Evaluating Native-Speaker Preferences on Machine Translation and Post-Edits for Five African Languages

Hiba El Oirghi[1] **Tajuddeen Gwadabe**[2] **Marine Carpuat**[1]
[1]University of Maryland, College Park [2]Masakhane Research Foundation
eloirghi@umd.edu

Abstract

Wikipedia editors undertake the task of editing machine translation (MT) outputs in various languages to disseminate multilingual knowledge from English. But are editors doing more than just translating or fixing MT output? To answer this broad question, we constructed a dataset of 4,335 fine-grained annotated parallel pairs of MT translations and human post-edit (HE) translations for five low-resource African languages: Hausa, Igbo, Swahili, Yoruba, and Zulu. We report on our data selection and annotation methodologies as well as findings from the annotated dataset, the most surprising of which is that annotators mostly preferred the MT translations over their HE counterparts for three out of five languages. We analyze the nature of these "fluency breaking" edits and provide recommendations for the MT post-editing workflows in the Wikipedia domain and beyond.

1 Introduction

The rapid expansion of Wikipedia content in low-resource, underserved African languages is heavily dependent on the accuracy of ContentTranslation[1], Wikipedia's out-of-English MT tool. However, translation quality remains inconsistent, especially for low-resource languages where MT does not adequately support both linguistic diversity and cultural suitability (Orife et al., 2020).

While standard Machine Translation (MT) evaluation metrics such as COMET (Rei et al., 2020), AfriCOMET (Wang et al., 2024), and xCOMET (Guerreiro et al., 2024) typically assume that human post-edits (HE) are inherently superior to raw MT output, the Wikipedia editing environment challenges this notion. Editors often work under time pressure, with varying levels of bilingual proficiency, and may prioritize encyclopedic formatting over translational fidelity.

This paper investigates how MT and HE translations differ on an aggregate segment level as well as on a fine-grained pairwise difference level. We present the following contributions:

- A curated dataset [2] of 4,335 English source, MT output, and HE output parallel segments, fully annotated for preference and error types;
- Empirical evidence that native speakers frequently prefer MT over human edits, driven largely by "fluency breaking" behavior in the post-editing process;
- Actionable recommendations for Wikipedia language communities and MT researchers.

2 Data and Annotation

2.1 Data Source and Global Statistics

We extract parallel English source, machine translation (MT) output, and human post-edit (HE) output segments from the 06/13/2025 Wikipedia data dumps[3] for the five following out-of-English Language Pairs (LPs): Hausa (eng-hau), Igbo (eng-ibo), Swahili (eng-swa), Yoruba (eng-yor), and Zulu (eng-zul).

Table 1 provides a statistical overview of this initial dataset. While Hausa and Igbo represent the largest corpora, a more telling metric is the Levenshtein character-level edit distance (Levenshtein, 1965) between their MT and HE pairs. Most importantly, all five languages have a null or near-zero average segment-level difference between the AfriCOMET (Wang et al., 2024) Quality Estimation scores of their MT and HE pairs. This surprising observation—that a quality estimation metric detects little to no quality differences after human editing—is the primary motivation for our deeper analysis.

[1] https://en.wikipedia.org/wiki/Wikipedia:Content_translation_tool

[2] https://github.com/hibaeloirghi/Wiki-Data

[3] https://dumps.wikimedia.org/other/contenttranslation/

LP	Edit Dist. (MT vs. HE)	AfriCOMET (MT)	AfriCOMET (HE)	AfriCOMET Diff	Token Diff
eng-hau (n = 177,387)	40.39	0.65	0.66	0.01	19.34
eng-ibo (n = 200,161)	89.35	0.59	0.58	0.00	38.71
eng-swa (n = 10,100)	93.50	0.72	0.72	0.00	-7.41
eng-yor (n = 9,988)	91.97	0.56	0.58	0.02	24.45
eng-zul (n = 7,319)	40.77	0.67	0.66	-0.01	-3.01

Table 1: Global statistics comparing machine translation (MT) and human-edited (HE) Wikipedia segments across five African languages. n denotes the number of parallel segments found in the 06/13/2025 Wikipedia dump. Global AfriCOMET (Wang et al., 2024) Difference refers to the difference between the HE and MT AfriCOMET scores. Global Edit Distance represents the Levenshtein edit distance (Levenshtein, 1965) between the MT and HE pairs. Global Token Difference refers to the word-count difference between the MT and HE pairs.

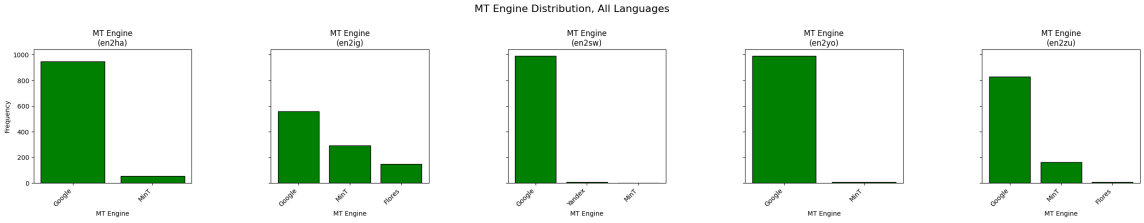


Figure 1: Distribution of MT engines used for out-of-English MT in the selected 5,000 segments for all five African languages.

2.2 Curation of the Annotation Dataset

From the large corpora described in §2.1, we curated a subset of 1,000 representative source-MT-HE triplets per language for annotation and analysis. To ensure the selected segments where informative and suitable for a detailed comparison, we applied the following filtering criteria:

- **Bounded Quality Difference:** The quality score differences between HE and MT must be between -0.5 and 0.5, as measured by AfriCOMET QE. We exclude segments with zero differences to avoid pairs with minimal variation.
- **Meaningful Edit Distance:** The character-level Levenshtein distance must be between 5 and 40 to filter out both cases with insignificant changes (e.g., punctuation edits) and complete re-translations which are hard to compare directly.
- **Sufficient Source Length:** The English source segments must contain at least 20 words to ensure segments are substantial enough to annotate.
- **High Absolute Quality:** The AfriCOMET QE score for both HE and MT must be greater than 0.5 to focus the analysis on higher quality translations.
- **Data Cleaning:** We remove duplicate segments and segments with excesive special characters (e.g. !]>*), which often signal lower quality segments.

We observe a diverse mix of MT engines in the source data (including Google Translate, NLLB, and potentially others), as illustrated in Figure 1.

2.3 Annotation Protocol

We used a customized version of the Appraise tool (Federmann, 2018; Kocmi et al., 2024)[4] to deploy our annotation protocol[5]. We recruited three native speakers per language through Masakhane[6] to annotate the same segments. Crucially, the identities of the candidates (MT vs. HE) were masked on the annotation interface. Annotators were asked to select the better translation and justify their choice using a fine-grained span-level mapping including labels for *Fluency*, *Adequacy*, and *Explicitation*. Appendix A shows a collage of screenshots of our annotation interface.

Inter-annotator Agreement Table 2 details overall high Fleiss' κ inter-annotator agreement (IAA) scores for the aggregate preference task. IAA is

[4] https://github.com/AppraiseDev/Appraise

[5] https://github.com/hibaeloirghi/Appraise-wiki

[6] https://www.masakhane.io/

LP	Fleiss' κ	N Segments
eng-hau	0.83	404
eng-ibo	0.72	1000
eng-swa	0.37	1000
eng-yor	0.57	835
eng-zul	0.70	1000

Table 2: Fleiss' κ (Fleiss, 1971) inter-annotator agreement scores for aggregate HE vs. MT preference across languages, considering only segments with exactly three annotators.

substantial for Hausa (0.83), Igbo (0.72), and Zulu (0.70) and moderate for Yoruba (0.57). However, IAA is lower for Swahili (0.37), indicating that judging overall translation quality was a more subjective task in this language context.

3 Findings

Our analysis of the annotated data reveals a complex and often counterintuitive relationship between MT and HE, challenging the assumption that post-editing Wikipedia content targets improvements against the original MT (see Figure 2). In Zulu, for instance, MT was preferred in 86% of cases, and in Igbo, 62%. In Yoruba, MT is preferred 27% of the time and in barely more than 4% of cases in Hausa. This variety suggests that the assumption of post-edit supremacy does not hold for the Wikipedia post-editing domain. Figures 3 and 4 summarize the distribution of factors cited for translation superiority and inferiority, respectively, across all five languages in our study. Appendix B contains a detailed key of the advantages and disadvantages selected by the annotators. We summarize findings and implications of those below.

The preference trends described above hold steady when we restrict our analysis to examples with a clear majority preference by filtering out segments where at least two out of three annotators spotted no difference between the candidate translations (selected "NoDiff" meaning no real difference in fluency or meaning between translation candidates as detailed in Appendix B). As shown in Figure 6 this filtering had a significant impact on the number of remaining segments, where a large number of post-edits are assessed as having no effect on quality.

3.1 Why is a translation better?

Our analysis reveals that **fluency** is overwhelmingly the dominant driver of annotator preference.

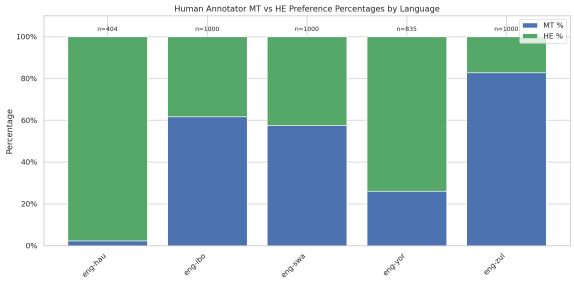


Figure 2: Human annotator MT vs. HE preference percentages. For Igbo, Swahili, and Zulu, annotators mostly prefer MT.

In all languages, the most frequently cited factors for preferring one translation are "Fluency_Natural" and "Fluency_Grammar_Spelling." Interestingly, machine translation (MT) outputs often outscore human post-edits on fluency grounds in Zulu, Swahili, and Igbo. In contrast, human-edited output tends to surpass MT in Yoruba and Hausa for fluency.

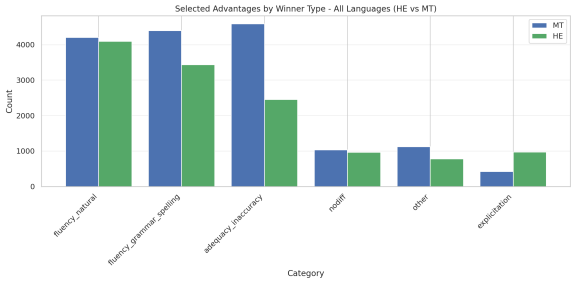


Figure 3: Distribution of justifications for the preferred translation. Fluency dominates the decision-making process.

Adequacy plays a complex role. Contrary to common assumptions, MT is also frequently cited as more adequate than the human post-edits—in all languages except Hausa. This suggests that in some cases, human interventions introduce errors or omit important information that MT maintains. Explicitation (the addition of beneficial context or clarifications) emerges as a notable factor, especially in Hausa, Swahili, and Yoruba. Finally, segments with "NoDiff" (no real difference in fluency or meaning between candidates as detailed in Appendix B) form a substantial subset in most languages, except for Igbo, where differences are more often perceived.

These patterns, visualized in Figure 3, underscore that *language-specific strategies may be necessary to improve translation workflows*. They also suggest that some post-edits may be superficial or

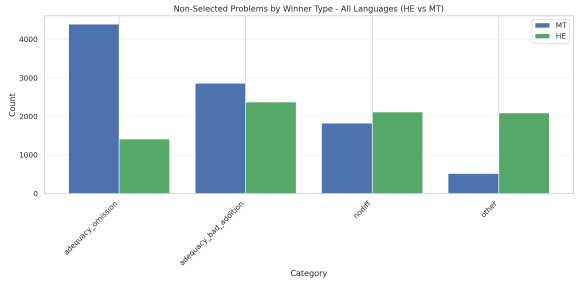


Figure 4: Distribution of factors cited by annotators for why a translation is judged worse than its counterpart. Post-edits often suffer from omissions or bad additions.

even detrimental to translation quality, particularly when editors make changes that reduce fluency or adequacy.

3.2 Why is a translation worse?

When annotators identify why a translation is inferior, a distinct language-dependent error profile emerges (Figure 4). Human editors are most frequently identified as omitting information present in the source, a pattern that is striking in Igbo, Swahili, and Zulu. In addition to omissions, human editors also sometimes add or elaborate information not present in the source text.

By contrast, the primary weaknesses of MT translations are less often tied to adequacy errors; more frequently, annotators cite general quality issues ("Other" or "NoDiff"). This indicates that, while MT typically remains closer to the original content, its output sometimes lacks the refinements or contextual adaptions made by human editors, though such adaptions may not always improve fidelity.

Collectively, these results suggest that MT is generally more **faithful to the source**, while humans are more likely to commit major errors per established MQM categories, particularly omissions or problematic additions. This raises important questions: Do human editors sometimes aim for goals beyond faithful translation, such as localization, summarization, or re-writing? If so, should such edits be distinguished in Wikipedia workflows (e.g., with a tag different from the standard "HE" marker), and how should downstream systems recognize and support these distinctions?

4 Recommendations and Implications

Our analysis strongly suggests that "post-editing" MT content on Wikipedia is not a monolithic task of error correction. Editors often engage in more complex activities that go beyond faithful translation, such as adding or removing content.

While these adaptations can be valuable, our analysis show they are also risky as they can degrade fluency. We offer the following actionable insights for the Wikipedia and the African NLP communities:

- **Distinguishing Editorial Roles:** Post-editing workflows could benefit from distinguishing between different types of post-edits. The workflow could be enhanced by allowing editors to tag their intent, for example, distinguishing a "Faithful Correction" from a "Cultural Adaptation". This distinction is critical for quality control; otherwise, valuable adaptations may be incorrectly flagged as translation errors, and the true intention behind the editor's work is lost.

- **Develop Language Specific Strategies:** The clear differences in editing patterns and fluency outcomes between languages (as shown in Figures 7 and 8) suggest that a one-size-fits-all approach to quality control is suboptimal. Workflows need to be adapted to the specific needs of each language community.

- **Implement Fluency Checks:** This study reveals that human edits can, paradoxically, decrease the fluency of machine-translated text, an issue mainly seen in Igbo, Swahili, and Zulu. To address this, the workflow could integrate lightweight automated checks that flag potentially awkward or ungrammatical sentences in post-edited content.

5 Conclusion

Our paper investigated native-speaker human preference between machine translation and human post-editing in the Wikipedia domain for five African languages: Hausa, Igbo, Swahili, Yoruba, and Zulu. We discovered that native speaker preferences are mainly driven by perceived fluency, and human edits can have varying effects on fluency of MT outputs. These findings highlight the need for better editor training and more domain-specific, human-preference-aligned automated metrics to support the creation of Wikipedia articles in low-resource Wikipedia.

Limitations

Our study focuses on five African languages; results may not generalize to other low-resource languages with different editing communities. Additionally, our analysis relies on the judgments of three annotators per language. While we filtered for majority agreement, subjective preferences regarding "naturalness" can vary by dialect and region. Finally, we did not explicitly model the intent of the editors (e.g., distinguishing between vandalism, partial edits, and genuine corrections), which adds noise to the "Human Edit" class.

Ethical Considerations

The annotators we hired are all native speakers of the five relevant African languages. We envision a participatory approach to MT evaluation and aim for our work to help expand the field's interest and understanding of MT for low-resource languages, with the ultimate hope that this will benefit the language communities of the five languages examined in this work: Hausa, Igbo, Swahili, Yoruba, and Zulu.

Acknowledgements

The authors are grateful to Eleftheria Briakou for her contributions, and to Vilém Zouhar and Tom Kocmi who answered our questions on the Appraise open source code.

References

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane – machine translation for africa. *Preprint*, arXiv:2003.11529.

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.

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.

A Appendix: Annotation Screenshots

B Appendix: Annotation Mapping Categories

B.1 Aggregate Annotation Mapping

Why is the selected translation better? (check all that apply)

- **Fluency_Natural**: Reads more naturally in the target language, regardless of meaning preservation.
- **Fluency_Grammar_Spelling**: Contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.), regardless of meaning preservation.
- **Adequacy_Inaccuracy**: Preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the unselected translation.
- **Explicitation**: The selected translation adds helpful context or clarifications beyond what is present in the source text. The non-selected translation does not.
- **NoDiff**: No real difference in fluency or meaning.
- **Other**: Other (please specify)

What problems does the non-selected translation have? (check all that apply)

- **Adequacy_Bad_Addition**: Inaccurately introduces words/phrases not in the source text.
- **Adequacy_Omission**: Omits information present in the source text.
- **NoDiff**: No real difference in fluency or meaning.
- **Other**: Other (please specify)

Which of the two candidate translations is adequate for a Wikipedia entry, even if it is not a perfect translation? (check all that apply)

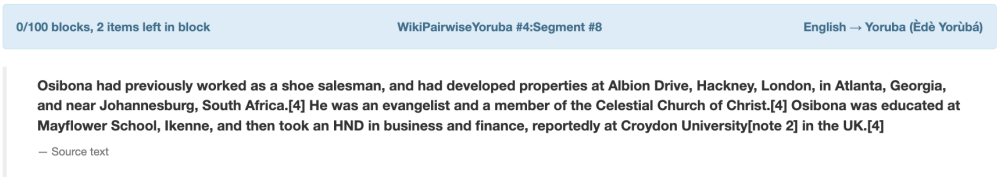
- **Wiki_Style_selected**: The selected translation.
- **Wiki_Style_non_selected**: The non-selected translation.
- **Neither**.

B.2 Span Annotation Mapping

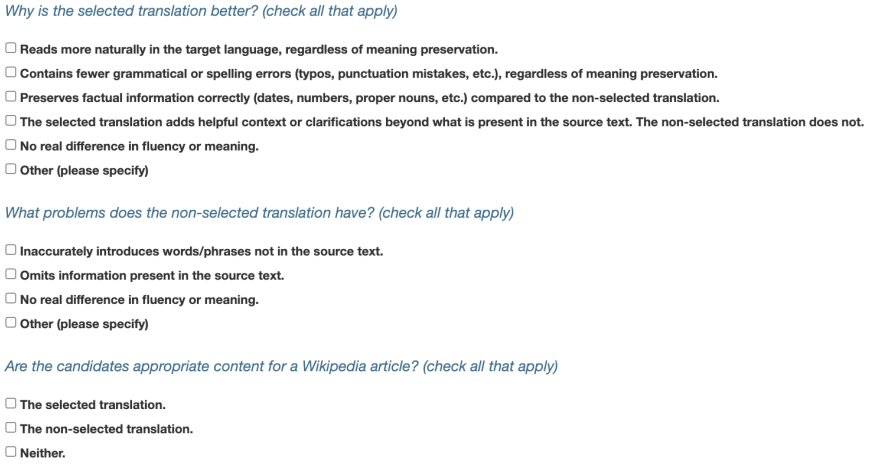
Why did you select this span? Please select one or more options below to explain your choice and share any additional thoughts.

- **Fluency_natural**: The selected span reads more naturally in the target language.
- **Fluency_grammar_spelling**: The selected span contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.).
- **Adequacy_inaccuracy**: The selected span preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the non-selected span.
- **Adequacy_untranslated**: The non-selected span is partially or fully untranslated.
- **Explicitation**: The selected span adds helpful context or clarifications beyond what is present in the source text. The non-selected span does not.
- **NoDiff**: No real difference in fluency or meaning.
- **Other**: Other (please specify)

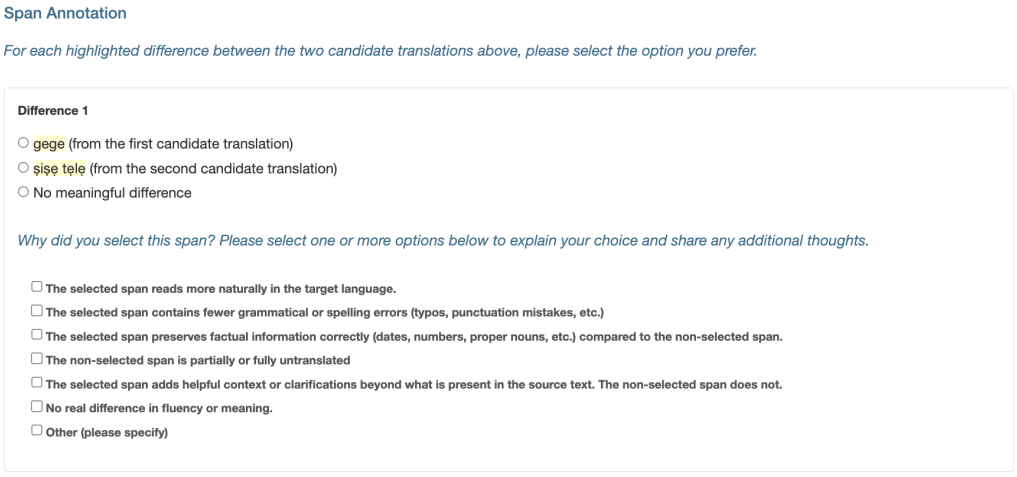
C Appendix: Filtered Dataset



(a) MT vs. HE Preference Annotation



(b) Aggregate MT vs. HE annotation



(c) Span-level annotation

Figure 5: Screenshots from one sample Yoruba segment annotation page of the annotation interface. Screenshot (a) shows the. Screenshot (b) shows two simplified MQM-style (Lommel et al., 2014) questions about the advantages and disadvantages of the candidate translations in (a). Screenshot (c) shows pairwise difference-level simplified MQM-style annotation.

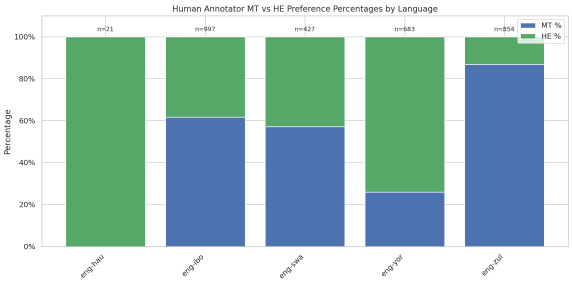


Figure 6: Human annotator MT vs. HE preference percentages. For Igbo, Swahili, and Zulu, annotators mostly prefer MT. Segments where at least two out of a total of three annotators marked that they see no difference between the two aggregate candidate translations are dropped.

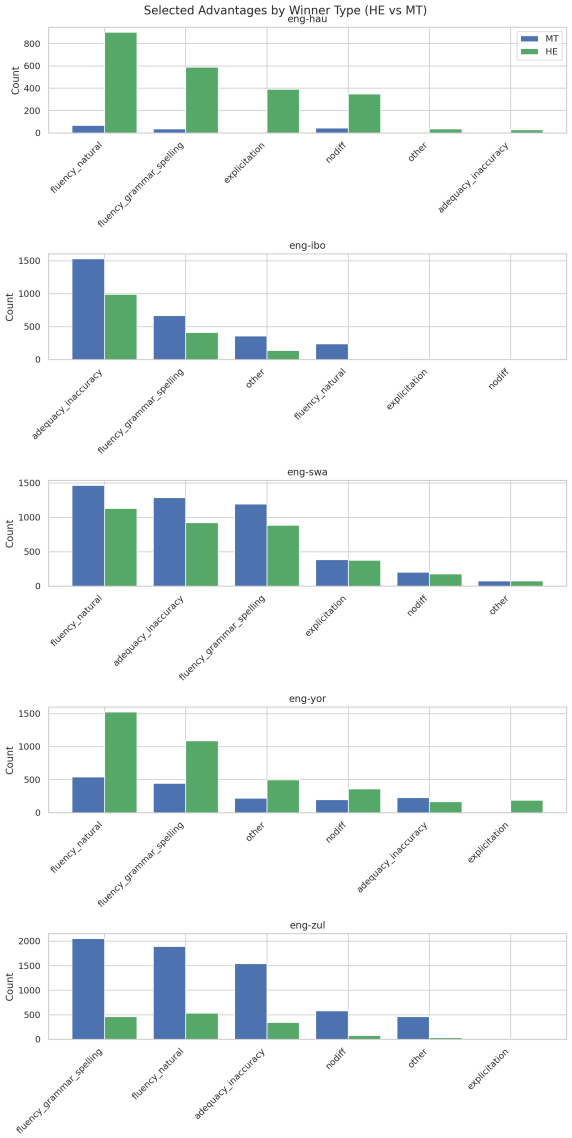


Figure 7: Distribution of justifications for the preferred translation for each language. Fluency dominates the decision-making process.

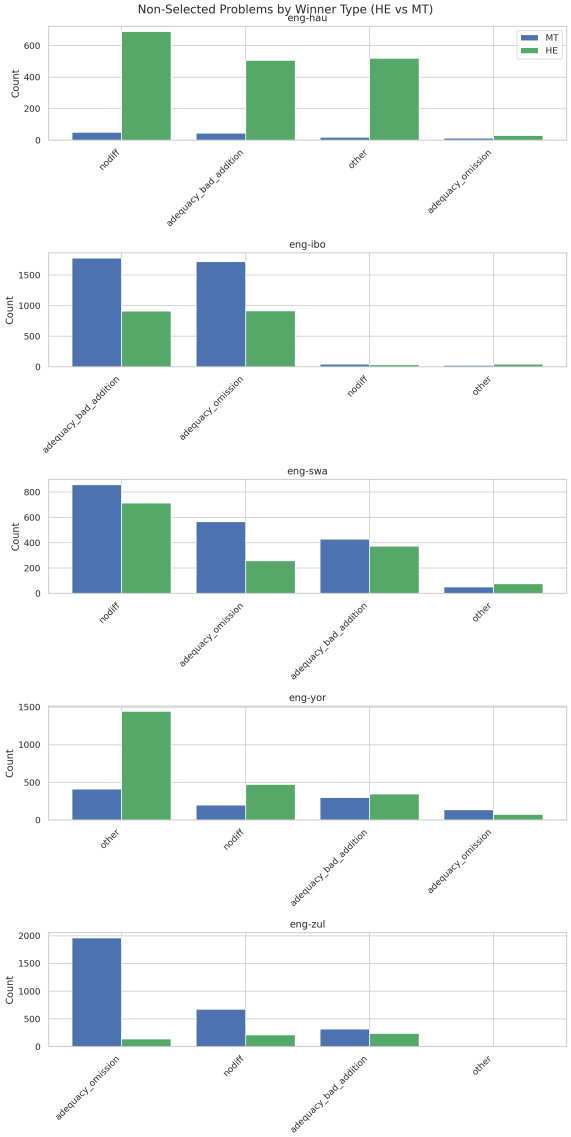


Figure 8: Distribution of factors cited by annotators for why a translation is judged worse than its counterpart for each language. Post-edits often suffer from omissions or bad additions.