English Studies at NBU, 2026
Vol. 12, Issue 1, pp. 196-219
https://doi.org/10.33919/esnbu.26.1.12

pISSN 2367-5705
eISSN 2367-8704
www.esnbu.org

INVESTIGATING NEURAL MACHINE TRANSLATION QUALITY: AN XCOMET-XL-BASED ANALYSIS ACROSS KATHARINA REISS'S TEXT TYPOLOGIES

Betül Özcan Dost

Ondokuz Mayıs University, Samsun, Türkiye

Abstract

This study examines AI-driven neural methods that reduce human involvement in machine translation evaluation. Traditional evaluation approaches face challenges related to consistency, scalability, and evaluator subjectivity. To address these limitations, the study employs xCOMET-XL, which evaluates translation quality within a shared semantic space. Within the scope of this study, fifteen Turkish source texts, classified according to Katharina Reiss's typology as informative, expressive, and operative, were translated into English using Google Translate and DeepL. The translations were evaluated using xCOMET-XL, followed by human evaluation. The originality of the study lies in its comparative design examining the alignment between xCOMET-XL scores and human judgments. After comparing results, alignment between the methods was analysed. Findings signal a relatively high degree of alignment with human judgments in certain texts, highlighting neural metrics' potential as a fast, scalable, and systematic alternative. The study contributes to Translation Studies by supporting alternative translation evaluation approaches.

Keywords: machine translation, xCOMET-XL, translation assessment, text types, Katharina Reiss

Article history:

Received: *02 May 2026*
Reviewed: *09 May 2026*
Accepted: *15 May 2026*
Published: *20 June 2026*

Citation: Özcan Dost, B. (2026). Investigating neural machine translation quality: An xCOMET-XL-based analysis across Katharina Reiss's text typologies. *English Studies at NBU, 12*(1), 196-219. https://doi.org/10.33919/esnbu.26.1.12

Dr. Betül Özcan Dost has a BA in Translation and Interpreting (English-Turkish) from Yeditepe University (2009), an MA from İstanbul University (2013) and a PhD in Translation and Cultural Studies from Gazi University (2018). Currently, she is Assistant Professor at Ondokuz Mayıs University. Proficient in English and Spanish, her expertise focuses on machine and audiovisual translation. She has authored numerous academic publications and presentations across various translation fields.

E-mail: betul.ozcan@omu.edu.tr https://orcid.org/0000-0003-3110-8017

The rapid pace of global communication and technology has moved translation into a new dimension, in which researchers and translators seek new methods to supplement the traditional methods and meet the growing demand for fast and high-quality translation. This trend has brought modern technologies such as Machine Translation (MT) and Artificial Intelligence (AI) to the forefront in translation. Although MT tools seek to provide a fully automated and smooth translation process and product by reducing the need for human intervention, their performance remains a subject of ongoing debate.

Within this context, translation quality evaluation has gained importance in Translation Studies (TS). In particular, it mainly relies on a systematic comparison of the Source Text (ST) with the Target Text (TT). This process becomes even more challenging in distant language pairs, such as the English–Turkish pair examined in this study, where comparison becomes more demanding. Therefore, both researchers and translation professionals are searching for reliable approaches that integrate both traditional and modern approaches. Accordingly, the emergence of neural evaluation metrics necessitates a reconsideration of existing practices to keep up with global advancements.

In line with this shift, this study aims to examine the translation quality of MT from Turkish into English through the neural evaluation metric xCOMET-XL, based on Katharina Reiss's text typology. For this purpose, three text types—informative, expressive, and operative—are utilized to examine MT quality and xCOMET-XL performance across different text types. The study also investigates whether modern metrics show a high alignment with human judgments. In line with the study's aim, the research questions are as follows:

- Is there a meaningful difference in MT performance across informative, expressive, and operative text types in terms of xCOMET-XL scores?
- To what extent does xCOMET-XL align with human judgments in assessing the quality of Turkish–English machine translations?

By comparing xCOMET-XL-based evaluations with human judgments, this study offers empirical insight into how effectively neural metrics capture meaning beyond surface-level correspondences. Furthermore, it addresses a significant gap in the

literature by providing a systematic analysis of neural evaluation metrics specifically within the context of the structurally distant Turkish–English language pair across diverse text types. In this respect, it contributes to discussions on the role of AI in translation assessment, highlights the importance of text type-sensitive evaluation, and supports the development of more consistent, scalable, and context-aware assessment models that bridge traditional and data-driven modern approaches.

Literature Review

Katharina Reiss's Text Typology

Text typology provides a framework for text analysis according to their functional, structural, and contextual features. Since texts serve different communicative purposes, they vary in content, linguistic structure, and discourse organization. Therefore, typologies help explain both formal features and communicative functions of texts. Due to its relevance across language-related fields, text types have been classified by various scholars according to criteria such as communicative purpose, linguistic features, and structural organization.

Within this theoretical background, as Delibaş (2023, p. 1020) states, translation involves recreating a text for a target audience while considering its type, communicative purpose, function, and cultural context. Central to this perspective are functionalist approaches, which prioritize the text's communicative role. In line with this paradigm shift in TS, the function-based classifications of text types have become prominent.

Reiss and Vermeer (2014, p. 181) argue that the wide range of texts in professional translation makes it useful to draw on text linguistics to organize this textual "cosmos" into a clear classification system. Such frameworks provide insight into translator behavior beyond intuition and offer a theoretical basis for explaining variations in translation. They highlight the distinction between text typology in text linguistics and Translation Studies based on the specific needs and priorities of each field. Unlike text linguistics, which focuses on monolingual analysis, TS is concerned with texts in translation, pre-translation, or post-translation contexts, which necessitates attention to their functional status in both source and target cultures.

Building on this functionalist perspective, Reiss is a prominent figure who developed a text typology grounded in the three functions of language identified by the German linguist Karl Bühler: informative, expressive, and appellative. She linked these functions to corresponding language dimensions and the text types or communicative situations (Reiss, 1977/1989, pp. 108-110), identifying informative texts, expressive texts, operative texts, and later expanded the classification to include audio-medial texts (Chen & Zhang, 2020, p. 34). By proposing a framework for translators that considers the functions, linguistic characteristics, and translation strategies associated with each text type (Uğurlu & Dolmacı, 2020, p. 87), she made a significant contribution to TS. Munday summarizes the main characteristics of each type as follows:

1. **Informative Text Type**: It is defined by factual content, including information, knowledge, and opinions, and primarily focuses on content, using logical and referential language to ensure clarity.
2. **Expressive Text Type**: It is characterized by creative composition and the aesthetic use of language, foregrounding the author and the stylistic form of the message.
3. **Operative Text Type**: It aims to elicit specific responses; texts such as advertisements and political speeches use persuasive and typically dialogic language to engage and influence the audience.
4. **Audio-medial texts**: This text type integrates visual, auditory, and other sensory elements alongside language, thereby complementing the other text functions and include texts such as films and audiovisual advertisements. (2016, pp. 115-116).

Extending this classification further, hybrid text types are also acknowledged, as texts may display features of more than one category. In such cases, Reiss (1977/1989, p. 109) argues that 'the transmission of the predominant function of the ST is the determining factor by which the TT is judged'. The following table (Table 1), adopted from Reiss by Munday, summarizes these functional characteristics and their corresponding translation methods.

Table 1

Functional characteristics of text types and links to translation methods

Text type	Informative	Expressive	Operative
Language function	Informative (representing objects and facts)	Expressive (expressing sender's attitude)	Appellative (making an appeal to text receiver)
Language dimension	Logical	Aesthetic	Dialogic
Text focus	Content-focused	Form-focused	Appellative focused
TT should…	Transmit referential content	Transmit aesthetic form	Elicit desired response
Translation method	'Plain prose', explicitation as required	'Identifying' method, adopt perspective of ST author	'Adaptive', equivalent effect

Note: Source: (Munday, 2016, p. 115)

In this study, Reiss's text typology was adopted for its systematic and functional relevance to translation evaluation. This framework enables clear distinctions between text types and supports a nuanced analysis of variation in translation quality within the xCOMET-XL based evaluation.

Machine Translation

Machine Translation (MT) enables automatic translation between languages and plays an increasingly important role in global communication. It seeks to provide accurate, fast and cost-effective translations, while reducing the need for human intervention. However, despite the continuous advancements, MT tools still face challenges arising from factors such as differences in language structures, word polysemy, contextual cues, and cultural nuances, which remain areas where human translators are generally considered to perform more effectively. In this context, MT tools are often considered in terms of their ability to preserve the integrity and semantic depth of the ST, along with word-level correspondence. A clear understanding of their main principles is essential for scholars and practitioners to comprehend their advantages and limitations, which subsequently contributes to their improvement in line with ongoing technological developments.

To better understand the principles and developments in the field, it is useful to consider how MT is defined in the literature. In their influential work, Hutchins and

Somers (1992, p. 3) define MT as "computerized systems responsible for the production of translation from one natural language to another, with or without human assistance". Beyond this definition, which establishes the fundamental scope of the technology, it is significant to consider MT's practical development and evolving capabilities over time. In this regard, Korkmaz (2019, p. 164) observes that MT has made significant progress in its relatively short history and evolved from simple dictionary-like systems to more complex platforms capable of processing large volumes of data rapidly and supporting multilingual translation. In line with these developments, Şahin and Duman (2013, p. 409) state that rapid developments in MT tools and their continuously improving database capacities are expected to reduce language barriers, leading to a more globalized world in which communication problems arising from language barriers are gradually minimized across domains of human interaction.

Although technology has recently become more prominent in translation, the use of technology in translation is not a new issue. The historical development of MT can be summarized as outlined by Hutchins (1995) in his study. The early systems appeared in the 1950s and 1960s. However, the ALPAC report affected the process adversely in the mid-1960s despite being criticized as "narrow, biased and shortsighted". In the 1970s, the revival period started mainly based on the needs of countries such as the United States, Canada and European countries that were among pioneers in the field. The 1980s witnessed the emergence of commercial and operational MT tools, together with new research in the field. At that time, there was also expansion and diversification in terms of MT use. In the 1990s, there was new research and subsequently, there was an increase in the use of the systems. (Hutchins, 1995). Building upon this historical foundation, it is clear that MT tools have constantly evolved. They appeared on the scene as simple machines that have decoding functions and an electronic dictionary look. Currently, they can analyze large amounts of data in a short time and provide translations in a variety of languages (Korkmaz, 2019, p. 164). Furthermore, the recent advancements in deep learning enabled MT technology to be combined with AI. This combination generated neural network machine translation, leading to the emergence of a novel phase concerning MT and artificial intelligence translation (Lin, 2023, p. 133). These game-changing Large Language Models (LLMs), which can be described as "AI models that are trained on vast amounts of text data in order to learn how to understand and generate human language", enable more accurate and effective translations thanks to their working

principle based on learning the patterns and relationships within a language by being trained on massive amounts of text data (Ray, 2023, p. 134).

Beyond its historical and technological development, the increasing role of MT in global communication should also be considered as it has increased with the rise of globalization. This growing demand has necessitated new approaches to translation, while technological advancement plays a pivotal role in this transformation. People can now access translation instantaneously and at lower cost compared to traditional methods. Considering its benefits, MT has been integrated into various professional, interpersonal, and social communication contexts worldwide (Orrego-Carmona, 2020, p. 323).

Çetiner (2021, p. 582) states that in recent years, MT has also profoundly influenced the translation industry, with both companies and individual translators increasingly relying on Neural Machine Translation (NMT), which uses AI to generate outputs that appear more fluent and natural than earlier Statistical Machine Translation (SMT) systems. Despite this apparent fluency, Çetiner highlights that NMT outputs often conceal errors that are not immediately visible, making it essential to identify and correct them to ensure professional-quality translations.

While the developments in the field highlight the growing impact of MT on the translation industry, they also point to ongoing challenges related to translation quality and reliability. In this regard, Palanichamy and Trojovský (2024, p. 1) summarize the situation by stating that “Machine translation facilitates cross-linguistic communication by converting text between languages”. However, producing contextually accurate translations remains a challenge. Metin Tekin (2025, p. 162) also states that despite the advantages of MT, it still presents several challenges in terms of accuracy and quality, particularly when dealing with distant or complex languages where errors may compromise meaning and clarity. She further adds that MT often fails to capture nuance, context, register, idiomatic expressions, cultural references, and context-dependent meanings, and therefore requires careful checking, critical evaluation, and revision before it can be used reliably. Building on this discussion, Palanichamy and Trojovský (2024, p. 4) state that “researchers are constantly exploring novel techniques to enhance MT models further, addressing challenges such as domain adaptation, low-resource

languages, and cross-lingual transfer learning", which signals the efforts to find robust evaluation frameworks, such as the neural-based metrics examined in this study.

This technological and industrial progress is also reflected in academic studies. Research on translation technology has increased significantly starting in 2019, with theses and articles introducing new perspectives on MT, translation technologies, and translation education. This growth continued in 2021-2023, with academic output on MT, AI, and related fields roughly doubling, reflecting a growing recognition within Turkish academia and the development of a stronger research tradition in translation technology. This rise in publications also highlights the need for more scholars in this area, in line with the evolving translation industry and the increasing focus on the profession's future (Moralı & Şahin, 2025, pp. 211–212).

Modern Tools for Measuring Translation Quality

Technological advancements on a global scale have not only revolutionized MT but have also reshaped the landscape of translation evaluation. The increasing speed of translations has also increased demand on rapid and reliable translation quality evaluations. Consequently, translation quality evaluation has started to play a more central role in both academic research and professional practice. While traditional methods are appreciated for their focus on factors such as semantic integrity and contextual accuracy, modern tools enable evaluations conducted in a more objective, systematic, and data-driven manner. By enhancing the speed and efficiency of the translation process, they provide more structured evaluation framework for both researchers and translators.

In this context, automatic evaluation metrics are evolving beyond their traditional role in model selection, now serving as tools for enhancing the performance of large language models (LLMs) across their entire development pipeline (Pombal et al., 2025, p. 1319). Furthermore, they are increasingly utilized to evaluate translation performance automatically (Perrella et al., 2024, p. 20689). However, this shift is not without challenges. Ahsan et al. (2025) state that the automatic evaluation of translation remains challenging due to the orthographic, morphological, syntactic, and semantic diversity among languages. While neural-based metrics have helped overcome some limitations of traditional string-based measures, their effectiveness is still limited by the scarcity of gold-standard evaluation data, especially for languages beyond the typical high-resource

pairs. Based on these challenges, they describe translation as a “chimera” and translation evaluation as a “daunting task” (Ahsan et al., 2025).

These limitations, however, also highlight the increasing necessity of improving MT tools and their evaluation mechanisms to ensure more reliable global communication through translation. Despite these challenges, enhancing global communication depends on improving the efficiency of MT tools and mitigating their existing limitations. This necessitates both the refinement of text-processing algorithms and the development of more robust quality evaluation metrics, which consequently enable translations to more accurately convey the meaning and contextual nuances of the ST (Khakimov, 2025, p. 2432).

In response to these challenges and demands, a range of translation quality evaluation tools has been launched by different companies. Tools such as BLEU, TER, ChrF and COMET are significant as they improve MT quality and reduce post-editing time by addressing diverse aspects of translation (Phrase, 2024). Among these tools, COMET and its variants have gained particular prominence due to their semantic sensitivity and higher correlation with human judgments. On its official website, COMET is described as “an open-source framework for MT evaluation that can be used for two purposes: to evaluate MT tools with our currently available high-performing metrics and to train and develop new metrics” (Unbabel, n.d.). Rei et al. (2020) state that COMET, as a metric based on contextual models, evaluates translation quality by considering sentence context, the TT, and existing quality assessments. The system functions by encoding the ST, the MT output, and the reference translation separately to produce sentence embeddings that are later combined to produce an overall quality score. Given its semantic-based approach, COMET has quickly become a preferred tool in academic research and industrial MT evaluation. The option of adding reference texts (reference translations) and the capacity to provide an automated but human-like assessments have also contributed to its popularity. Rei et al. (2020) proposed the COMET framework, which harnesses cross-lingual pretrained models and incorporates information from both the ST and the target reference translation, resulting in evaluation models that have a significantly higher correlation with human judgments. Furthermore, COMET (Guerreiro et al., 2024) also marks a significant advancement in fine-tuning-based evaluation frameworks by effectively modeling the relationship between STs, MT output, and reference translations.

This integrated approach enables a more sophisticated analysis that aligns more closely with human judgments of translation quality (Zhang et al., 2025).

In line with these advantages, this study employs xCOMET-XL, an extended and more recent variant of COMET, due to its improved robustness and enhanced correlation with human evaluations across diverse text types. Compared to standard COMET, xCOMET-XL also provides more detailed segment-level analysis and captures quality differences more effectively, making it more suitable for the purposes of this study.

Despite the advanced capabilities of these systems, it is important to note that no automated system is entirely free from limitations. As stated by Joshi et al. (2025), all automatic metrics have their own limitations. Therefore, research in MT evaluation aims to develop metrics or machine learning models capable of producing results that closely align with human judgments. Consequently, research on the topic contribute to the refinement of these tools, which indirectly improves MT quality. Moreover, such research might shed light on the performances of these metrics, providing valuable feedback for their creators.

Method

Research Design

This research employs a mixed-methods approach to evaluate the translation quality of MT from Turkish to English. Johnson et al. (2007, p. 120) state that "mixed methods research is the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study," which aligns with the integration of automated xCOMET-XL evaluation and human evaluation in this study.

Corpus Design

The sample was chosen based on Reiss's (1977/1989) typology, which categorizes texts according to their communicative functions. Reiss's framework was employed for its systematic and functional orientation, enabling a clear differentiation between informative, expressive, and operative texts and supporting the analysis of variation in translation quality across these categories. Within this scope, three text types—informative, expressive, and operative—were included while audio-medial texts were

excluded due to technical limitations and the primary focus of the research. A total of 15 STs in Turkish, consisting of 5 from each text type, were selected and then translated into English to facilitate a comparative analysis. The Turkish–English language pair was selected because Turkish is morphologically complex and syntactically distant from English, which enables a meaningful evaluation of MT performance. The direction from Turkish to English was chosen specifically because AI tools tend to develop more rapidly in English due to its global dominance and widespread use. The dataset includes 360 sentence-level segments, which constitute the unit of analysis for both xCOMET-XL and human evaluation. Each segment is considered an independent evaluation unit to ensure comparability across MT tools and text types.

Translation Procedure

Machine translation was carried out using Google Translate and DeepL. Google Translate was selected for its widespread use and support for Turkish–English translation, while DeepL was chosen for its context-aware translation capabilities, making it particularly suitable for evaluation in academic and professional contexts.

Evaluation Framework

To address the research questions, the study integrates automated evaluation using the xCOMET-XL, (Guerreiro et al., 2024), which is a Multidimensional Quality Metrics (MQM)-based metric, with human translators' evaluations. This hybrid approach combines neural-based scoring with the expertise of two human evaluators to ensure a comprehensive analysis of translation quality. In this framework, xCOMET-XL evaluates translations by comparing MT against both source texts and reference translations using scores between 0 and 1 and identifying MQM-based error types (minor, major, critical) to enable qualitative interpretation. Additionally, each sentence is treated as an independent evaluation unit to enable comparability across MT tools and text types.

The evaluation was conducted in two stages. First, numerical scores were obtained using the AI-powered neural evaluation tool xCOMET-XL (Guerreiro et al., 2024), which computes scores using cross-lingual semantic representations, providing an objective measure of translation quality based on comparative analysis with the STs. xCOMET-XL was selected for its capacity in translation quality assessment across multiple languages and its

ability to provide both source- and reference-aware scoring. This automated assessment laid the foundation for subsequent human evaluation. Second, the translations were assessed by human evaluators. The results were later aggregated to generate text-level scores for MT tool comparison across text types. This aggregation was conducted by calculating the mean of segment scores for each text and tool.

The combination of xCOMET-XL results and human evaluation enabled the cross-validation of translation quality from multiple perspectives, thereby enhancing the depth and originality of the study.

Human Evaluation and Reliability

In order to ensure reliability and avoid anchoring bias, two independent human evaluators, both of whom are expert translators and academics, assessed the dataset separately without access to each other's evaluations and xCOMET-XL's evaluation. Human evaluation was conducted using a structured form aligned with xCOMET-XL outputs. Evaluators rated each segment on a 0–1 scale considering semantic accuracy, fluency, and functional adequacy. A guideline scale was provided to ensure consistency, but final scores relied on expert judgment.

Data Analysis

Overall, the study evaluates the efficiency of neural translation evaluation platforms while analysing MT quality across text types and identifying the text types in which performance is strongest and weakest. By combining automated and human evaluations, this approach seeks to offer a thorough understanding of translation quality.

Human and machine scores were compared across MT tools and text types. Alignment was measured using score differences (|H–X|), and descriptive statistics (means and alignment comparisons) were calculated at both segment and corpus levels. In order to measure alignment, two metrics were used: mean score difference, representing the average difference between xCOMET-XL and human scores, and *winner agreement*, which refers to the percentage of texts in which both methods agree on which MT system performs better.

Findings and Discussion

This section presents the findings of the study and analyzes MT performance alongside the alignment between xCOMET-XL and human evaluator scores. The results highlight differences in translation quality across text types and the relationship between automated and human evaluations.

Overview of Corpus

The corpus consists of 15 STs and 30 machine-translated TTs produced by two MT tools (Google Translate and DeepL), resulting in a total of 45 texts, with each output subsequently segmented for analysis. Table 2 shows the number of segments per text.

Table 2

The number of segments per text

Text	Segments (n)
Informative T1	15
Informative T2	16
Informative T3	17
Informative T4	21
Informative T5	29
Expressive T1	32
Expressive T2	36
Expressive T3	22
Expressive T4	21
Expressive T5	30
Operative T1	22
Operative T2	28
Operative T3	18
Operative T4	31
Operative T5	22
Total	**360**

RQ1 asks whether there is a meaningful difference in MT performance across informative, expressive, and operative text types. Table 3 presents mean xCOMET-XL scores for both MT tools across all fifteen texts, grouped by Reiss type.

Table 3

xCOMET-XL mean scores by text and Reiss type

Text	n	Google Translate Score	DeepL Score	Winner
Informative T1	15	0.968	0.979	DeepL
Informative T2	16	0.926	0.937	DeepL
Informative T3	17	0.978	0.971	Google Translate
Informative T4	21	0.987	0.979	Google Translate
Informative T5	29	0.973	0.975	DeepL
Informative average	**98**	**0.966**	**0.968**	**DeepL**
Expressive T1	32	0.784	0.816	DeepL
Expressive T2	36	0.968	0.977	DeepL
Expressive T3	22	0.884	0.905	DeepL
Expressive T4	21	0.798	0.791	Google Translate
Expressive T5	30	0.830	0.877	DeepL
Expressive average	**141**	**0.853**	**0.873**	**DeepL**
Operative T1	22	0.945	0.957	DeepL
Operative T2	28	0.960	0.963	DeepL
Operative T3	18	0.939	0.960	DeepL
Operative T4	31	0.972	0.974	DeepL
Operative T5	22	0.976	0.979	DeepL
Operative average	**121**	**0.958**	**0.967**	**DeepL**

Variations Across Text Types

Text type has a substantial effect on xCOMET-XL scores. The combined averages are 0.967 for informative texts, 0.963 for operative texts, and 0.863 for expressive texts, indicating that expressive texts score 0.104 lower on the 0–1 scale, which is a meaningful difference in MT evaluation.

This pattern aligns with Reiss's typology. Informative texts use stable, predictable structures that are well suited to MT tools. On the other hand, operative texts are more challenging due to tone and audience sensitivity, but they still remain within a range that MT tools can handle. Expressive texts, however, prioritise authorial voice, figurative language, cultural specificity, and aesthetic form, which are difficult for MT tools to reproduce and are therefore assigned lower scores by xCOMET-XL.

Variation within expressive texts is also notable. *Expressive T2* achieves near-informative scores (0.968/0.977), whereas *Expressive T1* records the lowest results (0.784/0.816), reflecting differences in the linguistic and stylistic demands across expressive texts.

Google Translate vs DeepL Across Text Types

DeepL outperforms Google Translate in 12 of the 15 texts across all three Reiss categories. Google Translate marginally leads only in three texts: *Informative T3, Informative T4* and *Expressive T4*, and in each case the margin is minimal (maximum Google advantage: 0.0082 in *Informative T4*). These differences remain within a very narrow margin and are not substantial. This consistency indicates a general rather than text type-specific advantage for DeepL.

However, the margin of this advantage varies by text type. In informative texts, the gap is negligible (0.9664 vs 0.9682; difference: 0.002). In operative texts, it is small but more consistent (0.9584 vs 0.9665; difference: 0.008). In expressive texts, it is most pronounced (0.8528 vs 0.8732; difference: 0.0204). This pattern suggests that DeepL's advantage increases in parallel with translation complexity, particularly in handling register and stylistic nuance.

Error Span Analysis

Table 4 presents the total number of errors flagged by xCOMET-XL, categorised by severity and Reiss text type. These figures complement the score averages by showing not only overall translation quality, but also the types of errors that drive the scores.

Table 4

Total error spans by severity and Reiss text type

Text Type	Google Critical	Google Major	Google Minor	DeepL Critical	DeepL Major	DeepL Minor
Informative (5 texts)	0	6	45	1	3	44
Expressive (5 texts)	5	84	69	8	68	88
Operative (5 texts)	1	10	55	0	1	65
Total (15 texts)	6	100	169	9	72	197

The error distribution confirms and extends the score findings. Informative texts reveal very few major errors, with 6 for Google Translate and 3 for DeepL across five texts and 98 segments. Operative texts show a moderate number, with 10 for Google, 1 for DeepL. In contrast, expressive texts account for 84 of Google Translate's 100 total major errors and 68 of DeepL's 72. This pattern suggests a clear concentration and shows that the most serious MT failures are predominantly associated with translation of expressive texts. Critical errors follow the same pattern, with 5 of Google's 6 and 8 of DeepL's 9 occurring in expressive texts. This indicates that both informative and operative texts largely avoid errors that would render a seriously misleading or unusable translation.

Alignment Between xCOMET-XL and Human Evaluation

RQ2 asks to what extent xCOMET-XL aligns with human judgements. Table 5 brings together the xCOMET-XL scores and averaged human scores for all fifteen texts, and identifies whether the two methods agree on which system is better.

Winner Agreement

Across the fifteen texts, xCOMET-XL and human evaluation agree on the better-performing system in 10 out of 15 cases (66.7%), indicating that agreement is strongly linked to text type. For operative texts, both methods fully agree (5/5), which consistently identifies DeepL as superior. For expressive texts, agreement occurs in 4 out of 5 cases, with the only disagreement in *Expressive T4*, where xCOMET-XL marginally favours

Google Translate (0.798 vs 0.791), while both human evaluators prefer DeepL. For informative texts, agreement decreases to 1 out of 5 cases.

Table 5

xCOMET-XL and averaged human scores with winner agreement per text

Text	xCOMET-XL Google Translate	xCOMET-XL DeepL	Human Evaluation Google Translate	Human Evaluation DeepL	xComet-XL Winner	Human Evaluation Winner	Agreement
Informative T1	0.968	0.979	0.850	0.750	DeepL	Google	✗
Informative T2	0.926	0.937	0.800	0.900	DeepL	DeepL	✓
Informative T3	0.978	0.971	0.850	0.900	Google	DeepL	✗
Informative T4	0.987	0.979	0.800	0.900	Google	DeepL	✗
Informative T5	0.973	0.975	0.900	0.900	DeepL	Equal	✗
Expressive T1	0.784	0.816	0.700	0.850	DeepL	DeepL	✓
Expressive T2	0.968	0.977	0.600	0.800	DeepL	DeepL	✓
Expressive T3	0.884	0.905	0.750	0.900	DeepL	DeepL	✓
Expressive T4	0.798	0.791	0.650	0.800	Google	DeepL	✗
Expressive T5	0.830	0.877	0.700	0.900	DeepL	DeepL	✓
Operative T1	0.945	0.957	0.800	0.850	DeepL	DeepL	✓
Operative T2	0.960	0.963	0.800	0.900	DeepL	DeepL	✓
Operative T3	0.939	0.960	0.750	0.900	DeepL	DeepL	✓
Operative T4	0.972	0.974	0.750	0.850	DeepL	DeepL	✓
Operative T5	0.976	0.979	0.800	0.900	DeepL	DeepL	✓

The disagreements in informative texts highlight a limitation of automated evaluation. In four cases, xCOMET-XL's decisions are based upon very small score

differences. For instance, 0.9870 vs 0.9788 in the *Informative T4* (0.008 difference) and 0.9782 vs 0.9709 in the *Informative T3* (0.007 difference). These differences remain within a very narrow margin and are not substantial enough to indicate a clear performance difference. Human evaluators do not treat these marginal differences as meaningful and reach different conclusions. The clearest case is the *Informative T1*, where xCOMET-XL favours DeepL (0.979 vs 0.968), but both evaluators prefer Google Translate by pointing to "repetitions and some broken phrases" and "slight structural issues" that reduce clarity. These are holistic, text-level issues that sentence-level evaluation methods like xCOMET-XL do not fully capture.

Score Differences

While winner agreement indicates whether the two methods align in their ranking, mean score difference indicates the distance between their scores. Table 6 presents the average difference between xCOMET-XL and human scores, which are categorised by text type and MT system.

Table 6

Average score difference between xCOMET-XL and human evaluators by text type and MT system

	xCOMET-XL Google Translate avg	**Human Google Translate avg**	**Differences (Google Translate)**	**xCOMET-XL DeepL avg**	**Human DeepL avg**	**Differences (DeepL)**
Informative	0.966	0.840	0.126	0.968	0.870	0.098
Operative	0.958	0.780	0.178	0.967	0.880	0.087
Expressive	0.853	0.680	0.173	0.873	0.850	0.023
Overall	0.926	0.767	0.159	0.936	0.867	0.069

Overall, xCOMET-XL scores are consistently higher than human scores for both systems, with an overall difference of 0.159 for Google Translate and 0.069 for DeepL. This suggests that xCOMET-XL tends to assign higher scores than human evaluators, particularly for Google Translate, where the difference is more than twice that of DeepL on average and within the operative and expressive text categories.

The difference is most pronounced in the expressive category. For DeepL, both methods are closely aligned (0.023 difference), whereas for Google Translate the

difference increases to 0.173, with human evaluators scoring its output about 0.173 points lower. This suggests that xCOMET-XL captures local lexical and phrase-level errors but not the combined effect of limited stylistic variation, literal phrasing, and the loss of authorial voice.

A similar pattern appears in operative texts, where human evaluators rate Google Translate 0.178 and DeepL 0.087 points lower than xCOMET-XL. This points to evaluation differences at the discourse level, where tone, coherence, and persuasive effect play a central role but are not fully captured by sentence-level metrics, which suggests that all numerical comparisons should be interpreted with consideration of the relatively small margins observed in certain categories.

Inter-Rater Agreement

Before drawing conclusions from the human scores, it is important to note that the two evaluators showed a strong level of agreement. Across all thirty ratings, 63.3% of scores were identical and 83.3% differed by no more than 0.1 on the 0–1 scale. The average difference between the evaluators was 0.040, which is less than half of the overall gap between human and xCOMET-XL scores. This level of agreement confirms that the human scores are consistent and that the differences observed between human and xCOMET-XL evaluations reflect genuine methodological differences rather than inconsistencies among the human evaluators. Human evaluators tended to use rounded score values (e.g., 0.750, 0.800), reflecting an intuitive anchored rating approach rather than fine-grained continuous scaling.

Text Type as the Primary Driver of MT Quality

The xCOMET-XL results provide a clear answer to the RQ1, showing a meaningful difference in MT performance across the three text types. The difference is primarily driven by the expressive category. With a combined average score of 0.863, approximately 10 percentage points lower than informative and operative texts, expressive texts represent a distinct challenge for both MT tools. This aligns with Reiss's theoretical framework, as expressive texts prioritise aesthetic and stylistic form over content transmission, a dimension that current MT tools have difficulty in handling.

In contrast, differences between the non-expressive categories are relatively small and largely within the margin of normal variation. Both informative and operative texts show high and relatively stable scores across all texts in each category. The results suggest that across these text types that include texts such as technical documents, administrative

texts, and promotional materials, both Google Translate and DeepL reach a broadly adequate level of quality. However, they still consistently fall short of the quality expected from a professional translator, as confirmed by the human evaluation.

Alignment Between xCOMET-XL and Human Evaluation

RQ2 examines to what extent xCOMET-XL aligns with human judgements. The findings suggest that alignment depends strongly on text type. For operative texts, alignment is high, as both methods consistently identify the same system as superior in all five cases. This is also reflected in the evaluators' qualitative comments, which point to DeepL's stronger promotional tone, more natural phrasing, and better brand sensitivity, and these observations closely correspond to the patterns in xCOMET's scores. For expressive texts, alignment is moderate, as the two methods agree on four of five winner decisions. However, xCOMET-XL substantially underestimates the gap between Google Translate and DeepL as perceived by professional translators, particularly in its handling of Google's expressive text translation. For informative texts, alignment is low, as the score differences driving xCOMET-XL's conclusions are too small to be noticeable at the holistic level, and human evaluators' text-level impressions often signal a different direction.

These findings have a significant implication for translation quality. xCOMET-XL and human evaluation are not interchangeable methods, but rather complementary ones. xCOMET-XL provides fine-grained, reproducible, and scalable sentence-level analysis across large corpora, by identifying problematic segments and classifying error severity. Human evaluation, on the other hand, captures what xCOMET-XL cannot, including overall text quality, discourse-level coherence, tonal and stylistic adequacy, and the cumulative effect of translation decisions across a full text. A mixed-methods approach, as used in this study, therefore provides a more complete picture than either method alone.

Conclusion

This study evaluated 360 segments from fifteen Turkish texts translated into English by Google Translate and DeepL. Adopting Reiss's typology, the texts were equally distributed across informative, expressive, and operative categories. Quality assessment was conducted using xCOMET-XL and human evaluation.

The study shows that text type plays an important role in MT quality. Expressive texts consistently present greater difficulty for both systems, while informative and

operative texts yield more stable and higher-quality outputs. Overall, DeepL performs slightly better than Google Translate, with the difference becoming more visible in expressive texts. These results are consistent with Reiss's functional typology, which distinguishes text types according to their communicative priorities.

The comparison between xCOMET-XL and human evaluation indicates a generally similar pattern in identifying the stronger system, especially in operative texts, where alignment is highest. However, differences between the two approaches become more noticeable in expressive texts, where human evaluators tend to be more sensitive to stylistic and pragmatic aspects that are less visible at the segment level. In informative texts, both methods show the lowest alignment because the subtle numerical variations that are prioritized by xCOMET-XL do not always correspond with the holistic judgments of human evaluators.

Overall, the results suggest that automatic evaluation metrics such as xCOMET-XL are effective in capturing local and segment-level translation quality, particularly in large-scale comparisons. However, human evaluation remains necessary for capturing broader textual and stylistic dimensions of translation quality, especially in expressive texts where tone, style, and interpretive meaning play a central role. This highlights the complementary nature of automated and human-based evaluation approaches.

The study also shows that such metrics provide a useful and scalable alternative for translation quality assessment. At the same time, they do not fully replace human judgment, particularly when it comes to holistic reading and contextual interpretation of texts. Therefore, combining both approaches offers a more balanced and reliable framework for evaluating machine translation quality.

This study examined the applicability and effectiveness of the neural, AI-based evaluation metric xCOMET-XL in the assessment of MT quality, with a particular focus on Turkish-to-English translations that are categorized according to Reiss's text typology. By using the xCOMET-XL, the research conducted a systematic, semantic-aware evaluation of MT. By comparing xCOMET-XL scores with human evaluations, the study examined the extent to which automated, neural-based translation assessment aligns with human assessment. These findings highlight the growing importance of AI-based evaluation tools as complementary resources to traditional evaluation methods by offering insights into

how they can contribute to more objective, adaptable, and context-sensitive assessments of translation quality.

There are limitations of this study. Firstly, the analysis was restricted to Turkish–English language pair and a comparatively small corpus of texts categorized according to Reiss's typology. This might limit the generalizability of the findings. Secondly, while xCOMET-XL offers a robust semantic-based evaluation, it remains an automated metric and may not fully capture all the nuances of human judgment, such as cultural, pragmatic, or stylistic nuances. Thirdly, the study examined only two widely used MT tools, Google Translate and DeepL, and results may differ with other tools.

On the other hand, the results of the study might pave the way for further academic studies. Future research could address these limitations by extending the scope to additional language pairs and larger corpora. Comparative analyses incorporating multiple neural evaluation metrics alongside human assessments could further clarify the relationship between automated and human judgments. Finally, longitudinal studies on the integration of these metrics into professional translation workflows could also provide insights for both research and practice.

References

Ahsan, A., Mujadia, V., Mishra, P., Bhaskar, Y., & Sharma, D. M. (2025). *Crosslingual optimized metric for translation assessment of Indian languages*. arXiv preprint. https://doi.org/10.48550/arXiv.2509.17667

Chen, H., & Zhang, X. (2020). Application of Reiss's text typology in translation: A case analysis of Franklin D. Roosevelt's fourth inaugural. *East African Scholars Journal of Education, Humanities and Literature, 3*(2), 34-37. https://doi.org/10.36349/EASJEHL.2020.v03i02.005

Çetiner, C. (2021). Sustainability of translation as a profession: Changing roles of translators in light of the developments in machine translation systems. RumeliDE Dil ve Edebiyat Araştırmaları Dergisi, (Ö9), 575-586. https://doi.org/10.29000/rumelide.985014

Delibaş, H. (2023). Investigating the cross-cultural impact: An analysis of Turkish translations of Common European Framework of Reference (CEFR) through Reiss's text typology. *Ondokuz Mayıs University Journal of Faculty of Education, 42*(2), 1017–1034. https://doi.org/10.7822/omuefd.1322430

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error

detection. *Transactions of the Association for Computational Linguistics, 12*, 979-995. https://doi.org/10.1162/tacl_a_00683

Hutchins, J. (1995). Reflections on the history and present state of MT. In Proceedings of Machine Translation Summit V (pp. 431–435). Luxembourg.

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*. Academic Press.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133. https://doi.org/10.1177/1558689806298224

Joshi, N., Katyayan, P., & Arora, P. (2025). *Trainable reference-based evaluation metric for identifying quality of English-Gujarati machine translation system*. arXiv preprint. https://arxiv.org/abs/2510.05113

Khakimov, M. (2025). Metric for evaluating mathematical models natural language words for machine translation. *International Journal of Innovative Research and Scientific Studies, 8*(6), 2430-2447. https://doi.org/10.53894/ijirss.v8i6.10125

Korkmaz, İ. (2019). Makine çevirisinin kısa tarihçesi. *International Journal of Social Humanities Sciences Research (JSHSR), 6*, 155-166. https://doi.org/10.26450/jshsr.1030

Lin, Y. (2023). Analysis of machine translation based on neural networks. Proceedings of the 3rd International Conference on Signal Processing and Machine Learning (Vol. 5, Article 20230547). Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/5/20230547

Metin Tekin, B. (2025). Embracing technology: Perspectives of translation and interpreting students on machine translation (MT) in Turkish universities. *Istanbul University Journal of Translation Studies,* (23), 159-177. https://doi.org/10.26650/iujts.2025.1634780

Moralı, M., & Şahin, M. (2025). Research on translation and interpreting technologies in Türkiye through the lens of translation and interpreting studies: A bibliometric study. Söylem Filoloji Dergisi, Çeviribilim Özel Sayısı II, 202-218. https://doi.org/10.29110/soylemdergi.1600959

Munday, J. (2016). *Introducing translation studies: Theories and applications* (4th ed.). Routledge.

Orrego-Carmona, D. (2020). Machine translation in everyone's hands: Adoption and changes among general users of MT. In M. O'Hagan (Ed.), The Routledge handbook of translation and technology (pp. 323–340). Routledge.

Palanichamy, N., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience, 27*(10), Article 110878. https://doi.org/10.1016/j.isci.2024.110878

Perrella, S., Proietti, L., Huguet Cabot, P.-L., Barba, E., & Navigli, R. (2024). Beyond correlation: Interpretable evaluation of machine translation metrics. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 20689–20714). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.1152

Phrase. (2024). Machine translation. Phrase Support. https://support.phrase.com/hc/en-us/sections/5709646763420-Machine-Translation

Pombal, J., Guerreiro, N. M., Rei, R., & Martins, A. F. T. (2025). Adding chocolate to mint: Mitigating metric interference in machine translation. *Transactions of the Association for Computational Linguistics, 13*, 1319-1339. https://doi.org/10.1162/TACL.a.37

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Reiss, K. (1989). Text types, translation types and translation assessment (A. Chesterman, Trans.). In A. Chesterman (Ed.), Readings in translation theory (pp. 105-115). Oy Finn Lectura. (Original work published 1977)

Reiss, K., & Vermeer, H. J. (2014). *Towards a general theory of translational action: Skopos theory explained* (C. Nord, Trans.). Routledge.

Şahin, M., & Duman, D. (2013). Multilingual chat through machine translation: A case of English-Russian. *Meta: Translators' Journal, 58*(2), 397-410. https://doi.org/10.7202/1024180ar

Uğurlu, Y. G., & Dolmacı, M. (2020). Metin türleri. In E. Uluşahin & E. Eriş (Eds.), Çeviri bilimi üzerine kuramsal çalışmalar 1 (pp. 83–104). Nobel Akademik Yayıncılık.

Unbabel. (n.d.). COMET: Cross lingual Optimized Metric for Evaluation of Translation. https://unbabel.github.io/COMET/html/index.html

Zhang, R., Zhao, W., Macken, L., & Eger, S. (2025). *TRANSPROQA: An LLM-based literary translation evaluation metric with professional question answering*. arXiv preprint. https://doi.org/10.48550/arXiv.2505.05423

Reviewers:
1. Anonymous
2. Anonymous

Handling Editor:
Boris Naimushin, PhD
New Bulgarian University