Understanding How Editors Use Machine Translation in Wikipedia: A Case Study in African Languages

Eleftheria Briakou
University of Maryland

Tajuddeen Gwadabe
Masakhane Research Foundation

Hady Elsahar
Meta AI

Marine Carpuat
University of Maryland

Abstract

Our project envisions a participatory approach to understand how editors use machine translation (MT) in Wikipedia, with the end goal of helping Wikipedians develop best practices when using MT for creating content in African languages. We propose a three-step approach that factors in Wikipedia users, native speakers, and Natural Language Processing (NLP) researchers to study this question collaboratively.

Introduction

Wikipedia is clearly the largest multilingual encyclopedia, with its English edition being by far larger than any printed encyclopedia.[1] Despite the many languages it covers—i.e., 321 as of March 2023—94% of language editions contain fewer than a million articles.[2] Translating pages from one language to another appears as a promising step towards bridging the gaps in Wikipedia content across languages; however, it shifts the responsibility of content creation to multilingual editors. To simplify the translation process, *ContentTranslation* has been integrated into Wikipedia as an opt-in feature to translate articles in January 2015 (Laxström et al., 2015). The tool automates many of the tedious steps inherent in translating Wikipedia articles resulting—with the recent integration of the NLLB-200 (team et al., 2022) service—in publishing hundreds of translations into previously underrepresented African languages, such as Igbo, Hausa, Yorùbá, Swahili, and Zulu.

Despite the widespread application of the tool, the use of MT has been criticized by several Wikimedia communities,[3] while the Wikipedia consensus is that *"unedited machine translation is worse than nothing"*. To circumvent such issues, Wikipedia editions impose filters to test whether editors modify the machine translated texts.[4] However, such filters lump together cases where MT translation quality is good enough to be used as is, and cases where human editors let translation errors go through. Consider the following example of an English sentence (EN), its machine translation (MT), and a post-edited (PE) version of the machine translation provided by a Wikipedia editor in Yorùbá:

> EN: He married Bisi Towry-Coker but they are now separated. He has three children including a son, Olaotan.

[1] https://meta.wikimedia.org/wiki/Wikipedia_as_books

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] Some discussion on issues related to the tool are listed here: https://en.wikipedia.org/wiki/Wikipedia_talk:Content_translation_tool

[4] https://www.mediawiki.org/wiki/Help:Content_translation/Translating/Translation_quality

MT: O ti fẹ Bisi Towry-Coker ṣugbọn wọn ti pin bayi. O ni awọn ọmọde mẹta pẹlu ọmọ kan, Olaotan.[5]

PE: Ó fẹ́ ìyàwó rẹ̀ Bísífẹ Towry-Coker ṣùgbọ́n wọ́n ti pín yà báyìí. Ó bí àwọn ọmọ obìnrin mẹ́ta pẹ̀lú ọkùn rin kan Ọláòtán.[6]

Although the post-edited version modifies the machine translated texts significantly, it introduces content that does not exist in the source (i.e., [ìyàwó rẹ̀] which translates to [his wife] and [obìnrin] which translates to [female]), and additionally contains orthographic errors.[7]

This proposal aims to ***understand how editors use machine translation in Wikipedia*** and to design tools and data to help them make more informed decisions when editing machine translated texts in African languages. To that end, we propose a participatory approach that brings together Wikipedia users, native speakers, and Natural Language Processing (NLP) researchers and comprises three steps. **(Step I)** We will start by interviewing *Wikipedia users* to understand their needs and workflows when using machine translation in Wikipedia. For instance, what types of edits do they perform: do they primarily fix factual errors, disfluencies, or stylistic issues? **(Step II)** Drawing on insights from the previous step, we will design an annotation protocol to label such errors in machine translation output and the edited texts by recruiting *native speakers* of the studied languages. **(Step III)** The annotations collected will be used to understand the more prominent issues that arise when using machine translation in low-resource settings. This invaluable resource will be used to design guidelines and tools to help multilingual editors use machine translation more effectively.

Proposal Timeline:

Step I: Jul 1, 2023 - Sep 1, 2023
Step II: Sep 2, 2023 - Dec 1, 2023
Step III: Dec 2, 2023 - Jun 30, 2024

Related work

In the section, we start by discussing prior work on multilinguality in Wikipedia, then shift to analyzing patterns in (machine and post-edited) translations, and conclude with human and automatic evaluation of translation errors.

Multilinguality in Wikipedia Wikipedia editions differ widely across languages, not only in what they cover, but also in how they are used. Lemmerich et al., 2018 explore the readers' motivations behind using Wikipedia across several languages and show that specific Wikipedia use cases are more common in countries with certain socio-economic characteristics. Other works compare different Wikipedia editions across the dimensions of self-focus (Hecht et al., 2009), cultural bias (Callahan and Herring, 2011; Ribé and Laniado, 2018), geographical bias (Beytía, 2020), semantic similarity (Jian et al., 2017) and information covered (Samoilenko et al., 2017). Our proposal aims to explore the use of MT to bridge Wikipedia's knowledge and content gaps specifically for under-resourced African languages.

Analyzing Translation Patterns Prior analyses of patterns found in machine and post-edited translations in the MT literature focus on a small number of high-resource languages (see Figure 1). These studies have therefore

[5] Glossed as: "*He was married to Bisi Towry-Coker but they are now separated. He has three children including a son, Olaotan.*"
[6] Glossed as: "*He was married to his wife Biseffe Towry-Coker but they are now separated. He gave birth to three daughters with a rope walking in Oláotán.*"
[7] This example has been audited by a native speaker of Yorùbá.

neglected under-resource languages, even though MT achieves lower quality for these languages, and is more likely to produce problematic outputs.

Study	Languages
Garcia (2011)	Chinese
Carl et al. (2011)	Danish
Čulo and Nitzke (2016)	German
Carl and Schaeffer (2017)	German, Spanish
Daems et al. (2018)	Dutch
Toral (2019)	German, French, Spanish, Chinese
Stasimioti and Sosoni (2020)	Greek
Groves and Schmidtke (2009)	German, French
Şahin and Gürses (2019)	Turkish
Bizzoni et al. (2020)	German
Góis and Martins (2019)	French, German
Vanmassenhove et al. (2019)	French, Spanish
Fu and Nederhof (2021)	Czech, German, Russian
Guerberof-Arenas and Toral (2020)	Catalan
Corpas Pastor et al. (2008)	Spanish

Figure 1: Prior works study patterns in machine or post-edited translations focusing on high-resource languages.

Human Evaluation of Translations Standard protocols for human evaluation of machine translation have been developed for "generic" use, without accounting for specific contexts of use, such as the ones faced by Wikipedia editors. Human evaluation of machine-translated outputs was initially based on rating fluency and adequacy on a 5-point scale. This framing was adopted by the first evaluation campaign of the Conference on Machine Translation in 2006 (Koehn and Monz, 2006) and later replaced with a ranking-based (Vilar et al., 2007) evaluation schema until 2016. Subsequent evaluations were based on direct assessment on a continuous scale (Graham et al., 2013) focusing on adequacy (Bojar et al., 2017), or alternative evaluation dimensions, such as reading comprehension (Scarton and Specia, 2016) and believability (Martindale et al., 2021). More recently, human evaluation of machine translation has been based on the Multidimensional Quality Metrics (MQM) methodology carried out by skilled human annotators (Freitag et al., 2021a, b). The core idea in MQM is asking annotators to highlight each error in the (machine) translated text and label the error and its severity based on a predefined set of errors. While MQM is more accurate and informative than traditional ranking-based and direct assessment scoring methods, how to apply the framework in evaluating translations in the context of Wikipedia and in low-resource settings remains an open question that we aim to explore within the scope of this proposal.

Automatic Evaluation of Translations Recent work attempts to automatically compute the quality of a machine translated output without assuming access to a gold-standard reference translation (Rei et al., 2020; Sellam et al., 2020). Typically, those models are based on neural architectures trained with human supervision and evaluated for a handful of high-resource languages. Our own previous work shows that small meaning differences between English texts and French (human) translations can be automatically detected without human annotation (Briakou and Carpuat., 2019). In this project, we will build on this work to design and evaluate methods of quality estimation for low-resource languages.

Methods

In this section, we describe our preliminary analysis highlighting potential issues related to the use of machine translation in Wikipedias of African languages (see Preliminary Analysis) and conclude with the proposed research for the Wikimedia Research Fund 2023 (see Proposed Analysis).

Preliminary analysis

As of March 2023, the ContentTranslation tool supports out-of-English machine translation for five African languages: Igbo (IG), Hausa (HA), Zulu (ZU), Swahili (SW), and Yorùbá (YO). Based

on our preliminary analysis of Content Translation dumps from December 22 and March 2023, we see that most translated content published across all five languages is based on editing machine translation output rather than translating from scratch (Figure 2).

Crucially, when looking at the percentage of edit types—i.e., the amount of editing Wikipedia users perform on top of the machine-translated texts—we notice that, on average, the amount of editing is small (<20%) across all languages except for Yoruba that exhibits an editing rate of more than 40% (Figure 3).[8]

At the same time, we know machine translation quality varies across those languages. The NLLB machine translation service that powers translations for the African languages reports the following translation quality scores (based on BLEU (Papineni et al., 2002), which assigns higher scores to outputs that are more similar to reference human translations): 25.8 for Igbo, 33.6 for Hausa, 36.3 for Zulu, 37.9 for Swahili, and 13.8 for Yorùbá (team et al., 2022).[9] Based on those scores, one would expect Igbo translations to be more heavily edited compared to Hausa, Zulu, and Swahili, though the reported differences in Figure 3 are small.

Motivated by the above observations, we propose to conduct a more systematic analysis of the use of machine translation for African languages, as described below.

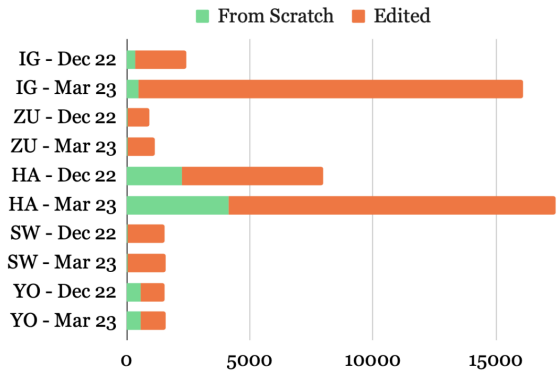

Figure 2: Number of published translations (i.e., roughly at the sentence level) for the five African languages supported by the ContentTranslation tool based on dumps of December 2022 and March 2023.

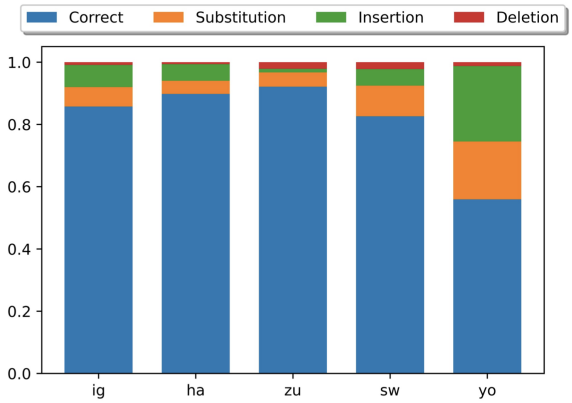

Figure 3: Percentage of edit operations when editing machine translation texts for 5 African languages (computed using TER, on Content Translation dumps of March 2023).

Proposed Research

We propose a three-step approach that brings together Wikipedia users, native speakers, and NLP researchers to understand the use of machine translation in Wikipedia (Figure 4). We propose studying the use of machine translation for 3 African languages (estimated based on maximum available budget). We will choose languages between Igbo, Hausa, Zulu, Swahili, and Yorùbá. Our choices will be drawn based on the availability of native speakers and Wikipedia editors among those languages, as well as based

[8] We extract Translation Error Rate (TER) labels Labels *automatically* using the official implementations of Snover et al. (2006).

[9] For interpretation of BLEU scores 10-20 roughly maps to *"hard to get the gist"*, 20-30 to *"the gist is clear but has significant grammatical errors"* and 30-40 to *"understandable to good translations"*.

on the updated statistics of the use of MT (see Figure 2) when we launch the project.
We describe each step in detail below.

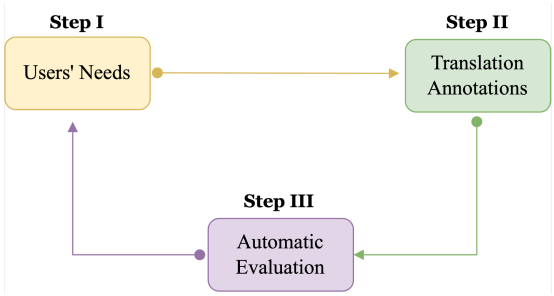

Figure 4: Outline of our three-step approach to understand and improve the use of machine translation in African Wikipedias.

Step I

Our first goal is to understand the annotation workflow and needs of Wikipedia editors who contribute to translating Wikipedia pages for the studied languages.

Method. We will connect with Wikipedia User Groups[10] , interview them, and ask them to complete surveys about their experience with the ContentTranslation tool. Finally, we would want to get insights on what are the issues they face, what are their goals, and any tensions that arise when they use the ContentTranslation tool. At a lower level, we will ask them questions about the types of machine translation errors they usually encounter or the edits they typically make. Based on their input, we will adapt the MQM typology of errors to the use case of translating Wikipedia in low-resource African languages. We will also use the take-aways from Step I in Step III to identify potential strategies to provide them with feedback on MT translation quality in a useful manner.

[10] https://meta.wikimedia.org/wiki/Wikimedia_user_groups

Participants. We will recruit participants from the Wikipedia User Groups of the language studied. To that end, we have already initiated discussions with two user groups: the Hausa Wikimedians User Group and Igbo Wikimedians User Group which have expressed interest in participating in our study. All recruited participants for Step I, will be compensated (see "Human Subject Incentives" in the budget spreadsheet).

Expected Outcome. The expected outcome of Step I is an MQM typology that reflects the types of errors that are most prominent in Wikipedias of the African languages. For instance, we might simply drop some error types that are not frequent in translation of Wikipedia, or add or specify others, such as errors in named entities, detached hallucinations, toxic language, etc. Additionally, we would also want to know at a higher level what are the kinds of things that editors care about and that frustrate them in the translation process.

Step II

Our second goal is to create datasets consisting of triplet instances as shown below:

{
..
English: [english text]
Machine Translation: [mt text]
Edited Text: [edited text]
...
}

along with human annotations of translation quality that reflect the types of errors and issues highlighted by the Wikipedia users in Stage I. These datasets will support a comprehensive quantitative analysis of Wikipedia-relevant MT errors and the development of tools to support editors. Below, we first describe the data curation process, then we outline a tentative annotation protocol with quality control

methods, and conclude with information about the participants.

Data Curation. We will sample triplet instances from the ContentTranslation dumps that Wikimedia provides.[11] The dumps are updated on an average biweekly basis, including more translations and possibly more languages. We will perform minimum filtering based on the following steps: 1. Length filtering; 2. Length ratio; 3. Untranslated texts; 4. Remove urls; 5. Remove texts with mostly numbers. Performing this filtering on the most recent dumps-20230317 results in 1,000 up to more than 15,000 translations as of March 2023 (see Figure 2). We aim to collect annotations of translation quality for approximately 1,000 translations in 3 languages.

Tentative Annotation Protocol. Given a triplet consisting of the original English sentence, the machine translation output, and the edited version, annotators will be asked to read them closely and highlight parts of the machine translation output and the edited text corresponding to errors. Additionally, once they highlight a span, they will be prompted to specify the type of error corresponding to it. The error types for annotation will be drawn from the MQM error typology that is derived from Step I. Finally, after highlighting and specifying errors in both the machine translation output and the edited texts, we would ask them to provide an aggregate judgment of translation quality by ranking them on a predefined range or a Likert quality scale.

Quality Controls. We will collect 3 annotations per triplet to allow for computing inter-annotator agreement statistics and allow for quality control. We will additionally employ attention checks and filter our annotations based on them.

Participants. We will recruit annotators from the Masakhane community. Participants will be proficient in English and native in one of the three languages studied. All recruited participants for Step II will be compensated (see "Annotators" in the budget spreadsheet).

Expected Outcome. The expected outcome of Step II is a dataset consisting of fine-grained annotated errors of the machine translated and the post-edited translations in 3 African languages.

Step III

Our goal in Step III of the project is to explore ways of increasing editors' awareness of potential MT errors, so they can use MT more reliably. The data annotation conducted in Step II will enable a wide range of approaches. The dataset itself can be used to identify prominent error types and to design training materials or editing prompts to help editors identify MT errors themselves more easily. In addition, the dataset can be used to design systems that automatically flag errors in a (machine) translated text so that Wikipedia users can potentially get actionable insights on improving the quality of a (machine) translated text.

Quality Estimation Model. To flag errors at scale, we propose to rely on automatic ways of estimating those errors that require computational approaches that compare and contrast the meaning across languages. To do so, we will build on our prior work on detecting and explaining small meaning differences (Briakou and Carpuat, 2019) across languages and build a model that, given as input an original English sentence and a (machine) translated text, will automatically extract highlights indicative of translation errors. To

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

ensure that our approach can potentially scale to other languages, we will not rely on human supervision for training purposes but instead draw on our findings from State II to generate synthetic supervision by mimicking the types of errors found frequently in Wikipedia based on our annotations.

Evaluation. We will evaluate our Quality Estimation Model against the annotations we collected at Stage II. Additionally, to understand the potential of our model to help Wikipedia editors edit machine translation errors thoroughly, we will conduct a small scale user study with the participants recruited at Stage I. In this small scale study, Wikipedia editors will be presented with English machine-translated pairs and optionally with the automatically extracted highlights. Then they will be asked to edit the machine translation output as they wish (i.e., we will not restrict them to editing only the highlighted parts, if provided). Finally, we will compare the edits made in the contrastive settings (i.e., with and without automatic highlights) to assess whether the highlights help Wikipedia editors be more thorough in catching and correcting machine translation errors. Finally, we will ask editors whether they would like to incorporate those annotations into their editing workflow.

Expected Outcome. The expected outcome of Step III is identifying prominent error types and potential training materials or editing prompts that help editors identify MT errors more easily, as well as a computational approach to detecting translation errors for the 3 African languages.

Expected output

We will share the results of our work with **Wikimedia communities** to inform them of the most critical issues concerning the use of machine translation for creating content in African languages (results of Step I & II) and explore the potential to revise their quality control methods (result of Step III). Concretely, we will:

- Present our findings at Wikimedia research conferences (e.g., Wiki Workshop, Wikidata Workshop, Wiki-M3L) and local conferences of Wikimedia communities and projects across Africa (e.g., WikiIndaba).
- Hold an active project page on MetaWiki:Research and update it regularly with findings of each of the outlined steps.

Additionally, we will share the results of our work with **academic communities** to advance our knowledge of low-resource quality estimation of machine translation. Concretely, we will:

- Release any data and computational tools produced as artifacts from this project to the public (e.g., on GitHub or other related open-access hosting platforms) and communicate any results from related surveys/interviews with Wikipedia communities via organizing office hours with them.
- Summarize our findings in open-access research papers submitted at top-tier Natural Language Processing conferences (e.g., the Annual Meeting of the Association for Computational Linguistics).

Risks

The most considerable risk associated with our proposal is that it requires bringing together Wikipedia communities and NLP researchers. To mitigate those risks, we have already contacted two Wikipedia User Groups and included Hady Elsahar who has experience working with the Wikipedia community and is a member of the NLLB team, as an advisor to our project to guide the process.

Community impact plan

Our project aims to help Wikipedia communities that serve under-resourced and under-studied languages develop best practices for using machine translation technology when they create content in their languages. This project represents a step towards bridging Wikipedia's knowledge and content gaps across languages while it brings together Wikimedia user groups and native speakers of African languages that are the direct target audience of the Wikipedia editions we will study.

Evaluation

Although we aim to produce artifacts for all outlined items in the Expected Output section, we will consider our proposed research successful if we can produce artifacts that at least have some benefit to the Wikimedia community. To ensure this is the case, we prioritized Steps I & II that will result in actionable insights about using machine translation for 3 African languages.

Budget

This information has been redacted.

References

Niklas Laxström, Pau Giner, and Santhosh Thottingal. 2015. Content translation: Computer assisted translation tool for Wikipedia articles. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation, pages 194–197, Antalya, Turkey.

team, N., Costa-jussà, M.R., Cross, J., cCelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G.M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K.R., Rowe, D., Spruit, S.L., Tran, C., Andrews, P.Y., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzm'an, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., & Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv, abs/2207.04672.

Declan Groves and Dag Schmidtke. 2009. Identification and analysis of post-editing patterns for MT. In Proceedings of Machine Translation Summit XII: Commercial MT User Program, Ottawa, Canada.

Mehmet Şahin and Sabri Gürses. 2019. Would MT kill creativity in literary retranslation? In Proceedings of the Qualities of Literary Machine Translation, pages 26–34, Dublin, Ireland. European Association for Machine Translation.

Masaru Yamada. 2019. The impact of google neural machine translation on post-editing by student translators. The Journal of Specialised Translation, pages 87–106.

António Góis and André F. T. Martins. 2019. Translator2Vec: Understanding and representing human post-editors. In Proceedings of Machine Translation Summit XVII: Research Track, pages 43–54, Dublin, Ireland. European Association for Machine Translation.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. Translation Spaces, 9(2):255–282

Ignacio Garcia. 2011. Translating by post-editing: Is it the way forward? In Machine Translation. European Association for Machine Translation.

Michael Carl, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011.

The process of post-editing: A pilot study. Copenhagen studies in language, pages 131–142.

Oliver Čulo and Jean Nitzke. 2016. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation, pages 106–114.

Michael Carl and Moritz J. Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. HERMES - Journal of Language and Communication in Business, 56:43–57.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2018. Translationese and post-editese: How comparable is comparable quality? Linguistica Antverpiensia, New Series – Themes in Translation Studies, 16.

Antonio Toral. 2019. Post-editese:an exacerbated translationese. In Proceedings of Machine Translation Summit XVII: Research Track, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Maria Stasimioti and Vilelmini Sosoni. 2020. Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In
Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation, pages 109–124, Virtual. Association for Machine Translation in the Americas.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In Proceedings of the 17th International Conference on Spoken Language Translation, pages 280–290, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In Proceedings of Machine Translation Summit XVII: Research Track, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Yingxue Fu and Mark-Jan Nederhof. 2021. Automatic classification of human translation and machine translation: A study from the perspective of lexical diversity. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age, pages 91–99, online. Association for Computational Linguistics.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification. In Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers, pages 75–81, Waikiki,
USA. Association for Machine Translation in the Americas.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In Proceedings on the Workshop on Statistical Machine Translation, pages 102–121.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E Banchs. 2007. Human Evaluation of Machine Translation Through Binary System Comparisons. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 96–103.

Carolina Scarton and Lucia Specia. 2016. A Reading Comprehension Corpus for Machine Translation Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3652–3658.

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.

Carolina Scarton and Lucia Specia. 2016. A Reading Comprehension Corpus for Machine Translation Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3652–3658.

Marianna Martindale, Kevin Duh, and Marine Carpuat. 2021. Machine Translation Believability. In Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 88–95, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondˇrej Bojar. 2021b. 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.

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 Process- ing (EMNLP), pages 2685–2702, Online. Associa- tion for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meet- ing of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1563–1580, 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.

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.

Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In International Conference on Communities and Technologies.

Marc Miquel Ribé and David Laniado. 2018. WikipediaCultureGap:Quantifying Content Imbalances Across 40 Language Editions. Frontiers in Digital Humanities 5 (2018), 12.

Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. Journal of the American society for information science and technology 62, 10 (2011), 1899–1915.

Yuncheng Jiang, Wen Bai, Xiaopei Zhang, and Jiaojiao Hu. 2017. Wikipedia-based information content and semantic similarity computation. Information Processing & Management 53, 1 (2017), 248–265.

Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. 2015. Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. In Web Science Conference.

Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. 2017. Analysing Timelines of National Histories across Wikipedia Editions: A Comparative Computational Approach. In International Conference on Web an Social Media.

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2018). Why the World Reads Wikipedia: Beyond English Speakers. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In International Conference on World Wide Web.

Pablo Beytía. 2020. The Positioning Matters: Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia. In Companion Proceedings of the Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 806–810. https://doi.org/10.1145/3366424.3383569