Quality in Human and Machine Translation: An Interdisciplinary Survey

Bettina Hiebl and **Dagmar Gromann**
University of Vienna, Austria
{bettina.hiebl, dagmar.gromann}@univie.ac.at

Abstract

Quality assurance is a central component of human and machine translation. In translation studies, translation quality focuses on human evaluation and dimensions, such as purpose, comprehensibility, target audience among many more. Within the field of machine translation, more operationalized definitions of quality lead to automated metrics relying on reference translations or quality estimation. A joint approach to defining and assessing translation quality holds the promise to be mutually beneficial. To contribute towards that objective, this systematic survey provides an interdisciplinary analysis of the concept of translation quality from both perspectives. Thereby, it seeks to inspire cross-fertilization between both fields and further development of an interdisciplinary concept of translation quality.

1 Introduction

Translation quality has been a source of debate in translation studies for decades (Koby et al., 2014), since it is considered highly subjective and dependent on how translation and quality are defined. One common denominator is the central role played by accuracy and fluency (Koby et al., 2014; Castilho et al., 2018), a view shared by the field of machine translation (Yuan and Sharoff, 2020; Koehn and Monz, 2006). An accurate semantic correspondence between source and translation as well as an adequate degree of fluency in the latter are expected. Aside from these shared notions, approaches to define, assess and measure translation quality differ substantially in the field of translation studies and machine translation. This interdisciplinary survey analyzes literature on translation quality from both perspectives.

The idea to join the theoretical basis of translation studies with the operationalized quality definitions of machine translation is not new (Čulo, 2014). However, existing surveys on the topic focus either on machine translation (Rivera-Trigueros, 2022; Han et al., 2021), post-editing (Koponen, 2016) or the perspective of translation studies (Koby et al., 2014). From a theoretical perspective, Castilho et al. (2018) present key quality theories from both fields and argue that the line between human and machine translation is increasingly blurring, especially in post-editing. The Multidimensional Quality Metric (MQM) (Lommel et al., 2014) proposes a comprehensive catalog of quality issues, which can be used to calculate a score for evaluating translations.

Inspired by the PRISMA method (Page et al., 2021) and guidelines by Kitchenham (2004), this paper presents a systematic literature review on translation quality in the field of translation studies and machine translation. Resulting publications are deduplicated and ranked by a keyword rating method that takes the number of occurrences across platforms and keywords into account. The resulting top 41 publications are presented based on the authors' fields and translation quality perspective. Thereby, the present survey contributes an overview of types of translation quality per field and interdisciplinary publications in the result set. It seeks to provide a basis for more cross-fertilization between human and machine translation quality analysis.

© 2023 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

2 Preliminaries

As a basis for the following discussion, we provide a very brief introduction to selected concepts of translation quality in translation studies and machine translation (see e.g. Castilho et al. (2018) for a more complete overview). An initial criterion of equivalence in translation studies, that is, a very close correspondence between source text and translation, was soon found too vague for a targeted quality assessment. Thus, a functionalist approach, the Skopos theory (Reiss, 1984), proposed to focus on preserving the purpose of the source text in the translation. House (2015) deems it difficult to exactly determine the purpose and proposes to divide a text into register and genre, each further subdivided, for a detailed analysis of category-based equivalence. With more attention on the recipient of the translation, criteria such as readability and comprehensibility were introduced. For instance, Göpferich (2008) proposes several dimensions of comprehensibility, that is, concision, correctness, motivation, structure, simplicity, and perceptibility.

In machine translation, the main differentiation is between automated and human quality measurement. In the former, some well-known evaluation metrics based on human reference translations are BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Other automated methods take linguistic features into account, e.g. syntactic features (Liu and Gildea, 2005) and semantic roles (Giménez and Màrquez, 2008). One major drawback is that these approaches rely on NLP techniques with limited availability for natural languages. With document-level approaches, criteria such as cohesion and coherence (Maruf et al., 2021) enter the field. To overcome the need for reference translations, machine translation quality estimation (MTQE) (Specia et al., 2018) has been proposed, especially for Neural Machine Translation (NMT). MTQE tasks include extracting features from source text and translation, selecting translations fit for post-processing, selecting the best translation between several MT systems, among others. Human quality assessment of MT focuses on categorizing segments or parts by specific criteria, e.g. comprehensibility and adequacy (Popović, 2020), however, is generally considered subjective and time-consuming and should be conducted by professional translators (Toral et al., 2018).

3 Method

The objective of this systematic literature review is to provide an overview of the state of translation quality research in the field of machine translation and translation studies and suggestions for possible joint approaches and future directions. To this end, the guidelines by Kitchenham (2004) and the PRISMA method (Page et al., 2021) served as a methodological basis. In a detailed review protocol, the main question, keywords for the search, search platforms, and inclusion/exclusion criteria were defined, which are explained below in the three main PRISMA stages, that is, identification, screening, and inclusion, illustrated in Figure 1.

3.1 Identification

To optimize the literature identification, the search was performed on three major scholarly platforms, i.e., Google Scholar, Web of Science, and Scopus. An initial list of domain-specific keywords and keyword combinations was identified, tested on domain-specific search platforms, and excluded on the basis of insufficient return of results. For translation studies journals such as Target and Translating and Interpreting Studies and for machine translation the journal of the same name and TACL as well as ACL proceedings were queried. Thereby, the following set of 12 keyword combinations was identified: “human translation” / “machine translation” AND “quality assessment” / “quality estimation” / “quality”; “translation quality”; “translation quality” AND “accuracy” / “assessment” / “comprehensibility” / “estimation” / “fluency”. To keep the amount of papers manageable by two experts and focus on recent work while including the change from statistical to neural MT, the search period was set from 2012 to 2022, assuming that this would include central concepts.

To rank the literature result set, two domain experts rated each keyword (combination) on a scale from 1, least important, to 10, most important, where the final keyword score represents the average of these two scores. The Spearman rank correlation is utilised to check the agreement of ratings between the two raters, which at 0.53 indicates a moderate agreement. The keyword score was multiplied by the times a publication was found based on this keyword (combination) on different search platforms, adding up all the occurrences across keywords and platforms. The final result set of literature was sorted by the resulting score.

3.2 Screening

Duplicates in the final result set were removed based on overlap of author(s), title, and year of publication, ranking the remaining set by the keyword-based score described in Section 3.1. Starting from the top-ranked publications, papers were screened regarding their relevance to translation quality and both authors and paper were categorized into translation studies, machine translation, or both.

3.3 Inclusion

The most central criteria for a final inclusion were the publication's relation to the topic of translation quality, quality control in form of peer reviewing, and English as a publication language. Quality control was ensured by the publication venues, where only venues with an explicit peer review process were considered. In case of preprint servers, especially arXiv, the final publication venue was double-checked manually.

4 Results

The number of records per stage of the literature survey is presented in Figure 1. During the identification stage, 12 keyword combinations were utilised to search and rate publications. The number of records returned from these was 13,762. After removal of duplicates, the keyword-ranking procedure produced results with a maximum score of 167 for the highest-ranked paper. The cutoff score for this article was determined at 77 after screening the results and determining their relevance for the research focus, taking into account the limitations caused by the number of experts of this study. In the screening process, 4 records were excluded because they were not peer-reviewed, 5 because they were superseded or results were presented elsewhere and 1 was a book review.

The 41 publications included in this review, were then divided into different thematic fields based on two dimensions: (i) background of authors in one or both fields, and (ii) field addressed in the publication. The background of authors was determined by affiliation(s), available biographic and educational descriptions, and their most common publication venues. In order to avoid confusions between the field of machine translation and approaches to machine translation, the former is referred to as computer science/computational linguistics in this section.

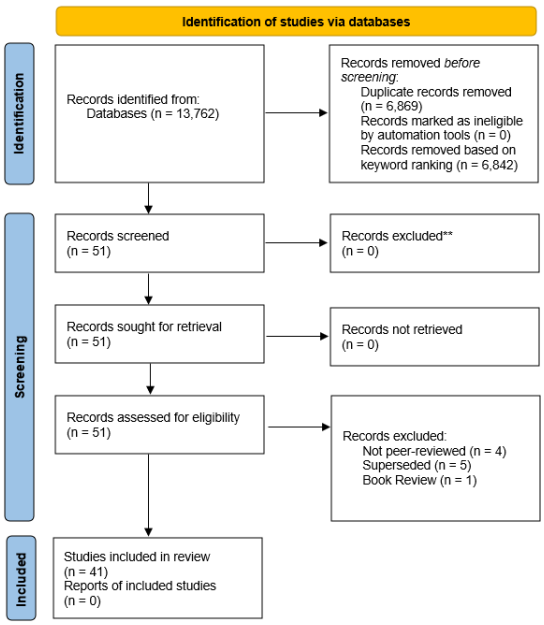

Figure 1: PRISMA 2020 Flow Diagram

4.1 Translation Studies/Languages (TS)

Out of the 41 works in the result set, 12 were assigned to the field of translation studies by the professional background of the author(s) and/or categorization of their contents. The main thematic fields in this category are (i) translation quality assessment in general; (ii) machine translation quality (assessment); as well as (iii) human translation quality, post-editing and revision.

TS - Translation Quality Assessment: The common topic in Doherty (2017), Krüger (2022) and Vela-Valido (2021) is translation quality assessment (TQA) and its performance by humans and machines from a theoretical point of view. Doherty (2017) discusses issues in TQA from the perspectives of TS, MT and the translation industry. The main identified issues are explicit definitions of quality, adhering to established tests for validity and reliability, greater awareness of human factors in evaluating quality, and improved transparency in shared translations. For testing validity and reliability, other fields should be taken into account, such as psychometrics.

Krüger (2022) focuses on providing input from the field of translation studies to methodologies for MT quality evaluation, as a means of contributing to the debate on quality of NMT compared to quality of human translations. Suggestions are that human reference translations should be approved, contextual factors should become more important when evaluating, translation er-

rors should be weighted by their severity, and MT should be integrated in settings where high quality of translations is of utmost importance for measuring the added value of professional translators.

Vela-Valido (2021) focuses on AI-based translation quality management in the translation industry, describing the steps performed before, during and after production. The main focus are AI-based tools in quality assessment and estimation as well as quality assessment workflows, presenting the support AI-based tools can give to humans and the need of humans to still take the final decisions.

These publications show the growing importance of MT in TS and the willingness of TS researchers to contribute their experience to MT quality definitions and approaches. However, a need to involve humans in the translation process is emphasized.

TS - Machine Translation Quality (Assessment): Different ways to perform machine translation quality assessment are presented by Chatzikoumi (2020) in a review of automated, semi-automated and human metrics for MT evaluation. Human evaluation categories are subdivided as to whether they present directly expressed judgements (DEJ) or not, a somewhat debatable categorization. While adequacy and fluency annotations present DEJ, error classification and post-editing are considered to merely state that the translation is not perfect without directly judging its quality.

The remaining works in this subsection are empirical studies on MT quality assessment, pointing to mistranslation as the most common error type across text types. Moorkens (2018) describes an evaluation of SMT as opposed to NMT by two cohorts of students on the basis of adequacy, post-editing productivity, and error taxonomy. With little surprise, a high preference for NMT in all three categories could be observed. A manual error annotation of an NMT-translated detective novel showed that the most frequent errors in this literary text were mistranslation, coherence, style and register (Fonteyne et al., 2020). Candel-Mora (2022) argues that different quality rating scales should be introduced for each type of text. In their study relying on the TAUS Dynamic Quality Framework (DQF), mistranslations but also punctuation errors were most common.

TS - Human Translation Quality, Post-Editing and Revision: While the focus of this subsection is on human translation, a growing influence of technological advances that impacts the concept of translation quality can be observed in TS. In contrast to editing or post-editing, revision involves an evaluation against the source text. Mellinger (2018) argues for re-thinking the concept of translation quality in the digital age and calls for a process-oriented perspective on translation quality, incorporating editing and revision tasks in TQA. The translation and revision workflow has changed with technological advances, such as Computer-Assisted Translation (CAT) and MT, allowing for asynchronous workload distribution and working on stored/draft translations. This view impacts the definition of translation quality as not merely determined by textual and linguistic features, but reliant on quality control to ensure compliance with (client) specifications, the purpose, and target audience. With the emergence of crowdsourcing and collaborative approaches, translation has evolved from a static, high-value to a dynamic, fit-for-purpose product (Jiménez-Crespo, 2017). Thus, different grades of quality can now be found in TS literature, e.g. low, medium, high or by amount of post editing required. This shifts the final responsibility for quality to the customers "who select the level of quality through a wide range of considerations, such as the available budget, permanence of the translation, potential risks involved, receiving audience, etc." (Jiménez-Crespo, 2017, 489)

Empirical studies in the result set include the utilization of automated metrics, e.g. BLEU or METEOR, to evaluate human translation (Karami et al., 2020). The basic idea was to test whether a higher number of translations increases the reliability of the score. This assumption could partially be confirmed, however, the increase in reliability depended on the specific reference translation that was added.

In a similar fashion, Ortiz-Boix and Matamala (2017) compare post-edited machine translations to human translations from parts of wildlife documentaries. 12 students translated and post-edited two excerpts, which were then assessed by 6 professional translators by means of grading, assessment with MQM, and questionnaires. The results confirmed the authors' assumption that there is no significant quality difference between translated and post-edited texts. Finally, Leiva Rojos (2018) assesses phraseological quality in com-

parison to the overall quality of texts in 14 original and translated museum texts based on the assumption that the level of phraseological quality of a text is directly related to its overall quality. While generally observed to be true, in most cases the results of the phraseological assessment are better than the overall results.

4.2 Computer Science/Computational Linguistics (CL)

From the result set, 21 publications were classified as belonging to computer science/computational linguistics. The main thematic fields are (i) translation quality, its assessment and crowdsourcing; (ii) machine translation and its quality assessment; (iii) machine translation quality estimation; and (iv) human translation quality estimation.

CL - Translation Quality Assessment: As in the field of translation studies, there is only a small number of works on TQA, describing or proposing quality assessment models. Whereas in TS the main suggestions are involving humans and machines as well as taking context into account, the publications in this section mostly present ideas for making translation quality easier to measure.

In a systematic survey, Han et al. (2021) present an extensive overview of human and automated methods of MT quality assessment, from basic criteria, such as intelligibility, to neural networks for TQA. They suggest that future TQA models should not only involve n-gram word surface matching but also deeper linguistic features, such as syntactic dependencies and semantic roles. Furthermore, they predict that MTQE will continue to attract attention due to its multiplicity of tasks. Lommel et al. (2013) present the much-used MQM, a flexible method for human TQA, which can be applied to human as well as machine translation. These metrics represent a system of core issue types, e.g. terminology, style, locale conventions, to which different subcategories can be added based on the task at hand. The MQM and its core issue types keep on being updated by a corresponding World Wide Web Consortium (W3C) community group[1].

CL - Machine Translation Quality (Assessment): Approaches in the result set on MTQA range from cross-sentence evaluations to crowdsourcing approaches. Popel et al. (2020) propose and evaluate a Transformer-based model against human translations and stress the importance of context-aware evaluation of translation quality, since cross-sentence contexts represented a major source for errors. On sentence-level, the model could even pass a Translation Turing Test, in which human participants failed to significantly differentiate human from machine translations. Licht et al. (2022) propose a new metric based on semantic text similarity called XSTS with five levels from full semantic equivalence to none that emphasizes adequacy rather than fluency. The metric is tested with human evaluators in 14 language pairs.

The result set further contained several use cases, such as in patent translation (Rossi and Wiggins, 2013) where automated metrics are compared to human evaluation of MT quality by terminology, missing or added information, and word order via an online interface. Graham et al. (2017) assess a new methodology for crowdsourcing human MTQA. They compare the assessments by the crowd with the WMT-12 evaluation and conclude that evaluation of MT systems by the crowd alone is possible.

Burchardt et al. (2021) argue that different purposes and user groups require different TQA methods and propose three and accompanying use cases: (i) a semi-automated method based on regular expressions, (ii) applying MQM, and (iii) a task-based user evaluation. Fomicheva and Specia (2016) assume that performing MTQA with reference translations may negatively bias human annotators. Using an online interface, they compared agreement between the annotators using the same human reference translation and those using different ones, showing that monolingual evaluation is affected by the reference provided. In a study on MT in foreign language education, He (2021) concludes that MT provides a good reference for learners, even though culture-specific aspects, such as tone, might not be represented equivalent to human translations. Way (2018) discusses quality expectations of MT. He views MT as enhancing the productivity of human translators and argues that with regards to the use cases of MT as well as their "shelf-life", the expectations of certain standards regarding quality need to be revised, while at the same time pointing out that humans are still crucial also with regards to MT.

CL - Machine Translation Quality Estimation: There are general works on MTQE and its future

[1] https://www.w3.org/community/mqmcg/

perspectives, such as Specia and Shah (2018), who review various fields in which QE at sentence-level was successful. They then discuss QE at word- and document-level as well as future perspectives. In the same direction, but with a more specific orientation, González-Rubio et al. (2013) present different dimensionality reduction methods and compare them against different reduction methods used in QE literature and they study how the performance of different learning models is influenced by these methods. Graham (2015) addresses issues which can arise during comparison of quality estimation prediction score distributions and gold label distributions. She proposes using a unit-free Pearson correlation and reruns parts of evaluations of WMT-13 and WMT-14 to demonstrate its use.

The remaining four publications in this category propose new MTQE methods, such as building on pretrained language models (Huang et al., 2020), RNN-based sentence-level methods (Ren, 2022), and reinforcement learning (Li et al., 2021). Chen et al. (2021) present a document-level QE model based on Centering Theory in order to tackle the problem of missing context information of previous sentence-level QE models.

CL - Human Translation Quality Estimation: A relatively new topic in the field of CL is Human Translation Quality Estimation (HTQE). Yuan et al. (2016; 2017) propose an evaluation framework based on feature sets extracted from and utilised to evaluate human translations. The focus is on predicting adequacy and fluency. Yuan and Sharoff (2018) investigate a slightly different topic, namely the influence of bilingual multiword units (BMWUs) on trainee translation quality. They assess the contribution of BMWUs to translation quality and show that normalised BMWU ratios can be useful for estimating human translation quality. Finally, in a comparison of neural-based sentence-level HTQE and prior feature-based methods (Yuan and Sharoff, 2020), the former outperform the latter.

4.3 Translation Studies/Languages & Computer Science/Computational Linguistics

In the result set, 8 publications represented joint work by TS and CL scholars. The thematic fields in this subsection are (i) translation quality assessment; (ii) machine translation quality (assessment) and post-editing; and (iii) human translation quality and post-editing.

TS & CL - Translation Quality Assessment: In the result set, only one publication was related to TQA explicitly. Castilho et al. (2018) reflect on TQA regarding both assessment of human as well as of machine translation from different perspectives, namely from TS, MT and the translation industry. They identify the following key issues regarding translation quality assessment: lack of standardisation in TQA usage, inconsistency in TQA, the differing relationship between human and automatic measures, the social quality and risk as well as education and training in TQA.

TS & CL – Machine Translation Quality (Assessment) & Post-Editing: Gaspari et al. (2015) conducted a survey of machine translation competences with 438 respondents, which included freelance translators, language service providers, translation trainers and academics. It shows that the importance of machine translation is growing and will be more and more part of workflows in the future, having an influence on the human translation process, e.g. the need of post-editing, and on translation training, e.g. the need for increased technical competencies.

Assessment of machine translation quality using the MQM highlights its usability in and adaptability for different contexts. Burchardt et al. (2016) focus on MT quality in the context of Audio-Visual Translation (AVT), trying to bridge the gap between the field of MT developers mainly focusing on high-quality MT for text production and the field of the tech-savvy AVT community. They propose to extend the MQM by AVT specific types, i.e., contextual for mistranslations in situative contexts and timing for translations presented out of synch with other modalities. Carl and Toledo Báez (2019) conducted an experiment in which translators annotate Spanish and simplified Chinese MT output using an MQM-derived error taxonomy. They investigated the effect of MT errors on post-editing efforts and found that accuracy errors influence production and reading duration. Additionally, they found that segments with MT accuracy issues in one language combination are likely to be difficult to translate to other languages, which they did not find to apply for fluency errors.

Analysis of different error types is also part of the studies carried out by Daems et al. (2017) and Vardaro et al. (2019). However, they both

also focus on the post-editing process and involve keystroke logging and eye-tracking. More specifically, in order to identify the MT error types with most impact on the post-editing effort, Daems et al. (2017) conducted a study, in which the post-editing process of student and professional translators was recorded and analyzed from the perspectives of acceptability and adequacy. They find that different types of errors affect different post-editing effort indicators and that coherence, meaning shifts and structural issues are good indicators of post-editing effort. Vardaro et al. (2019) conducted a study with translation experts from the German department of the European Commission's Directorate-General for Translation (DGT), analyzing how they identify and correct different error categories in NMT texts and the post-edited versions of these, which showed that the most common error types to correct are lexical errors. Differences of eye movements across error categories were not significant.

TS & CL – Human Translation Quality (Assessment) & Post-Editing: Munkova et al. (2021) and Jia et al. (2019) both compared from-scratch translation with post-editing of machine translated texts, both conducting analyses on the product and process level. Munkova et al. (2021) assess the influence of the quality of MT output on the translator's performance in translating journalistic texts. Product analysis was done as MTQA using the TAUS DQF and process analysis by measuring typing time during post-editing. Findings show that the translator's performance is influenced by MT quality and that post-editing compared to human translation is more effective. Jia et al. (2019) also compared from-scratch translation with post-editing of NMT of domain-specific and general language texts. The translation process and product data from 30 translation students were analyzed based on keystroke logging and screen recording, among other dimensions. The study's results regarding quality are that post-editing was significantly faster than translating from scratch with less cognitive effort, and that fluency and accuracy of post-edited texts was equivalent to those of translated texts.

5 Discussion

This systematic survey showed that translation studies and machine translation have more in common in reference to translation quality than accuracy and fluency. A growing influence of technological advances has shifted the translation workflow and conceptualizations of translation quality in TS. Alongside automated metrics and post-editing, the fit-for-purpose idea of translation quality has entered the field, shifting the burden of defining quality from translators to clients. On the other hand, quality criteria such as (cross-sentence) context, comprehensibility, and readability have entered the field of MT. Furthermore, the substantial number of joint publications by authors from both backgrounds indicates a convergence of both fields.

The results show that in both disciplines, new technological developments are of great interest. TS scholars become increasingly aware that MT can be useful in TS. In contrast, MT scholars realize that comparing outputs to a reference translation or without taking the context into account has considerable drawbacks. Publications in TS contain more theoretical contributions, ideas on how MT can be integrated in translators' workflows, studies on machine translation quality assessment as well as post-editing and revision. The fairly new concept of (machine or human) translation quality estimation seems to have not yet been considered in TS. In the field of MT, (machine or human) translation quality estimation is the main topic in more than half of the publications. Additionally, a continuously strong focus on automated metrics and technological advances can be observed. In a nutshell, TS can still contribute a strong theoretical basis, quality criteria, and especially definitions of translation quality to MT, while MT can facilitate more measurable and (semi-)automated approaches to translation quality to TS.

Several limitations of the present survey should be acknowledged. First, its scope was limited to 41 included results, which, given the scope of the topic, raises no claims as to completeness. In fact, several important publications, e.g. Toral et al. (2018) and Läubli et al. (2018), were not in the result set. Snowballing or considering citation scores should be future amendments of the method to counteract this issue. Secondly, categorizing publications by the authors' scientific field is a somewhat unusual and time-intensive approach. We opted for this approach, since we were particularly interested in the number of publications jointly authored by researchers from both fields and the view on quality concepts by each field. An additional

subdivision of publications by the main translation quality concept seeks to provide a transparent and comprehensible categorization method.

6 Conclusion and Future Research

This comprehensive survey on translation quality in the field of translation studies and machine translation showed that the main ideas in both fields still differ slightly, with translation studies still focusing more on theoretical and less measurable concepts and computational linguistics more on conducting studies and developing metrics. While, on the whole, quality concepts in the two fields are converging, the main challenge in the future will still be to design quality assessment metrics including less easily measurable criteria, such as context and purpose. A systematic catalog of translation quality definitions, criteria, and evaluations of their measurability would be interesting in this regard. Furthermore, we suggest to include the role of the translation industry and its viewpoint on translation quality in future reseearch.

References

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

Burchardt, Aljoscha, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja Popović. 2016. Machine translation quality in an audiovisual context. *Target*, 28:206–221.

Burchardt, Aljoscha, Arle Lommel, and Vivien Macketanz. 2021. A new deal for translation quality. *Universal Access in the Information Society*, 20:701–715.

Carl, Michael and M Cristina Toledo Báez. 2019. Machine translation errors and the translation process: a study across different languages. *Journal of Specialised Translation*, 31:107–132.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 9–38. Springer, Cham, Switzerland.

Chatzikoumi, Eirini. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Chen, Yidong, Enjun Zhong, Yiqi Tong, Yanru Qiu, and Xiaodong Shi. 2021. A document-level machine translation quality estimation model based on centering theory. In *Machine Translation: 17th China Conference, CCMT 2021, Xining, China, October 8–10, 2021, Revised Selected Papers 17*, pages 1–15, Xining, China.

Čulo, Oliver. 2014. Approaching machine translation from translation studies–a perspective on commonalities, potentials, differences. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 199–206, Dubrovnik, Croatia.

Daems, Joke, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8:1282.

Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In editor, The, editor, *Human Issues in Translation Technology*, pages 149–166. Routledge.

Fomicheva, Marina and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany.

Fonteyne, Margot, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an nmt-translated detective novel on document level. In *12th International Conference on Language Resources and Evaluation (LREC)*, pages 3783–3791, Marseille, France.

Gaspari, Federico, Hala Almaghout, and Stephen Doherty. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*, 23:333–358.

Giménez, Jesús and Lluís Màrquez. 2008. A smorgasbord of features for automatic mt evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198.

González-Rubio, Jesús, J Ramón Navarro-Cerdán, and Francisco Casacuberta. 2013. Dimensionality reduction methods for machine translation quality estimation. *Machine Translation*, 27:281–301.

Göpferich, Susanne. 2008. *Textproduktion im Zeitalter der Globalisierung: Entwicklung einer Didaktik des Wissenstransfers*. Studien zur Translation ; 15. Stauffenburg, Tübingen, 3. aufl. edition.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3–30.

Graham, Yvette. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China.

Han, Lifeng, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online.

He, Xinyu. 2021. Evaluation of machine translation quality based on neural network and its application on foreign language education. In *AIAM2021: 3rd International Conference on Artificial Intelligence and Advanced Manufacture*, pages 1395–1399, Manchester, United Kingdom.

House, Juliane. 2015. *Translation quality assessment: Past and present*. Routledge.

Huang, Hui, Hui Di, Jin'an Xu, Kazushige Ouchi, and Yufeng Chen. 2020. Ensemble distilling pretrained language models for machine translation quality estimation. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 231–243, Zhengzhou, China.

Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, 31(1):60–86.

Jiménez-Crespo, Miguel A. 2017. How much would you like to pay? reframing and expanding the notion of translation quality through crowdsourcing and volunteer approaches. *Perspectives*, 25(3):478–491.

Karami, Somayyeh, Dariush Nejadansari, and Akbar Hesabi. 2020. Reliability of human translations' scores using automated translation quality evaluation understudy metrics. *Journal of Foreign Language Research*, 10(3):618–629.

Kitchenham, Barbara. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Koby, Geoffrey S, Paul Fields, Daryl R Hague, Arle Lommel, and Alan Melby. 2014. Defining translation quality. *Tradumàtica*, 12:0413–420.

Koehn, Philipp 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, New York City, June. Association for Computational Linguistics.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2).

Krüger, Ralph. 2022. Some translation studies informed suggestions for further balancing methodologies for machine translation quality evaluation. *Translation Spaces*, 11(2):213–233.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Leiva Rojo, Jorge. 2018. Phraseology as indicator for translation quality assessment of museum texts: A corpus-based analysis. *Cogent Arts & Humanities*, 5(1):1442116.

Li, Feiyu, Yahui Zhao, Feiyang Yang, and Rongyi Cui. 2021. Incorporating translation quality estimation into chinese-korean neural machine translation. In *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, pages 45–57, Hohhot, China.

Licht, Daniel, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA, September.

Liu, Ding and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2), mar.

Mellinger, Christopher D. 2018. Re-thinking translation quality: Revision in the digital age. *Target*, 30(2):310–331.

Moorkens, Joss. 2018. What to expect from neural machine translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Mora, Miguel Ángel Candel. 2022. Fine-tuning machine translation quality-rating scales for new digital genres: The case of user-generated content. *ELUA: Estudios de Lingüística. Universidad de Alicante*, 38:117–136.

Munkova, Dasa, Michal Munk, Katarina Welnitzova, and Johanna Jakabovicova. 2021. Product and process analysis of machine translation into the inflectional language. *SAGE Open*, 11(4):21582440211054501.

Ortiz-Boix, Carla and Anna Matamala. 2017. Assessing the quality of post-edited wildlife documentaries. *Perspectives*, 25(4):571–593.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of Surgery*, 88:105906.

Papineni, Kishore, 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.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11:4381.

Popović, Maja. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069.

Reiss, Katharina und Vermeer, Hans J. 1984. *Grundlegung einer allgemeinen Translationstheorie*, volume 147. Max Niemeyer Verlag, Tübingen.

Ren, Beibei. 2022. Machine automatic translation quality evaluation model based on recurrent neural network algorithm. In *Cyber Security Intelligence and Analytics: The 4th International Conference on Cyber Security Intelligence and Analytics (CSIA 2022), Volume 1*, pages 1019–1026.

Rivera-Trigueros, Irene. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 56(2):593–619.

Rossi, Laura and Dion Wiggins. 2013. Applicability and application of machine translation quality metrics in the patent field. *World Patent Information*, 35:115–125.

Specia, Lucia and Kashif Shah. 2018. machine translation quality estimation: applications and future perspectives. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 201–235. Springer, Cham, Switzerland.

Specia, Lucia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality estimation for machine translation*. Synthesis Lectures on Human Language Technologies. Springer Cham.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics*, 6:41.

Vela-Valido, Jennifer. 2021. Translation quality management in the ai age. new technologies to perform translation quality management operations. *Revista Tradumàtica*.

Way, Andy. 2018. Quality expectations of machine translation. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 159–178. Springer, Cham, Switzerland.

Yuan, Yu and Serge Sharoff. 2018. Investigating the influence of bilingual MWU on trainee translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Yuan, Yu and Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1858–1865, Marseille, France.

Yuan, Yu, Serge Sharoff, and Bogdan Babych. 2016. MoBiL: A hybrid feature set for automatic human translation quality assessment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3663–3670, Portorož, Slovenia.

Yuan, Yu, Bogdan Babych, and Serge Sharoff. 2017. Reference-free system for automated human translation quality estimation. In *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5, Lisbon, Portugal.