particularly in low-resource languages, prioritize the active involvement of native speakers at every stage of the process to ensure linguistic accuracy and cultural relevance.

1 Introduction

Low-resource languages, especially from Africa, are greatly under-represented in the Natural Language Processing (NLP) landscape, and this is primarily due to the absence of sufficient resources for both training and evaluation (Adelani et al., 2022; Kreutzer et al., 2022). Various efforts have been made to create such resources and these include initiatives from organizations such as Lacuna[1] that fund new and qualitative open datasets, and communities such as Masakhane, HausaNLP, the University of Pretoria's Data Science for Social Impact (DSFSI) Research Group, and other individual initiatives (Abdulmumin et al., 2022; Parida et al., 2023). For machine translation evaluation, the FLORES dataset (Goyal et al., 2021; NLLB Team et al., 2022) is widely accepted as a benchmark for evaluation, especially because it was the first of its kind for many languages and enables many-to-many evaluation, making it easier to evaluate say a Hausa to Sepedi translation system without pivoting through a high resource language, e.g., English. Recently, the MAFAND dataset (Adelani et al., 2022) was created, but it only allows bilingual evaluation and is limited to the news domain.

While all these resources are being developed, it is imperative to review them for validation to ensure that they meet the expected standard of accuracy and representation. A revealing work by Kreutzer et al. (2022), albeit on mostly web-crawled datasets, found that many of the datasets that are being relied upon for low-resource languages are littered with significant errors such as misalignments, incorrect translations, and other issues. The significance of evaluation datasets make them even more deserving of such reviews especially by literate native speakers that know how these languages are written and spoken. This paper, therefore, presents a comprehensive review and correction of the public FLORES evaluation datasets for four African languages: Hausa, Northern Sotho, Xitsonga and isiZulu. We also provide the corrected datasets for future evaluation tasks[2].

2 The FLORES Evaluation Dataset

The FLORES evaluation dataset consists of the first FLORES-101 (Goyal et al., 2021) and the subsequent more expanded FLORES-200 (NLLB Team et al., 2022) that included more languages.

[1]https://lacunafund.org/

[2]https://github.com/dsfsi/flores-fix-4-africa

Proceedings of the Ninth Conference on Machine Translation, pages 570–578
November 15-16, 2024

FLORES-101: This was the original evaluation data and was created by translating the English dataset collected from Wikipedia, consisting of several topics and domains, into 101 mostly low-resource languages. The dataset was the first available evaluation benchmark for several low-resource languages and it enabled the evaluation of many-to-many translation systems without pivoting through another high-resource language such as English. Several quality control mechanisms were put in place to ensure that the final dataset was of acceptable quality. To determine if translations are good enough for inclusion in FLORES-101, a 20% sample of the dataset were reviewed by language-specific evaluators who assess the quality using a Translation Quality Score (TQS) on a 0 to 100 scale, with a score of 90% deemed acceptable. Errors such as grammar, punctuation, spelling, and mistranslation were examined, and each was assigned a severity level of minor, major, or critical. Three of the four languages in this paper were included in this dataset–Hausa (`hau`), Northern Sotho (`nso`) and Zulu (`zul`).

FLORES-200: This dataset expanded FLORES-101 to over 200 languages, including our fourth target language–Xitsonga (`tso`). In this data, a more comprehensive process was developed to ensure the quality of the translations. Specifically, professional translators and reviewers aligned on language standards before the translators translated the sentences. Afterwards, automated checks were first conducted and then followed by manual checks by independent reviewers. Translations that were found lacking quality were sent back for post-editing. Similarly to FLORES-101, translations scoring above 90% TQS were included in the FLORES-200.

2.1 Problems Identified in FLORES

Prior to this work, we have not found any published work that carefully reviews and attempt to correct mistakes in the FLORES evaluation dataset. However, some issues have been raised on the FLORES' public GitHub repositories.[3] Some of these issues include near-identical translations in several dialects of Arabic: Mesopotamian (`acm_arb`), Ta'izzi-Adeni (`acq_arb`), Najdi (`ars_arb`), and Moroccan (`ary_arb`) Arabic dialects were found to be too similar to Standard Arabic (`arb`),[4,5] unspecifying the "orthography" and "variety" used in Lombard (`lmo_latn`) and Sardinian (`srd_latn`),[6,7] unmatched quotation marks,[8] and using Mandarin Chinese in Traditional Chinese Script (`zho_Hant`) for Cantonese (`yue_Hant`) translations.

3 Focus Languages and Evaluation

3.1 Languages Covered

In this work, the public[9] FLORES dev and devtest splits of Hausa, Northern Sotho (Sepedi), Xitsonga and isiZulu were reviewed and corrected by native speakers of the languages. A description of each language is presented in Appendix A.

3.2 Correction Guidelines

For reviewing and subsequently correcting the identified errors in the datasets, the participants were given the following guidelines.

Reviewing: At this stage, the participants identified sentences in both data splits that require reviewing.

- **Read the original text:** carefully read the original English text to understand the intended meaning and context.
- **Compare with translated text:** compare each sentence or phrase in the original English text with its corresponding translation. Pay attention to both the overall meaning and the nuances of the language.
- **Check for accuracy:** look for errors, inaccuracies, or deviations from the original meaning in the translation. This includes mistranslations, omissions, additions, and grammatical mistakes.
- **Evaluate clarity and cohesion:** assess whether the translated text is clear and coherent in the target language. Ensure that it flows naturally and is easy for a target language-speaking audience to understand.

[3] https://github.com/openlanguagedata/flores

[4] https://github.com/openlanguagedata/flores/issues/8

[5] https://github.com/facebookresearch/flores/issues/64

[6] https://github.com/openlanguagedata/flores/issues/5

[7] https://github.com/openlanguagedata/flores/issues/6

[8] https://github.com/facebookresearch/flores/issues/36

[9] https://github.com/openlanguagedata/flores

Correcting the translations: To correct the translations, we followed the guidelines provided in the shared task description.[10] The participants were trained on and encouraged to follow these guidelines when correcting the identified incorrect translations.

3.3 The Annotators

The correction task was conducted by volunteer annotators that focused on their native languages. These annotators were a mix of university students and researchers holding first, second and third degrees in computing and linguistics.

3.4 Evaluating the Corrections

To determine the amounts of corrections and subsequent differences between the original and corrected data, we used the following metrics. The computations were conducted only on the instances that were corrected. We used the original dataset as the supposed predictions and for the reference translations, we used the corrected data. We used the Natural Language Toolkit (NLTK) (Bird and Loper, 2004) for all tokenization.

Token Difference: This is the difference between the number of all tokens in the original and corrected datasets.

Token Divergence: This was used to measure the difference or dissimilarity between two sets of tokens. Given T_o and T_c as the set of tokens in the original and corrected datasets respectively, the following formula was used:

$$\text{divergence} = \frac{|T_o - T_c| + |T_c - To|}{|T_o \cup T_c|} \quad (1)$$

The formula calculates the proportion of tokens that are different between the two sets relative to the total number of unique tokens across both texts. Higher divergence score indicates that the two texts are quite different, suggesting significant changes or corrections were made.

Translation Edit Rate: (Snover et al., 2006) is a metric used in machine translation and other natural language processing tasks to measure the number of edits required to change a system-generated translation into a reference translation, and is computed using the following formula.

$$\text{TER} = \frac{\#\text{ of edits}}{\#\text{ of words in ref. translation}} \quad (2)$$

The fewer the edits, the better the translation quality and a higher TER score indicates lower quality in the predicted translations.

BLEU: (Papineni et al., 2002) is an n-gram based metric that indicates the quality of generated machine translations. The BLEU is computed as follows:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (3)$$

where BP is the Brevity Penalty and is used to penalize instances where shorter translations are generated when the reference is comparably longer; p_n is the precision between the candidate translation and a set of ground truths; and w_n is the n-gram weights.

COMET: (Rei et al., 2020) is a metric that leverages pre-trained neural models and cross-lingual word embeddings to evaluate the quality of machine translation systems. We used the pre-trained models provided by Wang et al. (2024).

4 Error Analysis

Tables 1 and 2 present how similar, or different, the original sentences were to the corrections. Some of the errors found are analyzed below per language.

Hausa (hau) A significant part of the translations were suspected to have been automatically generated, as many of them appeared incoherent or unclear. To investigate this, we conducted a comparison with translations from the Hausa FLORES dataset and new translations generated by Google Translate. The comparison revealed that, although there were limited exact matches[11], several incorrect lexical choices in the dataset's translations aligned with those produced by Google Translate, supporting the suspicion that the translations may have been automatically generated. It is important to note that other translation tools may exist for Hausa that we did not evaluate. Furthermore, several sentence-level translations from Google Translate were found to be more qualitative and coherent

[10] https://oldi.org/guidelines#translation-guidelines

[11] Google Translate may have evolved since the creation of the dataset.

lang	dev (997 sentences)					devtest (1,012 sentences)				
	# corr. (%)	# tokens$_o$	# tokens$_c$	Δ tokens	% div.	# corr. (%)	# tokens$_o$	# tokens$_c$	Δ tokens	% div.
hau	632 (63.4)	17,948	18,073	125	24.7	70 (6.9)	2,006	1,978	28	49.2
nso	67 (6.7)	2,226	2,271	45	28.9	62 (6.1)	2,082	2,105	23	28.0
tso	-	-	-	-	-	83 (8.2)	2,919	2,947	28	27.4
zul	190 (19.1)	3,605	3,588	17	23.7	226 (22.3)	4,414	4,396	18	31.8

Table 1: Data statistics; # corr. (%) $\rightarrow$ number of sentences requiring at least one correction (percentage of original data); # tokens$_o$ $\rightarrow$ original token count; # tokens$_c$ $\rightarrow$ corrected token count; Δ tokens $\rightarrow$ token count difference; % div. $\rightarrow$ percentage of token divergence.

lang.	dev				devtest			
	TER		BLEU	COMET	TER		BLEU	COMET
	Score	# Edits			Score	# Edits		
hau	19.2	3,107	72.0	54.1	40.4	711	56.6	42.1
nso	22.4	472	68.5	55.2	21.2	409	71.8	55.9
tso	-	-	-	-	20.9	547	73.9	58.4
zul	17.2	524	76.3	53.0	23.6	879	70.6	53.0

Table 2: Similarities between the original and corrected FLORES evaluation data on the four African languages – original as predictions; corrected as reference translations.

than those in the current dataset. For an illustration, we examine sentences from the dev and devtest sets, see Table 3.

In several instance, named entities were translated instead of reusing them as they are due to the lack of their equivalents in Hausa. This is illustrated in the first example provided in Table 3. Planned Parenthood appears as an organization that was not supposed to be translated (and may only be explained as *hukuma mai kula da tsarin iyali*). The words in the organization name were translated as *Iyayen Tsararru*, with their literal word translations (*iyaye* $\rightarrow$ parents, *tsararru* $\rightarrow$ planned) instead of the name of the organization as a named-entity. In the second example, the phrase "standard business attire" was translated as *Kaya masu kala ɗaya su ne cikakkun tufafin mu'amala* instead of *kayan sawa na aiki da aka saba dasu*. The first translation is at best an incorrect explanation of the English phrase. And these are just two examples of the many we found in the dataset.

In addition to these severe mistakes, the dataset was littered with a lot of inconsistencies especially in the use of the standardized Hausa alphabets. Special characters are often ommitted and instead replaced with their normalized equivalents, e.g., ɓ $\rightarrow$ b, ɗ $\rightarrow$ d, etc. In some few places, the special ƴ is written as 'y which is acceptable.

Northern Sotho (nso) Several key challenges and areas for improvement were identified and corrected, focusing on vocabulary consistency, syntax, spelling, and the accurate conveyance of technical terms. Most of the text was accurately translated and, for the text with problems, only small changes were required to make it more accurate. Some of the words like "*safatanaga* and *disafatanaga*" have generally maintained lexical consistency although they were wrongly translated. These have been corrected to "*sefatanaga* or *difatanaga* (plural)".

Although sometimes Sepedi uses borrowed words for many technical and scientific terms, things such as pavement do have a translation which could be "*tsela ya maoto or tselanathoko*". These could have been used instead of borrowing the pavement term to say *pabamente*. The use of a borrowed term could have been from the available corpus or from learned behaviour for borrowing unknown English terms. Another example is the word college which was translated to *colleje*, but Sepedi has a standard borrowed translation: "*kholetšhe*".

Addressing spelling errors and ensuring proper spacing between words are vital for readability and comprehension. For instance, the word "*tswarelo*" was corrected to "*tshwarelo*" to reflect the proper spelling. Similarly, "*patlaladitše*" was adjusted to "*phatlaladitše*", and "*bontša*" to "*bontšha*". Additionally, "*mephutso*" should be spelt as "*meputso*", and "*delo*" should be corrected to "*selo*". Spacing was required when using "*begona*" so that it is "*be gona*" and similar adjustments were made. These adjustments are crucial to maintain lexical consis-

SN	English	Wrong Translation in FLORES	Corrected Translation
1.	Komen's policy disqualified Planned Parenthood due to a pending investigation on how Planned Parenthood spends and reports its money that is being conducted by Representative Cliff Stearns.	Manufar Komen ta hana Iyayen Tsararru sanadiyyar binciken kashe kudi kan yadda Tsararren Iyaye yake ciyarwa kuma ta ba da rahoton kuɗaɗɗinta wanda Wakilin Cliff Stearns ke gudanarwa.	Manufar Komen ta dakatar da chanchantar Planned Parenthood sanadiyyar binciken da akeyi akan yanda Planned Parenthood take kashewa da kuma bayar da ba'asin kuɗin ta wanda Wakili Cliff Stearns yake gudanarwa.
2.	Suits are standard business attire, and coworkers call each other by their family names or by job titles.	Kaya masu kala ɗaya su ne cikakkun tufafin mu'amala, kuma abokan aiki kan kira junansu da sunan iyalinsu ko da muƙaman aiki.	Kwat sune kayan sawa na aiki da aka saba dasu kuma abokan aiki suna kiran juna ne da sunan gidansu ko kuma matsayin da mutum yake kai.

Table 3: Some Hausa Examples of incorrect and inconsistent translations in FLORES dev and devtest.

tency and to ensure that translations are accurate and easily understood.

Some terms were left out, like "scientific" as "*tša bo ramahlale*" when scientific tools were talked about, and this greatly affected the meaning of the sentence. Additionally, in another instance, a sentence describing the use of Caesarean section to give birth to Nadia was misleading. Incorrectly, it implied that Nadia was both the baby being born and the individual undergoing the operation. This was corrected to have the intended meaning.

Xitsonga (`tso`) Some of the problems identified in the Xitsonga translations included problems to do with vocabulary accuracy and the use of borrowed words. Among the errors that were identified is the translation of "Type 1 diabetes" to "*vuvabyi bya chukela bya Type 1*". The correct phrase should therefore be "*vuvabyi bya chukela bya muxaka wo sungula*", which captures the type of diabetes and avoid misunderstanding. Similar trends raise the importance of using proper terms that might fit local context as opposed to directly translating English words.

Another problem was that translations were mostly uniform, without contextual variations. Even here, the words "*xiyenge xa tlilinikhali na sayense*" (clinical and scientific division) were used wrongly. The word actually is "*xiyenge xa vutshila ni ntokoto bya sayense*" (clinical and scientific division), but this clearly passes on the intended meaning. Moreover, the use of pluralization of terms was arbitrary. While the singular form of the term "worker" is "*mutirhi*", the plural form should be "*vatirhi*", and the singular form of "methods" is "*maendlelo*", which should be in plural throughout instead of appearing in single forms.

Spelling problems and the usage of borrowed terms can have a substantial influence on the correctness of Xitsonga translations. One of the most illustrative examples of such incongruity of terms is that the English word "channel" has been translated as "*chanele*". Instead, the work should have used the original term "*nongonoko*" in order to ensure a perfect linguistic and connotative translation. To avoid generation of wrong impressions, the phrase borrowed from IsiZulu as used to mean "President" had to be replaced by the word "*murhangeri wa tiko*" from Xitsonga. Deficient spelling, as in the case of writing "*dokodela*" instead of "Dr", and examples of slang such as using "*mwana wa*" instead of the formal "*muongori*" indicate how borrowing and spelling mistakes reduced the quality of the translations. Fluency and correct spelling as well as using the native language correctly are a necessity to maintain the translated material's effectiveness.

isiZulu (`zul`) Similar to the errors identified in the other languages above, isiZulu translations displayed several common challenges. These included inconsistencies in vocabulary, syntax errors, and issues with the accurate expression of technical and scientific terms. The agglutinative nature of isiZulu and its conjunctive writing style further worsen these issues, leading to specific translation errors related to morphology and orthography.

A closer examination of these challenges reveals issues such as in the translation of "Around 11:29, the protest moved up Whitehall, ..." which was initially rendered as "*Ngawo-11:29 ababhikishi baya Odongeni Olumhlophe, ...*". This translation contains two key issues. First, "*Ngawo-11:29*" should have been "*Ngabo-11:29*" to correctly match the grammatical structure for time

expressions in isiZulu. Second, the literal transliteration of "Whitehall" as "*Odongeni Olumhlophe*" failed to integrate properly into the sentence. The correct approach would involve incorporating the place name with the locative prefix "e-" to produce "e-Whitehall.". This prefix addition is required in conjunctive languages when using borrowed words or terms, but MT systems often fail to capture these variations. Additionally, another common issue was the unnecessary borrowing of words from English, despite the availability of standardized isiZulu terms. This was particularly evident with month names, scientific terms, and country names, where inconsistencies were frequent—one translation might use "January," another "*uJanuwari*," and yet another "*uMasingana*" Another example of this can be seen with the country name "Spain," which was inconsistently translated as both "Spain" and "*Speyini*" in different sections. Similar inconsistencies occurred with attempts to translate organizational names or acronyms, leading to partial translations that disrupted the linguistic flow.

To address the inconsistencies, standardized isiZulu terms were consistently applied throughout the translations. For instance, month names such as "*uMasingana*" replaced the inconsistent use of "January" and "*uJanuwari*" In dealing with organizational names and acronyms and countries' names, the approach was to fully translate these entities or retain their original form consistently, avoiding partial translations that could disrupt the flow.

In addition to the inconsistencies with terminology, other errors were also identified and addressed. These included issues with verb conjugation, where incorrect tenses or forms were initially used, and the improper handling of borrowed words that did not align with isiZulu's morphosyntactic rules. Minor spelling errors and incorrect use of prefixes or suffixes were also corrected to ensure that the translations were both grammatically accurate and easily understood.

5 Conclusion

In this work, we highlight the importance of qualitative evaluation datasets for low-resource languages and present our findings from a comprehensive review of the FLORES dataset for four African languages: Hausa, Northern Sotho, Xitsonga, and isiZulu. The original translations were marred by vocabulary inconsistencies, syntax errors, and inaccurate technical terms. After making necessary corrections, we measured the amount of edits and resulting difference between the improved datasets and the original using metrics like BLEU, TER, and COMET, which showed that significant improvements were made. The results presented highlight the need for ongoing refinement and human oversight in developing accurate translation datasets for underrepresented languages. For future work, we intend to expand the corrections to more African languages.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, 25 edition. SIL International, Dallas, Texas.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Kr-

ishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Ximbani Eric Mabaso. 2018. Xitsonga in south africa. *The Social and Political History of Southern Africa's Languages*, pages 311–330.

Stuart Mesham, Luc Hayward, Jared Shapiro, and Jan Buys. 2021. Low-resource language modelling of south african languages. *arXiv preprint arXiv:2104.00772*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Derek Nurse and Gérard Philippson. 2006. *The bantu languages*, volume 4. Routledge.

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.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. 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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

StatsSA. 2022. Statistics South Africa.

Elsabé Taljard and Sonja E Bosch. 2006. A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written bantu languages. *Nordic journal of African studies*, 15(4).

Jiayi Wang, David 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 Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, 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.

A Description of the Target Languages

Hausa (hau): Hausa is a widely spoken language across West Africa, particularly in Nigeria, Niger, Cameroon, and Ghana. It is spoken by approximately 77 million people worldwide, primarily in West Africa (Eberhard et al., 2022). Hausa ranks as the second most spoken language in Africa and 27th globally. The language belongs to the Chadic branch of the Afroasiatic language family, and it has a rich history of written communication. It was first written in Arabic script known as Ajami, reflecting the language's connection to Arabic, with many Hausa words borrowed from Arabic due to historical contact and influence. Today, the Boko script (also known as Roman script), which uses Latin characters, is the most common writing system for Hausa. This script excludes the letters p, q, v, and x, and includes additional consonants (ɓ, ɗ, ƙ, y, kw, ƙw, gw, ky, ƙy, gy, sh, ts) and vowels (long a, i, o, u, e, and two diphthongs ai and au). Hausa follows a Subject-Verb-Object (SVO) sentence structure.

Northern Sotho (nso): Northern Sotho, also known as Sepedi or Sesotho sa Leboa, is one of the official languages of South Africa and is spoken primarily by the Bapedi people in Limpopo Province. It is a Bantu language that belongs to the Sotho-Tswana group and shares linguistic similarities with Sesotho (Southern Sotho) and Setswana. Sepedi is known for its rich oral tradition that includes folklore, proverbs, and praise poetry that have played a significant role in the preservation of cultural heritage (Nurse and Philippson, 2006). Sepedi is written using the Latin alphabet, with the standard 26 letters and a few additional characters such as the "š" which are adapted to its unique sounds. The language primarily follows a Subject-Verb-Object word order in sentence structure.

Xitsonga (tso): Xitsonga, or Tsonga, is a Bantu language that is mainly spoken in South Africa and more especially in the Limpopo province and parts of Mpumalanga province. The language is estimated to be spoken by about 2.3 million people in South Africa. Xitsonga belongs to the Niger-Congo language family, specifically the Tshwa-Ronga subgroup, and is characterized by the extensive use of prefixes and suffixes to convey meaning (Mabaso, 2018). This linguistic feature can impact the accuracy of translations, especially when dealing with technical and scientific concepts. It also feature a complex system of writing and syntax, which are prerequisites to clear and concise language usage. Xitsonga is currently used in education and media section in South Africa, thus is regarded as relevant in cultural linguistic practices. That is why, the language being mentioned as a part of the country's multiple languages system emphasizes its relevance and application in different phases of the people's activity.

Language	Sentence
English	I know them
Hausa	Na san su
Northern Sotho	Ndza va tiva
Xitsonga	Ke a ba tseba
isiZulu	Ngiyabazi

Table 4: The grammatical structure of different languages.

isiZulu (zul): Zulu or isiZulu (in Zulu) is one of the 12 official languages in South Africa, and it is considered to be the most widely spoken indigenous language in the country. It constitutes approximately a quarter of the population, with around 15.1 million speakers out of the population of 62 million people (StatsSA, 2022). IsiZulu is part of the Nguni language family, which is made up of a group of closely related Bantu languages belonging to a larger Niger-Congo language family, and they are widely spoken across Southern Africa (Mesham et al., 2021). These languages are particularly notable for their complex morphology, characterized by agglutinative morphology and conjunctive orthography. Agglutinative morphology means that words are typically formed by combining multiple small meaning-carrying units, known as morpheme. Conjunctive orthography means that the morphemes are glued together to form a word, rather than writing them with spaces in between, as seen in disjunctive orthography, commonly associated with the Sotho group, as well as Tshivenda and Xitsonga in South Africa indigenous languages (Taljard and Bosch, 2006). To illustrate this distinction, consider the example in Table 4 which examines the different grammatical structures of the phrase *I know them*.

Table 4 shows that while the phrase's meaning is consistent across languages, the writing systems vary: in disjunctive orthography, morphemes are

separated by spaces, while in conjunctive orthography, as in isiZulu, they are joined into a single word. For example, in the phrase *I know them*, each morpheme serves a specific grammatical function– 'I' as the subject, 'know' as the verb, and 'them' as the object. In disjunctive orthography, these morphemes are written separately, making each unit distinct. In conjunctive orthography, they are combined into one continuous word, but the meaning remains intact. These orthographic variations pose challenges for machine translation systems, which must accurately process morphemes in different writing systems to produce accurate translations.