Evaluation Briefs: Drawing on Translation Studies for Human Evaluation of MT

Ting Liu 劉婷 [a, b] tliu109@uOttawa.ca
Chi-kiu Lo 羅致翹 [b] ChiKiu.Lo@nrc-cnrc.gc.ca
Elizabeth Marshman [a] Elizabeth.Marshman@uOttawa.ca
Rebecca Knowles [b] Rebecca.Knowles@nrc-cnrc.gc.ca
[a]School of Translation and Interpretation, University of Ottawa
[b]National Research Council Canada

Abstract

In this position paper, we examine ways in which researchers in machine translation and translation studies have approached the problem of evaluating the output of machine translation systems and, more broadly, the questions of what it means to define translation quality. We explore their similarities and differences, highlighting the role that the purpose and context of translation plays in translation studies approaches. We argue that evaluation of machine translation (e.g., in shared tasks) would benefit from additional insights from translation studies, and we suggest the introduction of an "evaluation brief" (analogous to the "translation brief") which could help set out useful context for annotators tasked with evaluating machine translation.

1 Introduction

The evaluation of translation quality remains a challenge in the fields of machine translation (MT) and translation studies (TS). Evaluation methods relying on human judgement have changed and developed alongside advances in machine translation technology. In MT, the longstanding goal of these evaluation approaches has been to provide a standardized and possibly even "objective" evaluation process. In this work, we will draw on complementary perspectives from MT and TS.[1] We will show that there are similarities and connections between the fields' views on evaluation, as well as areas where insights from TS could be used to inform and improve approaches to human evaluation of MT.

Controversies resulting from claims that MT quality has reached "parity" with humans (Hassan et al., 2018) as well as problems with human evaluation campaigns at the Workshop on Machine Translation (WMT) have led to MT researchers shining a spotlight on evaluation protocols and their challenges (Toral et al., 2018; Läubli et al., 2018; Knowles, 2021; Castilho and Knowles, 2024, i.a.)—this has also piqued the interest of researchers in TS (e.g., Krüger, 2022).

While many previous works in the MT literature on problems in human MT evaluation have examined questions like how to set up the evaluation process, how to incorporate context, and how to standardize annotator scores, it is rarer for them to focus specifically on the definition at the core of this process: what *is* translation quality?

In this paper, we investigate how perspectives on that question differ between researchers in the fields of MT and TS. We begin with MT researchers' perspectives on current methodologies in human evaluation, focusing on what attributes of quality these evaluations prioritize. We then present a view of theoretical and practical dimensions of translation quality assessment (TQA) within TS in academia and industry. We explore TQA

[1]This paper stems from ongoing research conversations about translation quality and its evaluation between researchers in MT and TS, with this work primarily aimed at an audience familiar with the MT literature.

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

models in academia such as House's TQA model, Williams' argumentation-centered approach, Colina's approach, and industry-driven approaches like Multidimensional Quality Metrics (MQM). Following a comparative analysis of these methodologies, we highlight one particular facet of quality evaluation that is present in the TQA models but frequently absent from the MT research approaches: an acknowledgement of the purpose of a given translation and the context in which it is produced and expected to be used.

We take the position that MT research could benefit from incorporating these TS perspectives, and we conclude this work by considering how this could be done in practice. Translators are sometimes given a "translation brief" describing the goals of the translation, the intended audience of the translation, and other important contextual information; we propose an analogous "evaluation brief" to serve a similar role in human evaluation of translation. We then discuss how this additional context could be implemented in practice in MT evaluation, including the importance of being aware of how the annotator population differs or is similar to the expected end users of the MT system being evaluated (e.g., in terms of subject area knowledge, dialect, context, etc.), as problems in evaluation could arise due to a mismatch.

2 MT Researcher Perspectives

Since early MT experiments, human evaluation has been positioned as the ideal form of evaluation, with automatic metrics seen as a necessary stand-in. Even BLEU ("bilingual evaluation understudy"), in its name, considers automatic metrics an "understudy" to human evaluators (Papineni et al., 2002). How those human evaluations should be produced has been an open question, with a rotating cast of proposed methodologies and definitions of quality. The methodologies, procedures, and interfaces used for collecting evaluations include sliders with a continuous scale, discrete scales, ranking, annotations of the text, among others. These are often discussed in conjunction with the aspects of quality being considered (e.g., discrete scales for adequacy and fluency[2]), with terminology surrounding methodologies and interfaces blurring the line between the interface itself and the questions that annotators are being asked about quality. But in practice these are orthogonal concerns; various interfaces could be paired with any number of questions about different aspects of quality. This paper categorizes human evaluation methodologies used in MT evaluation into three broad (and sometimes overlapping) groups: manual scoring, semi-automatic (or, from another perspective, this could be viewed as semi-manual), and task-based. We also touch on how these groups of evaluation methodologies typically address questions of quality, and which aspects of quality are regularly considered in MT human evaluation.

We define "manual scoring" evaluations as evaluations in which an annotator directly provides a score or ranking to one or more systems. Early evaluations at shared tasks asked annotators to judge adequacy and fluency on 5- or 7-point scales (LDC, 2002; Koehn and Monz, 2006; Callison-Burch et al., 2007, i.a.) and this approach is frequently revisited in other proposed variants, such as the rating of semantic faithfulness to source text in Licht et al. (2022). In later system ranking tasks (Vilar et al., 2007; Callison-Burch et al., 2007, 2008, 2009, i.a.), annotators were asked "to rank the translations from best to worst (ties are allowed)" (Bojar et al., 2016), without specific guidance about what aspects would make one translation better or worse than another. Recent WMT annotation campaigns have used direct assessment (DA; Graham et al., 2013a, 2014, 2016), where annotators provide a score from 0-100 on a sliding scale. These began with asking for adequacy-oriented human judgements, but were gradually replaced with questions including both meaning and grammar (Kocmi et al., 2023).[3] We include some of the exact questions for human annotators in WMT evaluation campaigns for refer-

[2]Adequacy is defined in terms of the amount of meaning carried over from the source sentence to the translation, while fluency focuses on whether the target language text is grammatical or natural-sounding regardless of semantic content. At times these have been referred to by other terms as well. In earlier stages of MT development, adequacy and fluency were found to be highly correlated, and evaluations shifted to focus only on adequacy (Callison-Burch et al., 2007). More recent research argues that "accuracy and fluency are positively correlated at the level of the corpus but trade off at the level of individual source segments" (Lim et al., 2024).

[3]Notably, the most recent scale design in Kocmi et al. (2022, 2023, 2024) violates best practices in measurement theory and questionnaire design by incorporating these two distinct aspects into a single rating scale (Fowler, 2013, p. 81-82).

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

ence in Appendix A. In addition to their similarity in terms of annotators directly providing some sort of score for a translation, these approaches have all been challenged at various points due to issues relating to inter- and intra-annotator consistency.

We define our second category of "semi-automatic" evaluations as ones in which an annotator provides some sort of annotation (or transformation like post-editing) to the text, and then a score for the MT is computed (automatically) based on the annotations. This includes approaches such as Human-targeted Translation Edit Rate (HTER; Snover et al., 2006), where the edit rate (an automatic measure of the number of changes/difference) of MT output is measured against the (human) post-edited version, rather than a generic reference, with the expectation that that MT output with higher translation quality requires fewer edits in order to produce an acceptable post-edited translation. Other approaches involve the annotation of errors using an error typology (e.g., Multilingual Quality Metrics, MQM; Burchardt, 2013) followed by computing a score based on the number and severity of errors. Similar approaches, such as HMEANT (Lo and Wu, 2011) and HUME (Birch et al., 2016), involve annotating the shallow semantic structures/units and translation correctness of each semantic unit in the MT output, followed by aggregating these correctness annotations into a score for the translation quality at sentence level. Semi-automatic approaches tend to have more well-defined instructions for annotators. However, even these may have ambiguities in the interpretations of the evaluation task. Al Sharou and Specia (2022) described challenges in consistency of annotating critical errors using an error typology, noting the importance of annotator training while also acknowledging that ambiguities and confusion may nevertheless persist. Lo and Wu (2014) and Birch et al. (2016) both showed that there are compounding disagreements between annotators at the end of the evaluation task using HMEANT and HUME.

The third group consists of task-based evaluations. In this type of extrinsic evaluation, annotators are asked to use MT output to perform a task, e.g., template filling in Laoudi et al. (2006), question answering in Jones et al. (2007), semantic parsing in Moghe et al. (2023), etc.; the performance on the downstream task is scored. These scores are interpreted as the usefulness of the MT output for the downstream task and used to form a score or ranking of the translation quality of the underlying MT system. These task-based approaches typically do not ask annotators to directly judge aspects of MT quality. Instead, they emphasize the utility/usefulness aspect of the translation and implicitly ask "Is the quality of the MT good enough for the annotator to perform the requested task?", "Does one MT system better enable annotators to complete the task than another MT system?", or similar questions.

As far as we can tell, the specific form of the questions posed/directions given to annotators are (with a few exceptions) rarely studied by MT researchers in order to ensure their validity or reliability. In general, despite its goals of producing "objective" scores, human evaluation in MT research has tended to focus either on high-level and potentially undefined or underspecified aspects like generic "quality", divided quality into adequacy and fluency, defined MT performance based on downstream task performance, or used error typologies. Graham et al. (2012) raised the question of whether identifying the "components" of quality that annotators used in their decisions could help to improve the reliability and validity of future evaluations.

3 Translation Studies (TS) Perspectives

We now explore perspectives on translation quality and translation quality assessment (TQA)[4] from TS academia and industry. Academic research in TS often explores theoretical frameworks, pedagogical implications, and methodological innovations (Jakobsen, 2017; Carl, 2021, i.a.), and the translation industry tends to focus on operational efficiency, quality assurance, and client satisfaction, frequently employing quantitative measures and standardized processes to ensure consistency and reliability in translation outputs (Williams, 2004; Pym, 2019; Bowker, 2019, i.a.). Drugan (2013) and Castilho et al. (2018) note the challenges of TQA in practical settings, with an eye toward real-world applicability, often within the constraints of tight timelines and specific client needs. Although there is a significant body of research within TS

[4]TQA is a branch of translation criticism (Holmes, 1988, p. 78), concerning "how to tell whether a translation is good or bad" (House, 2015).

that is process-oriented (Dimitrova, 2010; Saldanha and O'Brien, 2014, i.a.), in this paper we focus on product-oriented aspects of translation, as these align more with the areas of MT evaluation research that we also examine.

3.1 Definitions of Translation Quality

In TS, quality has been conceptualized through diverse contexts and perspectives and has been the subject of many debates. Koby et al. (2014) characterized translation quality in terms of two major senses: narrow and broad. In the narrow sense, translation is text-centric, requiring a full transfer of the source text's message to the target language with correct grammar and cultural appropriateness. Early understandings focused on linguistic fidelity and equivalence, such as textual equivalence (Catford, 1965). This line of work emphasized accurate replication of meaning and structure from source text to target text to ensure the translation closely mirrored the original. Building on this, House (1997) conceives of translation as a double-constrained text, bound to both the source text and the target audience's communicative conditions. Translating involves substituting one language's text with another language's equivalent that serves the same purpose. This functional equivalence is significantly affected by two empirically established categories of translation: overt and covert translation (House, 1997). In an overt translation, the original text's cultural context and linguistic features are preserved so the target audience can experience the original cultural nuances (House, 2001). For example, translating ancient Greek poetry while maintaining references to Greek mythology and cultural practices is a type of overt translation. However, covert translation seeks to create an equivalent text that functions seamlessly in the target culture as if it were an original. Translations with a "cultural filter" adapt the content to the target audience's expectations and cultural norms, creating a text that appears to have been written in the target language originally. Translations of marketing and advertising materials typically fall into this category. This often involves adapting idiomatic expressions, cultural references and humour to align with local tastes and expectations. Thus, translating overtly or covertly depends on the text's nature, the purpose of the translation, and the intended audience.

The broad sense of quality described by Koby et al. (2014) encompasses the narrow sense but adds compliance with negotiated specifications and consideration of end-user needs, ensuring translations meet measurable standards and fit their purpose. This broader perspective first aligns with functionalist approaches which define translation quality as whether a translated text fulfills its intended purpose for the target audience in the given circumstances, ensuring linguistic accuracy and appropriateness in context. Vermeer (1978, 2021) introduced Skopos theory, which argued that linguistic solutions, such as lexical choices and syntactical adjustments, cannot address all translation issues, including maintaining the original text's intent and adapting to cultural differences. Skopos theory considers translation as a purposeful action based on the source text, where the translator must consider the intent of the original text and adapt it to the target culture.

Building on Skopos theory and the process by which translations are commissioned (Vermeer, 1978, 2021), Nord (1997a, p. 46-48) introduced the translation brief or "Übersetzungsauftrag". A translation brief typically includes "the target-text addressee(s), the prospective time and place of text reception, the medium over which the text will be transmitted, and the motive for the production or reception of the text" (Nord, 1997a). ISO (2015) listed 22 key and supplementary elements in a translation brief, which included information about the source content, source and target languages, linguistic specifications (e.g., language variants), audience, purpose, style guide, locale conventions, reference materials, etc., on top of some project management specifications. Similarly, Esselink (2003) introduced a translation kit (or localization kit) as a package of files that includes all necessary information to meet the client's quality standards. More recently, Calvo (2018) used the term "specifications" to reflect the complexity of modern translation projects. Here, "Skopos", "brief", and "specifications" determine the communicative function and quality of the translation.

Chesterman and Wagner (2002, p. 80-84) added a view on quality from an industrial context. Here, translation quality is viewed from different perspectives: as a product judged by end quality, as a process dependent on correct execution, as a service measured by customer satisfaction, and as a

copy to be assessed by accuracy and faithfulness to the original text. The view of translation quality as customer/end-user satisfaction is also discussed by Pym (2019, p. 437-452).

Another view on the definition of translation quality stems from the management quality framework in Garvin (1984), which encompasses five perspectives: transcendent, product-based, user-based, manufacturing-based, and value-based. Fields et al. (2014) introduced this framework to the translation industry to increase overall translation effectiveness and satisfaction by balancing stakeholder expectations and addressing diverse quality dimensions.

To conclude, the understanding of quality in TS has evolved significantly from early emphases on linguistic fidelity to a more inclusive understanding that considers functionalist, industrial, and management perspectives. This inclusive view acknowledges that translation quality is multifaceted, considering both linguistic accuracy and the fulfillment of the translation's intended purpose for its end users. This does not cover the full range of definitions of translation quality; the ones we selected for discussion here are especially pertinent.

3.2 Concepts of Translation Quality Assessment (TQA)

Bowker (2000, p. 183) described TQA as "the most problematic area of translation," citing descriptions like "a great stumbling block" (Bassnett-McGuire, 1991), "a complex challenge" (Mahn, 1987), "a most wretched question" (Malmkjær, 1998), and "a thorny problem" (Snell-Hornby, 1992). Historically, TS has favoured "translation criticism over empirical measurement" (Moorkens et al., 2018, p. 12), with a particular emphasis on literary works.

Equivalence is a cornerstone of early TQA (House, 2015, p. 21-22). The concept of equivalence in TS describes the relationship between the source text (ST) and the target text (TT), in which the TT aims to match the ST in terms of meaning, function, and effect. However, equivalence at all levels is often impossible due to linguistic and cultural differences. The concept has evolved through various scholars, from Vinay and Darbelnet (1958) focusing on the stylistic impact, to linguistic categorization by Jakobson (1959), and Nida and Taber (1969) distinguishing between formal and dynamic equivalence. Early TQA models focused on achieving textual and formal equivalence between source and target texts (Lauscher, 2000). However, equivalence-based TQA approaches have often been criticized for being too rigid and not accommodating the diverse functions translations can serve. For a more in-depth overview of the concept, see Appendix B.

Compared to the concept of equivalence, functionalism in translation (Vermeer, 1978; Honig, 1997) emphasizes the purpose and function of translations within their specific contexts over strict equivalence to the source text (Lauscher, 2000). For example, under Skopos theory, translation quality is assessed by how well translations achieve their intended purpose. In doing so, assessments will consider the cultural and situational appropriateness of the translation to ensure it resonates with the target audience and serves its intended function.

The introduction of translation technologies to the translation industry, such as MT and computer-aided translation (CAT) tools, further impact the assessment of translation quality. Bowker (2019, p. 453-468) emphasizes evaluating translations based on their suitability for their intended purpose rather than adhering to a one-size-fits-all notion of quality. This perspective helps translators navigate the "Triple Constraint" of quality, cost, and time, ensuring that translations meet specific end-user needs. By informing clients about the significance of defining the translation's purpose and agreeing on specifications, translators ensure that their work focuses on both linguistic merit and overall effectiveness in fulfilling intended purposes.

3.3 Modern TQA Models and Methodologies

Modern TQA methods can be categorized into quantitative and qualitative dimensions. Quantitative TQA models aim to provide measurable standards and numeric descriptions of translation quality. Qualitative TQA models look at how well the translation conveys the original message, fits within the cultural and contextual setting, and meets the needs of its intended audience. Quantitative TQA models, during the assessment, may break down the translation work into smaller units, e.g., paragraphs, sentences, or even phrases (an approach that is also common in MT); qualitative TQA models usually look at the complete work of the translation as a whole. We begin with a brief discussion of quantitative approaches, as those more closely resemble

the MT-style evaluations, before examining qualitative approaches.

Many quantitative models are based on error typologies. Canadian Language Quality Measurement System (Sical) and the Canadian Translators, Terminologists and Interpreters Council (CTTIC) certification exam (CTTIC, 2021) emphasize a structured, numerical approach to quality evaluation. The CTTIC's error-based assessment has a "Marking Scale" that differentiates between major "Translation" errors, like significant misinterpretations (-10 points), and minor "Language" inaccuracies (-5 points). We should note, though, that the CTTIC exam is arguably an evaluation/grading of a *translator*, rather than an evaluation of translation or translated texts more generally.

Other types of error typologies have grown from the intersections of technology and translation. MQM introduces over 100 issue types, arranged in a hierarchical structure (Lommel et al., 2014). They use five main branches: Fluency, Accuracy, Verity, Design, and Internationalization and evaluate translations according to specific project requirements and communicative purposes by selecting relevant issue types. MQM supports multiple levels of granularity and includes tools for calculating quality metrics and is used for both human and machine translation evaluations.

Another type is introduced in ISO (2024), which focuses on segment-based comparisons and detailed error typology to promote objective and reliable quality assessment. The error categories cover points such as: terminology (e.g., inconsistent use of terms), accuracy (e.g., mistranslation, omissions), linguistic conventions (e.g., grammar, spelling), style (e.g., register, unidiomatic style), locale conventions (e.g., formats of dates and currencies), audience appropriateness (e.g., cultural references), and design and markup (e.g., character formatting, layout). Error annotations are made based on the relevant translation project specifications and translation evaluation specifications. To further assist users in analyzing their evaluation needs, that document contains appendices with guiding questions to help users determine their evaluation needs and think about how to best implement an evaluation setup for their situation, covering translation use cases, evaluation purposes, and constraints.

Many modern qualitative TQA models draw on linguistic and functionalist approaches. The model in House (1977, 1997, 2015) is rooted in functional pragmatics. It employs a register analysis (an analysis of the variety of language used in a particular situation/for a particular purpose) to assess how well the source and target text match in terms of these dimensions. In particular, the model analyzes field, tenor, and mode (Halliday, 1973; Halliday and Hasan, 1989)—roughly domain, relationship between the translation participants, and medium of communication—as well as genre.[5]

The functionalist/componential (i.e., breaking quality down into components) approach, described in Colina (2008), evaluates various components of translation quality separately based on their functions or purposes. The evaluation tool—similar to a grading or evaluation rubric—includes descriptive statements for different categories such as linguistic form, functional adequacy, meaning, and specialized content. Raters select descriptors that best match the text's quality in each category, which are then converted into numerical scores for analysis. By separating the evaluation into distinct, well-defined components, the componential nature of the tool likely contributed to the better inter-rater agreement observed in this study, reducing ambiguity and subjectivity. Another aspect of this study was that all raters were given training and an explanation of the methodology before participating, which may have contributed to their confidence and high levels of agreement.

Qualitative TQA models also draw on an end-user-focused approach to complement error typologies. Bowker (2009) used recipient evaluation—surveying the target audience about how well various translation options meet their needs and expectations—to assess quality. This approach positions the end-users of translation at the centre, examining how different language communities may have different use cases, needs and requirements. Similarly, Saldanha and O'Brien (2014) proposed using diverse research instruments, such as questionnaires and eye-tracking, to make a more flexible and precise TQA method to adapt to genre, text function, and translation briefs. Han (2020) also highlighted this integration of various methods to enhance reliability, validity, and practicality to emphasize the

[5]See Appendix C for more details.

need for robust and pragmatic assessment methods to address challenges in evaluating translation quality.

There are also models, such as the argumentation-centred TQA from Williams (2004), which combine both qualitative and quantitative methods. This model focuses on assessing instrumental translations[6] by evaluating how well reasoning and arguments are transferred. It uses two main components: argument schema, including elements like claims, grounds, and rebuttals, and rhetorical topology, which encompasses organizational relations, propositions, and narrative strategies. This model employs a detailed framework for deconstructing arguments to ensure the factual content and the persuasive force are accurately conveyed. It also assigns numerical values to various parameters, including core and field-specific elements like terminology and formatting, to provide a comprehensive quality assessment.

To conclude, TQA is complex in both the academic and industry sectors of translation and localization (Castilho et al., 2018). The lack of a universally agreed-upon measurement standard for quality underscores a broader debate on TQA methodologies, particularly with the increasing integration of MT and human translation in various contexts. As a whole, TQA models in TS advocate for a holistic and context-sensitive evaluation of translation quality, acknowledging that different contexts and purposes require different quality standards, but they differ in how to assess and evaluate these.

4 Overlaps and Differences in MT and TS

The fields of MT and TS do have points of commonality when it comes to human evaluation and assessment of MT quality, while the areas where they differ may have their origins in the underlying objectives and methodologies of the fields. This is noted in Castilho et al. (2018), who point out that many researchers in TS "have argued that evaluation is directly associated with the underlying translation theory that one subscribes to," citing in particular the quote that "different views of translation lead to different concepts of translational quality, and hence different ways of assessing it" (House, 1997). We also note that even within each research community—and in their areas of overlap—there is not a broad consensus on how to define quality or which aspects ought to be considered most important.

There are two main forms of evaluations that overlap between the two fields: error typology-based and task-based evaluations. MQM, developed in the translation industry and TS and recently adopted by MT for some evaluations (Freitag et al., 2021a; Anastasopoulos et al., 2022; Agarwal et al., 2023), breaks MT quality down into a typology of errors. Task-based and recipient evaluations have also been used in both fields. In task-based evaluations, we see the use of MT presented in a particular context, with the users asked to either perform a task or evaluate it from the perspective of their use case. These may come the closest to examining whether or not translations are appropriate for the situations and contexts for which they are intended.

Both fields have recognized the challenge of defining and assessing translation quality, though they have largely taken different approaches in exploring this. At various points in MT evaluation history, we have seen quality broken down into component parts at different levels of granularity (e.g., adequacy and fluency, or error typologies like MQM). Perspectives from TS provide other ways of categorizing the components that come together to make up notions of "quality". These include extra-textual factors that influence quality, as well as borrowing and incorporating understandings of quality from different disciplines, such as functionalism, industry, and management. But in all of these efforts, we see that quality is multidimensional (i.e., made up of various contributing aspects) and situation-dependent; there is not a straightforward simple or

[6] "Instrumental translation" refers to a type of translation where the target text functions independently and serves as an instrument for communication within the target culture (Nord, 1997b). Unlike documentary translation, which focuses on reflecting the source text's original context and form, instrumental translation adapts the source text to meet the communicative needs of the target culture. While this concept bears some resemblance to the distinction between covert and overt translation described in House (1997), the two should not be conflated. Covert translations, like instrumental translations, aim to blend seamlessly into the target culture. However, instrumental translation places a particular emphasis on the functional adaptation of the text to serve the target audience effectively, sometimes requiring significant modifications to the source text. This approach is especially relevant in technical, pragmatic, and other context-sensitive translations where functional equivalence is prioritized.

universal definition.

MT research often seeks a single "objective" metric for MT "quality", which can be used to compute a simple ranking of systems. This is connected to the leaderboard and competition aspects that are common to MT and other areas of machine learning. It also directly relates to the fact that system optimization is a major focus in MT research: optimization towards a single objective is substantially easier than optimization towards multiple (potentially conflicting) objectives. In MT, the response to observing annotator variation in evaluation has often been to modify the evaluation protocols (e.g., changing from a rating scale to ranking to direct assessment) or to seek ways of standardizing annotator scores.

Both MT and TS have considered the question of who should perform annotations. In large-scale MT evaluations, this has often been constrained by the cost of annotation, with interest in crowdsourcing (Callison-Burch, 2009; Denkowski and Lavie, 2010; Bentivogli et al., 2011; Graham et al., 2013b, i.a.) and comparing crowdsourced results against language and translation experts. In TS, Colina (2008) also examines this question of who should perform annotations, finding greater levels of inter-annotator agreement within homogenous groups (e.g., groups of all professional translators or groups of bilinguals who are not translators).

In TS, we see more attention paid to the meaning of quality itself and how best to define that, influenced by definitions and descriptions of quality from different disciplines. TS also tends more towards exploring the notion of subjectivity, with a greater focus on the specific use cases and users of a particular translation and how that translation serves its purpose. This focus on a specific use case can be seen as a difference from MT research, which often purports to aim for a broad or universal use case (see, e.g., the framing of tasks at WMT, such as "News" or "General" translation, without reference to a specific audience for the news/general translations).[7] In TS, there is a significant focus on who the translation is for, what it is intended to do, and the specific circumstances surrounding its creation and use (Bowker, 2009; Chesterman and Wagner, 2002; Colina, 2008, i.a.). This approach ensures that translations are tailored to meet the needs and expectations of their target audience. This contrasts with some MT research proposals of a translation that can be used in any context by anyone. This MT perspective may be tied to underlying assumptions of invertibility as a desired component of MT (since round-trip translation performance has frequently been used as a benchmark of success by MT researchers), a view which is not shared in all of TS.

A recent concept from the MT perspective assumes that a single translation can meet all purposes or that there exists a general-purpose translation, which is often unrealistic given the diversity of language use and cultural contexts.[8] While the concept of a universal translator has long been a goal of some researchers, we note that, over time, MT research has taken various views on how best to approach translation. Early MT successes such as the METEO systems (Chandioux, 1976, i.a.) occurred through focused efforts on limited and specific domains: purpose-built MT. The late 1990s and early 2000s saw the widespread availability of free public online MT systems, such as AltaVista Babel Fish (Yang and Lange, 1998), allowing anyone with an internet connection to (attempt to) translate anything within a limited set of language pairs. MT research has seen both these research tracks—the purpose-built task-specific translation system and the goal of a universal system—pursued in parallel. When researchers or users treat online MT systems, for example, as a box into which any source text can be placed with the expectation of receiving the desired translation, conflict and disappointment are likely to arise. Users of MT technologies are in fact using MT with a purpose, and two users of the same MT system may be using it with two different and

[7] We do note some exceptions to this, such as the specification that the 2024 English–Spanish task is intended to translate into Latin American Spanish, specifically (WMT, 2024), though one could argue that this still covers a wide range of language variants.

[8] For example, we know that it is frequently the case that sentences in isolation may have ambiguities that would require additional context to resolve for translation (Castilho et al., 2020); MT systems that translate at the sentence level will struggle with this. Similarly, if we do not specify language variant well enough, we may produce text that is suitable for one linguistic community that speaks a language but not another (e.g., orthographic, writing system, or vocabulary differences). While most MT evaluations omit such factors as design, layout, formatting, and markup, these factors are more frequently considered in the TS perspective. Consider, for example the task of subtitle or closed-caption translation, which places additional constraints, such as length, on the translation, which we are now also beginning to see addressed in MT.

conflicting purposes. The emphasis on universality that is often present (implicitly or explicitly) in MT research may overlook the specificities that TS scholars deem crucial for high-quality translations. We argue that MTR research should be considering these purposes and specificities when performing evaluations, whether by explicitly highlighting specific use cases, language variants, and so on, or by being clear about how to handle conflicting preferences in translation quality.

5 Evaluation Briefs

In this work, we have looked at how both MT and TS have explored questions of what it means to evaluate the "quality" of a translation. While we have seen that MT has explored some aspects of quality (e.g., adequacy and fluency), TS has enumerated a wider range of aspects that contribute to perceptions and judgments of MT quality; TQA involves decisions that take into account many factors beyond the source and target text, such as the intended target audiences and their linguistic and cultural background, the purpose of translation, and the medium of reception. Without access to these relevant details, human evaluators are reasoning under uncertainty. This leads us to ask: can insights from TS suggest to us aspects that are missing from many of the current implementations of human evaluation of MT from the MT research side?

We argue that the concepts of the purpose and intended audience of a translation are some of the central aspects that have been underexplored in the MT literature. This is also one of the major research areas identified in human-centered MT evaluation by Liebling et al. (2022). Trying to incorporate this into MT evaluation (e.g., of the sorts performed at WMT or other large-scale evaluations) will require MT researchers to first settle on more concrete and well-defined goals for their MT systems. That includes the considerations of the intended use case, the language variants, and the intended audiences.

This is certainly not a new call; Church and Hovy (1993) pointed out that "if the application is not clearly identified (or worse, if the application is poorly chosen), then it is often very difficult to find a satisfying evaluation paradigm." That claim was made in an era of "crummy" MT, but we argue for its continued relevance in an era of improved MT. Among several other goals, Church and Hovy (1993) argued that an appropriate application should "set reasonable expectations" and "should be attractive to the intended users". Now that we have access to much-improved MT for many language pairs and domains, how should we push forward?

We propose being explicit with a "translation brief" (for the use of both the translators producing reference translations and the researchers building MT systems[9]) as well as expanding this to an "evaluation brief". An "evaluation brief" would provide the human evaluators with a wider context and detailed instructions about how to evaluate the translation. This is similar to the "role" or "persona" described in Graham et al. (2012), which annotators are asked to take on when evaluating MT output; that work also highlights the importance of taking great care with the design of such instructions. With the evaluation brief, human evaluators could situate themselves in the use case of the translation and as the intended users of the translation to consider the users' needs and expectations. As for what to include in an evaluation brief, we could draw inspiration from the translation brief: source and target languages (including language variants), relevant information about both the author/speaker and the audience, purpose, style guide, and so on (ISO, 2015).

For example, we can consider two different types of medical texts: medical information that is intended for healthcare workers (domain experts) and medical information in public health announcements that is intended to be accessible to a broad audience (non-experts).[10] An appropriate evaluation

[9]While human translators will make use of the translation brief directly, i.e., deciding on levels of formality, language variants, technical language, and so on to use in their translations, MT researchers are likely to use this more indirectly, such as by selecting which data sources to train on, deciding whether to incorporate model features such as tagging (e.g., for multi-domain or multilingual systems), considering issues such as robustness to input variations, and so on, with the goal of producing a translation system that in turn will follow the translation brief. We could also imagine employing translation briefs when experimenting with large language model-based translation, as part of the instructions provided to the large language model.

[10]We consider here primarily the case in which the expertise level of the audience is held consistent from the source to the target (i.e., translating text for domain experts from a source language into target language text also intended for domain experts); the transformation of text from expert to non-expert (or vice versa) introduces additional challenges.

brief would, at a minimum, indicate which audience and purpose was intended, and perhaps also other relevant concerns like whether there were terminological conventions that should be followed. Importantly, the translation brief (for translators producing reference translations and for MT researchers building systems) and the evaluation brief should generally be in agreement; while there may be some situations (e.g., challenge sets or analyses of MT robustness) where it is appropriate to evaluate MT systems on things outside of the purview of the translation brief, to be fair to the participants of a shared task, the evaluation should match the stated objectives of the task itself.

However, an evaluation brief is likely insufficient on its own; MT researchers also need to think about recruiting human evaluators with skills, knowledge, and cultural expertise appropriate for the specific goals of the translation. In the case of translations that are intended to be acceptable across a wide range of language variants (e.g., dialects, spelling conventions), whether the evaluator pool reflects this diversity would affect the validity of the evaluation results. Similarly, in the case of translations for a highly-technical domain (intended for use by experts), e.g., biomedical translations in (Neves et al., 2023), employing subject matter experts as translation evaluators is necessary for a meaningful evaluation. A lack of such experts may lower evaluation consistency (Freitag et al., 2021b). Importantly, the evaluation brief (and any translation brief) should be reported (e.g., in the appendices of publications), along with relevant information about the annotators (e.g., language skills, expertise, etc.). Current practices often report only high-level information (e.g., whether annotators were translators or non-translator bilinguals); one may wish to consider expanding this to cover a broader range of relevant demographic information about annotators.

6 Conclusion

In this work, we have examined perspectives on both MT quality and how to evaluate MT from the perspectives of MT research and translation studies. We argue that future MT evaluation could benefit from drawing on insights from translation studies. In particular, this includes an increasing focus on the purpose, intended audience, and context of translation. More broadly, we encourage MT researchers to seek collaborations and conversations in TS and beyond. In order to better design the questions that MT researchers ask of evaluators, the field would likely benefit from more interactions with research best practices in measurement theory, survey research methods, human-computer interaction, and more.

Acknowledgments

We thank Malcolm Williams for discussion of his TQA model, and Gabriel Bernier-Colborne for feedback on a draft of this paper.

References

Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., and Zevallos, R. (2023). FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Salesky, E., Federico, M., and Carpuat, M., editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Al Sharou, K. and Specia, L. (2022). A taxonomy and study of critical errors in machine translation. In Moniz, H., Macken, L., Rufener, A., Barrault, L., Costa-jussà, M. R., Declercq, C., Koponen, M., Kemp, E., Pilos, S., Forcada, M. L., Scarton, C., Van den Bogaert, J., Daems, J., Tezcan, A., Vanroy, B., and Fonteyne, M., editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

Anastasopoulos, A., Barrault, L., Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., Currey, A., Dinu, G., Duh, K., Elbayad, M., Emmanuel, C., Estève, Y., Federico, M., Federmann, C., Gahbiche, S., Gong,

H., Grundkiewicz, R., Haddow, B., Hsu, B., Javorský, D., Kloudová, V., Lakew, S., Ma, X., Mathur, P., McNamee, P., Murray, K., Nădejde, M., Nakamura, S., Negri, M., Niehues, J., Niu, X., Ortega, J., Pino, J., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Turchi, M., Virkar, Y., Waibel, A., Wang, C., and Watanabe, S. (2022). Findings of the IWSLT 2022 evaluation campaign. In Salesky, E., Federico, M., and Costa-jussà, M., editors, *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Baker, M. (1992). *In Other Words: A Coursebook on Translation*. Routledge.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bassnett-McGuire, S. (1991). *Translation Studies*. Routledge, London, revised edition edition.

Bentivogli, L., Federico, M., Moretti, G., and Paul, M. (2011). Getting expert quality from the crowd for machine translation evaluation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Birch, A., Abend, O., Bojar, O., and Haddow, B. (2016). HUME: Human UCCA-based evaluation of machine translation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Bowker, L. (2000). A corpus-based approach to evaluating student translations. *The Translator*, 6(2):183–210.

Bowker, L. (2009). Can machine translation meet the needs of official language minority communities in Canada? A recipient evaluation. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 8:123–155.

Bowker, L. (2019). Fit-for-purpose translation. In O'Hagan, M., editor, *The Routledge Handbook of Translation and Technology*. Routledge.

Burchardt, A. (2013). Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In Koehn, P. and Mihalcea, R., editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) evaluation of machine translation. In Callison-Burch, C., Koehn, P., Fordyce, C. S., and Monz, C., editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J., and Fordyce, C. S., editors, *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J., editors, *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Calvo, E. (2018). From translation briefs to quality standards: Functionalist theories in today's translation processes. *Translation & Interpreting*, 10(1).

Carl, M., editor (2021). *Explorations in Empirical Translation Process Research*. Springer International Publishing.

Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (2018). *Approaches to Human and Machine Translation Quality Assessment*, pages 9–38. Volume 1 of Moorkens et al. (2018).

Castilho, S. and Knowles, R. (2024). A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, page 1–31.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Catford, J. (1965). *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Language and language learning series. Oxford U.P.

Chandioux, J. (1976). METEO, an operational system for the translation of public weather forecasts. In Hays, D. G. and Mathias, J., editors, *Foreign Broadcast Information Service Seminar on Machine Translation*, pages 27–36, Virginia.

Chesterman, A. and Wagner, E. (2002). *Can Theory Help Translators?: A Dialogue Between the Ivory Tower and the Wordface*. Routledge.

Church, K. W. and Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

Colina, S. (2008). Translation quality evaluation: Empirical evidence for a functionalist approach. *The Translator*, 14(1):97–134.

CTTIC (2021). CTTIC translation certification examination marker's guide. Accessed: 2024-06-02.

de Waard, J. and Nida, E. (1986). *From One Language to Another: Functional Equivalence in Bible Translating*. Nelson.

Denkowski, M. and Lavie, A. (2010). Exploring normalization techniques for human judgments of machine translation adequacy collected using Amazon Mechanical Turk. In Callison-Burch, C. and Dredze, M., editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 57–61, Los Angeles. Association for Computational Linguistics.

Dimitrova, B. E. (2010). Translation process. In Gambier, Y. and van Doorslaer, L., editors, *Handbook of Translation Studies*, volume 1, pages 406–411. John Benjamins Publishing Company.

Drugan, J. (2013). *Quality in Professional Translation: Assessment and Improvement*. Bloomsbury.

Esselink, B. (2003). *Localisation and translation*, page 67–86. John Benjamins Publishing Company.

Fields, P., Hague, D., Koby, G., Lommel, A., and Melby, A. (2014). What is quality? A management discipline and the translation industry get acquainted. *Tradumàtica: tecnologies de la traducció*.

Fowler, F. (2013). *Survey Research Methods*. Applied Social Research Methods. SAGE Publications.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (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.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021b). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Garvin, D. A. (1984). What does "product quality" really mean? *Sloan Management Review*, 26(1):25–43.

Graham, Y., Baldwin, T., Dowling, M., Eskevich, M., Lynn, T., and Tounsi, L. (2016). Is all that glitters in machine translation quality estimation really gold? In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Graham, Y., Baldwin, T., Harwood, A., Moffat, A., and Zobel, J. (2012). Measurement of progress in machine translation. In Cook, P. and Nowson, S., editors, *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 70–78, Dunedin, New Zealand.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013a). Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013b). Crowd-sourcing of human judgments of machine translation fluency. In Karimi, S. and Verspoor, K., editors, *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 16–24, Brisbane, Australia.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2014). Is machine translation getting better over time? In Wintner, S., Goldwater, S., and Riezler, S., editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Halliday, M. (1973). *Explorations in the Functions of Language*. Arnold, London.

Halliday, M. and Hasan, R. (1989). *Language, Context and Text: Aspects of Language in a Social Semiotic Perspective*. Oxford University Press, Oxford.

Han, C. (2020). Translation quality assessment: A critical methodological review. *The Translator*, 26(3):257–273.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018).

Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Holmes, J. S. (1988). *Translated! Papers on Literary Translation and Translation Studies*. Rodopi, Amsterdam.

Honig, H. G. (1997). Positions, power and practice: Functionalist approaches and translation quality assessment. *Current Issues In Language and Society*, 4(1):6–34.

House, J. (1977). *A Model for Translation Quality Assessment*. Narr, Tübingen, 2nd edition.

House, J. (1997). *Translation Quality Assessment: A Model Revisited*. Tübinger Beiträge zur Linguistik. G. Narr.

House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta: Journal des traducteurs*, 46(2):243–257.

House, J. (2015). *Translation Quality Assessment: Past and Present*. Routledge.

ISO (2015). *International Standard ISO 17100:2015 Translation services - Requirements for translation services*. International Organization for Standardization.

ISO (2024). *International Standard ISO 5060:2024 Translation services - Evaluation of translation output - General guidance*. International Organization for Standardization.

Jakobsen, A. L. (2017). Translation process research. *The Handbook of Translation and Cognition*, page 19–49.

Jakobson, R. (1959). On linguistic aspects of translation. In *On translation*, pages 232–239. Harvard University Press.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M. (2007). ILR-based MT comprehension test with multi-level questions. In Sidner, C., Schultz, T., Stone, M., and Zhai, C., editors, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80, Rochester, New York. Association for Computational Linguistics.

Knowles, R. (2021). On the stability of system rankings at WMT. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. Association for Computational Linguistics.

Koby, G. S., Fields, P., Hague, D., Lommel, A., and Melby, A. (2014). Defining translation quality. *Revista Tradumàtica: tecnologies de la traducció*, 12:404–412.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., and Shmatova, M. (2023). Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., and Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., and Shmatova, M. (2024). Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In Koehn, P. and Monz, C., editors, *Proceedings on the Workshop on Statistical Machine*

Translation, pages 102–121, New York City. Association for Computational Linguistics.

Koller, W. (1979). *Einführung in die Übersetzungswissenschaft*. UTB 819. Quelle & Meyer.

Koller, W. (1989). Equivalence in translation theory. In Chesterman, A., editor, *Readings in translation theory*, pages 99–104. Oy Finn Lectura Ab, Helsinki.

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

Laoudi, J., Tate, C. R., and Voss, C. R. (2006). Task-based MT evaluation: From who/when/where extraction to event understanding. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Lauscher, S. (2000). Translation quality assessment: Where can theory and practice meet? *The Translator*, 6(2):149–168.

LDC (2002). Linguistic data annotation specification: Assessment of fluency and adequacy in Arabic-English and Chinese-English translations.

Licht, D., Gao, C., Lam, J., Guzman, F., Diab, M., and Koehn, P. (2022). Consistent human evaluation of machine translation across language pairs. In Duh, K. and Guzmán, F., editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Liebling, D., Heller, K., Robertson, S., and Deng, W. (2022). Opportunities for human-centered evaluation of machine translation systems. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 229–240, Seattle, United States. Association for Computational Linguistics.

Lim, Z. W., Vylomova, E., Cohn, T., and Kemp, C. (2024). Simpson's paradox and the accuracy-fluency tradeoff in translation. *arXiv preprint arXiv:2402.12690*.

Lo, C.-k. and Wu, D. (2011). MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA. Association for Computational Linguistics.

Lo, C.-k. and Wu, D. (2014). On the reliability and inter-annotator agreement of human semantic MT evaluation via HMEANT. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 602–607, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 1(12):455–463.

Mahn, G. (1987). Foreign language proficiency criteria in translation. In Rose, M. G., editor, *Translation Excellence: Assessment, Achievement, Maintenance*, pages 44–45. SUNY, Binghamton.

Malmkjær, K. (1998). Linguistics in functionland and through the front door: A response to hans g. hönig. In Schäffner, C., editor, *Translation and Quality*, pages 70–74. Multilingual Matters, Clevedon.

Moghe, N., Sherborne, T., Steedman, M., and Birch, A. (2023). Extrinsic evaluation of machine translation metrics.

Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (2018). *Translation Quality Assessment From Principles to Practice*. Springer.

Neves, M., Jimeno Yepes, A., Névéol, A., Bawden, R., Di Nunzio, G. M., Roller, R., Thomas, P., Vezzani, F., Vicente Navarro, M., Yeganova, L., Wiemann, D., and Grozea, C. (2023). Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Nida, E. (1964). *Towards a Science of Translating*. BRILL.

Nida, E. and Taber, C. (1969). *The Theory and Practice of Translation*. Helps for translators. E. J. Brill.

Nord, C. (1997a). Defining translation functions: The translation brief as a guideline for the trainee translator. *Ilha Do Desterro*, 33:39–54.

Nord, C. (1997b). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. St. Jerome.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pym, A. (2019). Quality. In O'Hagan, M., editor, *The Routledge Handbook of Translation and Technology (1st ed.)*. Routledge.

Pym, A. (2023). *Exploring Translation Theories*. Routledge.

Saldanha, G. and O'Brien, S. (2014). Product-oriented research. In *Research Methodologies in Translation Studies*, pages 50–108. Routledge.

Snell-Hornby, M. (1992). The professional translator of tomorrow: Language specialist or all-round expert? In Dollerup, C. and Loddegaard, A., editors, *Teaching Translation and Interpreting: Training, Talent and Experience*, pages 9–22. John Benjamins, Amsterdam.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Vandepitte, S. (2017). Translation product quality: A conceptual analysis. In Svoboda, T., Biel, L., and Łoboda, K., editors, *Quality aspects in institutional translation*, pages 15–29. Language Science Press, Berlin.

Vermeer, H. J. (1978). Ein rahmen für eine allgemeine translationstheorie. *Lebende Sprachen*, 23:99–102.

Vermeer, H. J. (2021). Skopos and commission in translational action. In Chesterman, A., editor, *The Translation Studies Reader*. Routledge, 4th edition.

Vilar, D., Leusch, G., Ney, H., and Banchs, R. E. (2007). Human evaluation of machine translation through binary system comparisons. In Callison-Burch, C., Koehn, P., Fordyce, C. S., and Monz, C., editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

Vinay, J. and Darbelnet, J. (1958). *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

Williams, M. (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*. University of Ottawa Press.

WMT (2024). Shared task: General machine translation. https://www2.statmt.org/wmt24/translation-task.html. Accessed: 2024-08-07.

Yang, J. and Lange, E. D. (1998). SYSTRAN on AltaVista. In Farwell, D., Gerber, L., and Hovy, E., editors, *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Tech-*

nical Papers, pages 275–285, Langhorne, PA, USA. Springer.

A Questions for DA Annotators in WMT

In the original work of Graham et al. (2013a), annotators were asked questions about fluency using the DA sliding scales. In the Mechanical Turk setup in Bojar et al. (2016, 2017, 2018); Barrault et al. (2019) annotators were asked: "Read the text below. How much do you agree with the following statement:" where the statement was "The black text adequately expresses the meaning of the gray text in English." (*English* was replaced with a different language where appropriate). The Appraise interface used the question "How accurately does the above candidate text convey the original semantics of the reference text? Slider ranges from "Not at all (left) to Perfectly (right)."—though in bilingual assessment, "reference" was replaced with "source" (Bojar et al., 2017, 2018). In Barrault et al. (2019) the Appraise setup asked annotators: "For the pair of sentences below: Read the text and state how much you agree that:" where the statement was "The black text adequately expresses the meaning of the gray text in German (deutsch)." (replaced with a different language where appropriate). Later evaluations with a different interface added clarifications about the location of the two texts (Barrault et al., 2020).

The most recent two WMT shared tasks have used an approach that they call DA+SQM; that interface uses a continuous slider to assign scores on a 7-point (0 to 6) scale, with the 0, 2, 4, and 6 tick marks attached to labels. These labels and their descriptions explicitly include both meaning and grammar, as we see in this example: "4: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies." (Kocmi et al., 2023). This decision to use DA+SQM and these particular labels was supported by "internal preliminary experiments" (Kocmi et al., 2022) that showed that it may produce more stable scores across annotators; the results and supporting data have not been released publicly.

B Translation Equivalence in TS

Table 1 shows different understandings of equivalence in TS, acknowledging that the target text (TT) can never be equivalent to the source text (ST) on all levels (Vandepitte, 2017, p. 151). Vinay and Darbelnet (1958, p.32) suggested that the stylistic impact in translation is critical. Jakobson (1959, p. 233) took a linguistic approach, discussing different types of translation (intralingual, interlingual, and inter-semiotic). Nida (1964); Nida and Taber (1969); de Waard and Nida (1986) drew on Bible Studies and distinguished between formal and functional equivalence, stressing the importance of message over form. Catford (1965, p.27) introduced the concept of "textual equivalent", which refers to a text or section of text in the TT that, in a specific situation, is deemed equivalent to a corresponding text or segment in the ST. This work underscores the challenges of achieving equivalence across languages and the critical role of context in defining linguistic meaning, distinguishing between "textual equivalence" and "formal correspondence" based on their respective roles in translation. House (1997) emphasized functional equivalence between the ST and TT. Koller (1979, 1989) identified five types of equivalence, ranging from denotative to pragmatic. Baker (1992), from a linguistic perspective, elaborated on text-level and pragmatic equivalence. Pym (2023, p. 10-12) framed equivalence as a relationship of 'equal value' between segments of the ST and TT from form to function. While languages and cultures may differ, translations can achieve equivalence by preserving some aspect of value, whether it be in terms of meaning, function, or effect. The work emphasized that equivalence involves "transformation", aiming to preserve or reproduce a certain value from the ST in the TT. This perspective emphasizes the translator's role in navigating cultural differences and making deliberate choices to ensure the translation fulfills its intended purpose, whether that be informing, persuading, or entertaining the target audience.

Exploring these perspectives provides a context for the evolution of TQA approaches. The discussions by Vinay and Darbelnet (1958), Jakobson (1959), Nida and Taber (1969), and others laid the groundwork for what was predominantly a qualitative assessment of translations, rooted in linguistic, functional, and stylistic parameters. This era's

Representative work	Key understandings of equivalence
Vinay and Darbelnet (1958)	Replicate the same message with different wording (p. 32); an emphasis on the stylistic impact in the target text (TT) (p. 256)
Jakobson (1959)	Three kinds of translation: intralingual, interlingual, and inter-semiotic, with interlingual translation as the focus in TS; there is no full equivalence between code-units; translation from one language into another substitutes messages in one language not for separate code-units but for entire messages in some other language (p. 233).
Nida (1964); Nida and Taber (1969); de Waard and Nida (1986)	Two basic types of equivalence: (1) formal equivalence (fidelity to the original text) and; (2) dynamic equivalence; a translation is to seek equivalence of the message rather than conserving the form of the utterance; meaning is given priority over structure; style, though secondary to content, must still be preserved (1986, p. 36)
Catford (1965)	"Translation is an operation performed on language: a process of substituting a text in one language for a text in another. Then, any theory of translation must draw upon a theory of language – a general linguistic theory." (p. 1) Textual equivalence is "any target language text or portion of text which is observed on a particular occasion to be equivalent of a given ST or portion of text" (p. 27) Formal correspondence is "any TL category (unit, class, structure) which can be said to occupy as nearly as possible the same place in the economy of the TT as the ST given category occupied in the ST" (p. 27)
House (1997)	An emphasis on functional equivalence between the ST and the TT.
Koller (1979, 1989)	Five different types of equivalence: denotative (extra-linguistic factors), connotative (verbalized through source text), text-normative (textual and linguistic norms), pragmatic (concerning the receiver of the target text) and formal (the formal-aesthetic qualities of the source text).
Baker (1992)	Word-level equivalence (p. 9-49); grammatical-level equivalence (p. 92-129); textual-level equivalence (cohesion and thematic structure) (p.131-228); pragmatic level equivalence (mainly with implications which refers to the implied not the literal meanings) (p. 230-271).
Pym (2023)	Transformation-based equivalence (p.12)

Table 1: A timeline of understanding "translation equivalence" in TS.

TQA was characterized by its reliance on human expertise, with scholars advocating for various frameworks to grapple with the intangible qualities of a "good translation". These early debates and theories remain influential, offering a point of departure for understanding how the advent of technology has reshaped the methodologies and tools of TQA.

C House's TQA Model (2015)

The House TQA model (House, 2015, p. 127) employs a register analysis derived from the framework in Halliday (1973) and Halliday and Hasan (1989), utilizing the categories of field, tenor, and mode. It includes six parameters:

- **Field**: This refers to the domain of knowledge or social practice that the text relates to and the activities that it refers to. It answers the question of "what is happening" or "what is being talked about." For example, a scientific report on climate change will have a different field than a personal letter, affecting the choice of technical versus everyday language.

- **Tenor**: This includes the participant relationships, the author's provenance, social relationships, social attitudes, and participation. It reflects the social roles and relationships between the participants (e.g., teacher-student, doctor-patient, friend-friend) involved in the communicative event, including the author, the reader, and the translator. Tenor influences aspects of language such as the level of formality, use of pronouns, and the choice of modal verbs expressing obligation, possibility, or permission, reflecting the nature of interpersonal interactions.
- **Mode**: Mode refers to the medium of the text, the channel of communication (spoken or written), and the complexity or simplicity of the language, as well as its connectivity. It refers to how the text is presented and how it establishes a connection with the reader.
- **Register**: This is a central concept that draws together the elements of field, tenor, and mode, to describe the language variety used for a particular purpose. For instance, an academic lecture employs specialized vocabulary and complex structures (field), within a formal relationship between lecturer and students (tenor), delivered through a monologic presentation (mode). Conversely, a casual conversation between friends features everyday topics (field), marked by an informal, equal-status interaction (tenor), in a spontaneous, spoken format (mode).
- **Genre**: Genre is understood in terms of socially ratified forms of texts, like a novel, a legal document, or a poem.
- **Corpus Studies**: This is not traditionally part of House's model but suggests a methodological approach to TQA through the use of corpora to analyze translations in a larger, more empirical context.

House's approach to TQA is functionalist and descriptive. A quality translation is functionally equivalent to the ST, meaning it should enable the reader to understand and do the same things as they would with the ST, taking into account the cultural context and the communicative situation of the TT. The emphasis of the model is on the equivalence of the communicative functions of the texts rather than a word-for-word correspondence.