Computer Science and Information Systems 23(2):827–860 https://doi.org/10.2298/CSIS251020021K

A Pilot Study of Multi-Method Evaluation of Machine Translation in Macedonian

Jana Kuzmanova[1] and Katerina Zdravkova[2]

[1] Faculty of Computer Science and Engineering
jana.kuzmanova@finki.ukim.mk
[2] Faculty of Computer Science and Engineering
katerina.zdravkova@finki.ukim.mk (corresponding author)

Abstract. This pilot study offers a linguistic evaluation of six machine translation systems: GPT-4o, GPT-5, Gemini 2.5 Flash, Google Translate, Microsoft Translator, and NLLB-600M applied to the translation of a short excerpt of Orwell's "1984" into Macedonian. The analysis consisted of three interconnected experiments: manual annotation of translation errors and comparison with human output, evaluation using eight popular MT metrics, and sentence-level similarity analysis via cosine similarity, Jaccard similarity, and Levenshtein distance. Manual annotation revealed that stylistic errors (48.47%) and linguistic errors (34.54%) were the most common. The LLMs outperformed other systems, particularly GPT-5, while NLLB-600M performed poorly, often introducing incomprehensible sentences or non-existent words. Metrics-based evaluation showed that lexical metrics sometimes penalized fluent and accurate translations that deviated from the reference. Sentence similarity analysis confirmed that accurate translations were more consistent, while wrong–wrong sentence pairs were more divergent, especially in Levenshtein scores. The findings underscore the importance of combining manual and metric-based evaluation to fully understand MT quality, particularly in low-resource settings.

Keywords: Machine Translation, Manual Evaluation and Annotation, Linguistic Similarity, Low Resource Language

1. Introduction

Bibliophiles enjoy reading novels in the original language when they are fluent in it, but they also appreciate high-quality translations, particularly when they capture the spirit of the original. In some cases, translation can even surpass the source text, especially when it brings creativity and cultural insight to the work [1]. An excellent translation is both an expertise and an impressive reproductive art [2]. It not only conveys the meaning and nuances of the original text, but also flows naturally, reflecting the translator's deep understanding of the culture, history, and linguistic subtleties of the source [3,4].

Unfortunately, achieving high-quality translation is not always an easy activity [5]. Many readers have, at some point, given up reading an otherwise excellent literary work due to poor translations [6]. A major turning point in improving translation quality has been the development of machine translation (MT) tools. These tools can significantly facilitate the process of converting a text from one language to another, while striving to preserve the meaning of the original.

Recently, MT systems have improved rapidly, particularly due to their shift to neural machine translation (NMT). NMT is mainly based on deep learning (DL) and large training datasets [7]. DL involves training artificial neural networks with many layers, enabling the models to automatically learn hierarchical features and complex representations from vast amounts of data.

New MT systems have become an essential support for readers who attempt to understand texts written in languages entirely unfamiliar to them. Yet, a critical question remains: can MT reach the quality of a skilled human translator? This study seeks to explore this question by comparing a human translation of Orwell's "1984" into Macedonian with translations produced by six leading MT systems.

Building on this inquiry, our pilot research investigates how the following MT systems: ChatGPT (GPT-4o and GPT-5), Gemini (2.5 Flash), Google Translate, Microsoft Translator and NLLB-600M handle literary language, by focusing on a small excerpt from a well-known and widely translated novel "1984" by George Orwell. This novel was selected not only for its literary and political significance, but also because it forms the basis of the MULTEXT project [8]. Moreover, as part of the MULTEXT-East project [9], a professionally produced human translation of "1984" has been part of speech (POS) [10] and morpho-syntactically annotated and mutually aligned multiple languages, including Macedonian [11,12]. These resources allowed us to conduct a detailed manual evaluation of machine-translated outputs against a trusted human reference. The Macedonian reference translation was used with written permission from the publisher for academic research purposes within the MULTEXT-East framework. In accordance with these conditions, the study does not redistribute the full text and includes only short illustrative examples for analysis.

For each machine-generated sentence, we assessed whether the translation was accurate or contained errors. By an accurate translation, we refer to one that faithfully preserves the meaning, intent, and relevant nuances of the source text in the target language, without distortion, omission, or unwarranted addition [13,14].

The identified translation errors were grouped into four clusters, reflecting both the linguistic levels and key dimensions of translation quality:

1. Linguistic accuracy, referring to grammatical accuracy regardless of the meaning;
2. Stylistic errors, addressing issues such as inappropriate word choice, tone, or awkward phrasing;
3. Lexical and semantic errors, evaluating how well the intended meaning of the source text is preserved;
4. Fluency and naturalness, evaluating the readability and native-like quality of the translation.

Inspired by Tolstoy's famous opening line from Anna Karenina "All happy families are alike; each unhappy family is unhappy in its own way" [15], we adapted the sentiment to reflect our findings: "Every accurate translation is alike; every wrong translation is inaccurate in its own way." This phrase emerged from our manual evaluation of machine translation results, where we noticed that accurate translations tended to converge in form and quality, while inaccurate ones differed significantly in their errors.

This motivated us to explore sentence-level similarity analysis using three different methods:

1. Lexical/surface similarity (BLEU, TER, chrF)
2. Embedding-based similarity (BERTScore, Sentence-BERT, cosine similarity)
3. Edit-based distance (Levenshtein)

We further examined the variance in similarity scores, hypothesizing that accurate translations would be more consistent and clustered closely, while erroneous outputs would show greater divergence. In addition, we investigated how specific error types relate to automated metric scores and assessed the extent to which metric-based rankings align with human judgments. We also evaluated the literalness of machine translation, comparing it to both automatic quality estimates and human evaluations.

Finally, we visualized these results to reveal the underlying patterns and clarify the relationships between translation quality, error characteristics, and translation similarity.

2. Review of Recent Studies that Evaluate MT Systems

Several studies have evaluated and compared commercial and publicly available MT systems. As an example,[16] performed an automatic evaluation of the systems submitted to the WMT22 General Machine Translation task, employing chrF (Character n-gram F-score) [17], BLEU (Bilingual Evaluation Understudy) [18], and COMET (Crosslingual Optimized Metric for Evaluation of Translation) [19] metrics. A total of 185 systems were evaluated across 11 language pairs and 21 translation directions. These pairs included Czech – English, Czech – Ukrainian, French – German, German – English, English – Chinese, English – Croatian, English – Japanese, English – Russian, English – Ukrainian, as well as more distant and low-resource combinations such as Russian – Yakut and English – Livonian. The systems were ranked using the same three metrics after which a statistical significance testing was performed. Additional statistics were computed for each system, including: the number of sentences identical to the reference; the number of sentences that differed between the MT output and the reference, despite being the same in the source; and sentence normalization to assess how punctuation changes affect BLEU and COMET scores. The results revealed significant discrepancies between BLEU and COMET scores for some systems. It is worth mentioning that COMET did not agree with the top-ranked BLEU and chrF systems in 11 out of 21 translation directions. BLEU and chrF gave identical rankings for only 5 language pairs, although their rankings were generally similar. Statistical significance testing indicated that a difference of 0.9 BLEU points was considered meaningful.

Similarly,[20] found that a difference of 2 to 3 BLEU points is statistically significant. However,[21] demonstrated that achieving a pairwise accuracy of 85%, defined as the proportion of system pairs for which the automatic metric ranks systems in the same order as human evaluators, requires a minimum difference of 3.35 BLEU points between system scores. It is important to note that BLEU never reaches 90% pairwise accuracy, even when the score differences are larger. Because different metrics operate on different scales, the score improvement required to reach a given pairwise accuracy threshold can vary significantly. For example, chrF can achieve 90% pairwise accuracy with an improvement of 3.05 points, while COMET can reach 95% accuracy with an increase of only 1.18 points for some models. In contrast, for lexical metrics like BLEU and chrF, the required score difference is typically larger, and pairwise accuracy decreases when comparing unrelated systems. Additionally, Kocmi et al. (2021), who first proposed the pairwise accuracy

metric, noticed that in cases where human evaluations disagree with BLEU rankings, the median BLEU score difference between systems is 1.3 points.

[22] compared two versions of ChatGPT (3.5 and 4), Google Translate, and DeepL using BLEU, chrF, and TER (Translation Edit Rate) [23] for all translation combinations between English, German, Chinese, Japanese, and Romanian. They also computed word prediction accuracy based on frequency, grouped scores by sentence length, and conducted a human evaluation of the number of errors each system produced. Human evaluation revealed that GPT-4 was the best-performing model, despite having lower BLEU scores than Google Translate.

[24] conducted a study that analyzed the quality of literary translations from English into Dutch. A 500-word short story was translated using three open-access NMT engines: DeepL, Systran, and Google NMT. The target translations were evaluated for accuracy, fluency, and style. The evaluation was performed by human reviewers and through automated metrics, including BLEU scores, supplemented by a Dutch literariness algorithm. The study also highlighted specific words and phrases in the source text that demanded careful handling across the assessed categories. Consistent with the findings of Jiao et al., a significant discrepancy was observed between human judgments and BLEU scores.

[25] also evaluated generative pre-trained transformer (GPT) models using a combination of automatic metrics and human judgment. The evaluation covers translations between English and a range of other languages, including French, German, Czech, Icelandic, Chinese, Japanese, Russian, Ukrainian, and Hausa. The evaluation metrics included two versions of COMET: COMET-22 and the reference-free COMETkiwi, as well as BLEU and chrF scores. Translations were assessed at both the sentence and document level. The results were grouped according to prompt design and selection, language resource level (high vs. low), and domain. Consistent with previous studies, human evaluation indicated that traditional lexical metrics do not fully capture improvements in translation quality, whereas COMET exhibited a stronger correlation. The study further analyzed distinctive characteristics of GPT-generated translations compared to those produced by NMT systems, focusing on dimensions such as non-monotonicity, fluency, punctuation insertion, and dropped or inserted content.

Other studies also examine the distinguishing characteristics of MT systems. Focusing on German-to-English translation,[26] investigate translation artifacts that differentiate human, NMT, and LLM-generated outputs. They classify sentences as either original or translated and employ explainability methods to identify the features contributing to these classifications. Although BLEU and COMET scores are reported, no correlation is found between classification accuracy and translation quality. The authors apply leave-one-out and integrated gradient techniques to analyze both feature overlap and feature frequency. Feature overlap is assessed by identifying the most influential features and calculating Jaccard similarity scores to compare them across systems. Feature frequency is analyzed through POS distributions grouped by sentence length. The findings indicate that LLM translations exhibit artifacts more similar to those observed in human translations than in NMT outputs, although notable differences remain.

A similar comparison between machine and human translations is conducted by[27]. They perform word-based and arc-based analyzes of English–Chinese, English–French, and English–German translations. They identify POS patterns in the source text and examine the corresponding target-side patterns, focusing on distribution, conditional entropy,

and convergence. These metrics are aggregated according to frequency. Their findings suggest that translation quality and structural divergence are not directly related; however, human translations consistently exhibit greater variability and divergence than machine-generated ones. [28], meanwhile, investigate the literalness of translations produced by GPT models in comparison to NMT systems. Literalness is measured by the number of unaligned source words and by non-monotonicity, defined as deviations in word order. Their analysis spans English – German and English – Russian translations in both directions and is supported by human evaluation. The results show that GPT translations contain more unaligned source words overall, while higher non-monotonicity is observed only when translating from English into other languages.

Given the limitations of automated metrics noted above, MT evaluation is typically carried out by human annotators. However, the WMT24 Translation Task employed a preliminary system ranking based on automated metrics, due to the high volume of submissions [29]. Two top-performing metrics from the WMT23 Metrics Task were used: MetricX-23-XL and COMETkiwi-DA-XL. The latter, a reference-free metric, was included specifically to mitigate reference bias. The evaluation covered a wide range of language pairs, including translations from English into Czech, German, Spanish, Hindi, Icelandic, Chinese, Japanese, Russian, and Ukrainian, as well as translations between Czech and Ukrainian and between Japanese and Chinese.

However, human evaluation is also far from straightforward. Scoring practices and evaluator expertise can significantly influence results. Direct Assessment (DA), i.e., the practice of assigning scores on a 0–100 scale, can be unreliable; however, this limitation can be partially mitigated by supplementing the scale with descriptive labels. A more comprehensive evaluation framework is the Multidimensional Quality Metrics (MQM) system [30], which focuses on error annotation. Freitag et al. (2021) [31] proposed an MQM-based scheme that incorporates categories for accuracy, fluency, style, and locale, along with error severities and a method for computing aggregate scores.

[32] adapted MQM for Slavic languages by incorporating agreement features such as person, number, gender, and case into the core tagset. Both studies found MQM to yield higher inter-annotator agreement than DA or Scalar Quality Metrics (SQM). However, MQM annotation is time-consuming and costly. An alternative approach, error-span annotation [33], focuses only on marking problematic segments and their severity. This method achieves strong alignment with MQM rankings, demonstrates higher inter-annotator agreement, and is more efficient to conduct.

The human evaluation of literary translation is also explored by[34], with a focus on the performance of large language models (LLMs). The study evaluates translations involving English, Polish, Russian, Czech, French, Japanese, and Chinese as source languages, with English, Japanese, and Polish as targets. Evaluations are conducted at both the sentence and paragraph levels. Annotators are presented with two MT outputs (one generated at the sentence level, the other at the paragraph level) and are asked to mark error spans from categories including mistranslation, untranslated segments, grammar, inconsistency, register, and formatting. They are also required to select their preferred version and provide freeform justifications. To investigate potential data memorization, the authors assessed whether masked named entities can be predicted by the model, which was generally not the case. While paragraph-level translation yields improvements over sentence-level output, critical errors may still persist.

An alternative method for MT evaluation involves the use of test suites, as demonstrated by[35]. Their approach employs a curated set of linguistic phenomena alongside regular expression rules to automatically detect correct and incorrect translations, supplemented by manual evaluation where necessary. The study focuses on translations from English into German and Russian. Accuracy is calculated as the proportion of correct translations. Statistical significance is also assessed. The results are aggregated by linguistic category and phenomenon. For both languages, case agreement, prepositional multi-word expressions, and date formatting are among the most accurately translated categories. Conversely, idioms and semantic role assignments remain challenging. Additionally, German translations struggle with rare verb tenses, while Russian systems face difficulties with compounds and verbal multi-word expressions.

The above-mentioned papers, particularly recent WMT shared-task evaluations [36], illustrate substantial variation in MT performance across language pairs, depending on resource availability. While some pairs involving traditionally lower-resource languages (e.g., Czech–Ukrainian) achieve relatively strong results with human evaluation, likely aided by linguistic relatedness, other pairs such as English–Icelandic exhibit markedly lower performance compared to high-resource benchmarks. These results highlight that translation quality in lower-resource settings is influenced not only by data size but also by factors such as language similarity and available training corpora.

Recent research reviewed in this section highlights both the potential and the limitations of current MT systems, particularly in literary and multilingual contexts. While LLMs and NMT systems show notable progress, traditional lexical metrics often fail to capture nuanced improvements in quality. Human evaluation remains essential, although methods vary in reliability and cost. Emerging alternatives such as error-span annotation and test suites offer promising, more scalable solutions. Overall, fine-grained, linguistically informed evaluation remains crucial for assessing translation quality across diverse languages and text types.

Building on the insights and challenges identified in the literature, the following section presents our methodology and methods, which draw upon and extend the approaches reviewed to provide a comprehensive evaluation framework for MT quality.

3. Methodologies and Methods

Our study was initially designed to evaluate the quality of five MT systems, selected for their widespread use and demonstrated proficiency in the Macedonian language. Based on an extensive review process, these five systems were confirmed as the focus of our initial evaluation: GPT-4o [3], Gemini 2.5 Flash [4], Google Translate [5], Microsoft Translator [6], and NLLB-600M [37]. Unfortunately, one of the most popular NMT services in Europe, DeepL [7], still does not support the Macedonian language. However, in early August 2025, the highly anticipated release of GPT-5 [8] was announced, generating significant interest

[3] https://chatgpt.com/
[4] https://gemini.google.com/
[5] https://translate.google.com/
[6] https://translator.microsoft.com/
[7] https://www.deepl.com/en/translator
[8] https://chatgpt.com/

in both academic and technological communities. Recognizing its potential impact, we promptly expanded the scope of our research to include GPT-5, ensuring that our study reflects the most up-to-date trends in the field.

All MT systems were used through their respective web interfaces as available in August 2025, with the exception of NLLB, which was used through HuggingFace. The systems were used with their default parameters. The prompt used for GPT-4o, GPT-5, and Gemini was a simple zero-shot prompt as follows: "Translate the following sentences from English into Macedonian. Return one sentence per line:", followed by the sentences, each one in a new line.

To avoid exceeding the free translation limits of these MT systems, we focused on the first 100 sentences of Orwell's novel. An additional reason for selecting these specific sentences is that, without exception, they are aligned 1:1 in the human translation from English to Macedonian. For the LLM prompts, the sentences were chunked in two batches of 44, and a remaining batch of 22 sentences.

The linguistic evaluation consisted of two phases: manual annotation of machine translations and sentence-level similarity assessment. Each phase is explained in more detail in the following subsections.

3.1. Manual Annotation of Translation Errors

Manual error annotation and classification were performed by a single expert annotator, who was in continuous consultation with two senior linguists from the Institute of Macedonian Language, who acted as domain experts and external reviewers. The evaluation was further supported by systematic cross-linguistic comparison with existing human translations in Serbian, Croatian, and Bulgarian available within the MULTEXT-East framework [9].

All translations were compiled in a Google Docs spreadsheet consisting of six worksheets, one dedicated to each MT system. Each worksheet included five columns: the source English sentence, the corresponding MT output, the official human translation into Macedonian, a column for mnemonic labels identifying observed errors, and a section for detailed comments explaining those errors.

The corpus of 100 sentences was examined in detail, and all observed errors were initially recorded and labelled in the comments. These errors were then reviewed, filtered, and consolidated through an iterative process, resulting in a final set of 39 distinct error types that were characteristic of all evaluated translations, many of which were language-specific (LS).

To facilitate evaluation, the identified errors were categorized into four main clusters, as outlined in the introduction: Linguistic inaccuracy, Stylistic errors, Lexical and semantic errors, and Fluency and naturalness. Each cluster was further divided into three subgroups, based on shared characteristics and recurring patterns observed across the corpus.

The resulting taxonomy was established through a systematic comparison of existing error-classification frameworks, with particular attention to their cognitive and systemic characteristics. Although this comparative analysis informed the structure of the taxonomy, it was developed independently and was not directly validated against the MQM framework.

According to Polio (1997) [38], linguistic accuracy refers to the degree to which the rules of a language are correctly applied, which in the case of MT includes grammar, vo-

cabulary, and syntax. We determined that the corresponding cluster of errors encompasses the following broad categories: morphological disagreement, verb and tense accuracy, and syntactic structure.

Morphological disagreement refers to mismatches in grammatical categories between words that are expected to agree [39]. Within this category, we identified four recurring error types: gender disagreement, incorrect plural formation, and number disagreement.

Verb and tense accuracy relate to the correct use of verb forms to ensure clarity and temporal consistency in translation [40]. Irregular verb form refers to the incorrect use of the active voice when the passive voice would be more appropriate. Past tense selection in Macedonian language depends on whether the speaker directly witnessed or participated in the event or is simply reporting it indirectly [41]. This subcategory includes not only tense and verb form errors, but also person disagreement, which refers to a mismatch between the subject and the verb's inflectional marking.

Syntactic structure is a broad term encompassing grammatical rules, word order, and the construction of well-formed sentences [42]. We limited this group to four specific error types: missing accusative clitic, word order errors, wrong definiteness, and wrong preposition. The key reason why missing accusative clitics form part of this classification is that they are usually associated with the incorrect use of prepositions, whose function in sentence structures is crucial for its linguistic accuracy.

Stylistic errors are writing choices that impair clarity or appropriateness, even if they are not strictly ungrammatical [43]. These errors make the text harder to read or less natural, although the intended meaning usually remains understandable.

The use of calques, literal translations, foreign language insertions, and pleonasms falls under the subgroup of literalism. These errors involve transferring structures or redundant expressions from the source language, resulting in unnatural or inaccurate target-language output [44].

Punctuation and formatting errors include missing and wrong punctuation, unnecessary use of quotation marks, and use of lowercase letters in formal expressions. These errors occur when the translator does not follow the target language's conventions or fails to appropriately adapt the source text's style [45]. They can significantly impair readability, alter meaning, and undermine the professionalism of the translated text.

Finally, we include spelling errors, misused synonyms, and incorrect phrasing under the subgroup of vocabulary inaccuracies. Vocabulary inaccuracy refers to the incorrect use of words in terms of spelling, form, or contextual meaning. Such errors can lead to miscommunication, obscure ideas, and indicate an imprecise understanding of the target vocabulary [46].

Translation errors sometimes stem from a lack of knowledge, carelessness, or insufficient competence. MT systems are prone to such errors, as they are often embedded in the training data they rely on [47]. These types of errors are grouped in the third cluster named Lexical and semantic errors. It comprises three broad subcategories: terminological errors, untranslatability, and contextual misinterpretation.

Terminological errors involve the incorrect use of terms specific to specialized domains [48]. When such terms are mistranslated, the result is a word or phrase with an inaccurate or inappropriate meaning in the target language. In our experiment, these errors appeared in the form of incorrect adverbs, medical term, misinterpretation of Orwell's Newspeak, and improper word usage.

Untranslatability refers to the difficulty or even impossibility of conveying the exact meaning of a source text in the target language [49]. In our case, this subcategory includes inconsistent translations of the same word or phrase, untranslated terms that alter the original meaning, and instances where transliteration was used in place of accurate translation.

Contextual misinterpretation occurs when the translator fails to accurately interpret the cultural, situational, or linguistic context of the source text [3]. This can result in translations that are inaccurate, misleading, or even inappropriate. In the MT systems we analyzed, such errors typically involved issues with sentence structure, such as incorrect POS or mistranslations caused by a lack of information from preceding context.

The last group deals with errors related to fluency and naturalness in translation. These criteria assess how well the translated text can be understood in the target language. Fluency refers to grammatical correctness and the smooth flow of a sentence [50], while naturalness encompasses broader cultural and idiomatic appropriateness, ensuring that the translation reads as if it were originally written in the target language, rather than as a literal translation from another language [50].

Grammatical correctness is often compromised by omissions. In the MT systems we examined, the omissions included missing conjunctions, prepositions, verbs, and entire phrases. Notably, many observed errors also involved the invention of non-existent words.

We grouped such errors alongside untranslated phrases and words into a subcategory labeled non-existent or untranslated items. The final subcategory, unnatural expressions, included meaningless translations, overly descriptive phrasing, and misuse of conjunctions, all of which disrupt the natural flow and tone of the target text.

After manually evaluating and annotating all the six MT systems, we compared machine and human translations by estimating sentence similarity. In parallel, the five best-performing MT were also compared with one another using sentence similarity techniques. These techniques are introduced in the following subsection.

3.2. Assessing Translation Quality with Automated Metrics

To assess the quality of MT outputs produced by various systems, we employed a comprehensive set of both lexical and learned evaluation metrics. Lexical metrics operate primarily at the surface level, focusing on the overlap between the generated translation and a human reference. Among these, we included BLEU [18], a precision-oriented metric that measures the n-gram overlap between system output and reference translations. BLEU computes a geometric mean of modified n-gram precisions (typically up to 4-grams) and applies a brevity penalty to penalize overly short outputs.

In addition to BLEU, we used chrF [17], which calculates F-scores based on character-level n-gram matches. chrF has been shown to correlate better with human judgments than BLEU for morphologically rich languages and translations involving high lexical variation, as it captures finer-grained patterns of correspondence that word-level metrics may miss.

Another lexical metric we used was TER [23], which quantifies the number of edits: insertions, deletions, substitutions, and shifts needed to convert a system translation to reference. For these three metrics, their sacrebleu implementation [51] was used.

We also considered METEOR (Metric for Evaluation of Translation with Explicit Ordering [52]), which aligns words using not only exact matches, but also stems, synonyms, and paraphrases, and incorporates a fragmentation penalty to account for word order.

Beyond surface-level lexical evaluation, we also integrated a set of embedding-based and learned metrics that leverage deep neural models to estimate translation quality with a stronger focus on meaning and contextual understanding. Among these, we used COMET [19], a regression-based metric trained on human quality judgments. We also employed XCOMET-XL [53], a recent extension of COMET that leverages larger pretrained language models, showing improved correlation with human evaluation, especially for high-resource language pairs and longer documents.

To further complement our evaluation suite, we included COMETkiwi [54], a reference-free variant of COMET. This metric estimates the quality of a translation using only the source and the system output, without requiring a human reference.

Finally, we used BERTscore [55], which computes similarity between token embeddings derived from BERT models, aligning words in the hypothesis and reference based on cosine similarity. BERTscore captures semantic similarity more effectively than traditional lexical metrics and has been shown to correlate well with human judgments at both the sentence and the system levels.

3.3. Sentence Similarities Between Different Machine Translation Systems

In addition to evaluating translation quality through standard metrics, we also investigated the pairwise sentence-level similarities between the outputs of different MT systems. The goal was to check if systems fail in the same or different ways when they make errors, as well as to profile the overall diversity between the machine translations of different systems.

To quantify these similarities, we employed three complementary metrics: cosine similarity based on multilingual MPNet sentence embeddings (Masked and Permuted Pretraining for Language Understanding) [56], Jaccard similarity, and Levenshtein distance.

The cosine similarity metric was computed using sentence embeddings generated by a multilingual variant of MPNet, a transformer-based model known for its effectiveness in capturing contextual semantic information. MPNet combines masked language modelling and permuted language modelling for improved sentence representation. A higher cosine similarity indicates that two translations share similar meanings, even if they use different words or structures, making this metric especially useful for identifying cases where different systems arrive at semantically equivalent translations via different surface forms.

In contrast, Jaccard similarity operates on a set-based lexical level, measuring the overlap between the unique tokens (e.g., words or character n-grams) in two translation hypotheses. It is calculated as the size of the intersection divided by the size of the union of the token sets. This metric reflects vocabulary-level similarity and is sensitive to synonymy and word order changes, often underestimating similarity in semantically equivalent but lexically diverse translations. Nevertheless, it provides a simple and interpretable measure of how much two systems reuse similar words, which can be helpful in studying system redundancy or diversity.

The third metric, Levenshtein distance (also known as edit distance), quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one translation into another. We normalized this value to allow comparability across sentences of varying lengths. While Levenshtein distance is primarily a surface-level similarity metric, it offers an intuitive view of how closely aligned the outputs are in terms of literal sequence. In cases where translations are nearly identical except for minor

variations or errors, Levenshtein distance provides a direct estimate of edit effort, which is especially relevant in post-editing scenarios or for assessing system robustness.

To summarize the relationships between system outputs, we computed the mean similarity scores across all sentence pairs for each metric, effectively creating a system-level similarity matrix. Furthermore, to explore how output similarity correlates with translation quality, we analyzed the distribution of similarity scores within sentence pairs grouped according to manual evaluation labels (e.g., accurate vs. inaccurate translations).

4. Results and Discussions

The linguistic evaluation of machine translations from English to Macedonian, which is the subject of this paper, consists of three interconnected experiments. The first experiment involved defining the key error categories and the clusters they naturally belong to. Following this, all translations were manually annotated and compared with human translations. In the second experiment, the quality of each system was assessed using the eight metrics, which were introduced in Section 3.2. The third experiment focused on evaluating sentence similarity through complementary linguistic metrics: cosine similarity, Jaccard similarity and Levenshtein distance.

The results of all three experiments will be presented and discussed in a unified section, with the findings of the first experiment illustrated in detail using examples from both machine and human translations.

4.1. Analysis of Manually Annotated Translation Errors

This section is organized into four subsections, each presenting the results and corresponding discussion for one of the three experiments.

Linguistic Inaccuracy The number of manually annotated errors across all ten types of linguistic inaccuracies is presented in Table 1. In the remainder of this subsection, each type is explained in detail and, where necessary, illustrated with representative examples, demonstrating that these errors not only compromise grammatical correctness but also hinder comprehension of the translated text.

Linguistic inaccuracy accounts for 34.64% of all observed errors, with the majority of 72.99% originating from Microsoft Translator and NLLB-600M. Most of these errors stem from incorrect tense usage, which is not entirely unexpected. Even native Macedonian speakers occasionally confuse the so-called L-form with the traditional past tense, such as the aorist. Although both refer to past actions, the aorist is used when the speaker directly witnessed or participated in the event, whereas the L-form implies the speaker learned of the action indirectly [41]. In the context of Orwell's 1984, Winston recounts events he personally experienced, making the use of the L-form inappropriate. Incorrect tense selection also appeared in GPT-4o during the final ten sentences, suggesting a LLM decline in consistency toward the end of the output. Interestingly, GPT-5 corrected nearly all tense selection errors, with only one exception, similarly to Gemini and Google Translate. These systems appear to favor the L-form, which is stylistically appropriate for the narrative prose of the novel and therefore does not result in frequent verb selection errors.

Table 1. Distribution of linguistic inaccuracy

Linguistic inaccuracy		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Morphological disagreement	Gender disagreement	1	2	1	0	7	20
	Incorrect plural formation	0	0	0	1	0	0
	Number disagreement	0	0	0	0	1	2
	Person disagreement	0	0	0	1	2	1
Verb and tense accuracy	Incorrect verb voice	0	2	0	1	1	0
	Wrong tense selection	8	1	1	2	73	46
	Missing accusative clitic	1	1	2	2	1	2
Syntactic structure	Word order	3	3	0	2	2	4
	Wrong definiteness	5	5	7	12	2	8
	Wrong preposition	1	0	0	2	1	8
Total errors		19	14	11	23	90	91

In contrast, plural formation was generally accurate, except for the noun *виор (transliterated as vior / English: eddy)*, where Google Translate "invented" the plural form *виорови (viorovi)* instead of the correct *виори (viori / eddies)*. Namely, the plural suffix -ови is rarely used in Macedonian and typically appears with monosyllabic masculine nouns, such as *бит – битови (bit – bitovi / bit – bits)*, *рог – рогови (rog – rogovi / horn – horns)*, and *слон – слонови (slon – slonovi / elephant – elephants)*. The noun *виор* is disyllabic *ви-ор* [57].

Number and person agreement were handled successfully by all three LLMs: both GPT versions and Gemini. The three MT systems also performed confidently. However, gender disagreement between personal pronouns, adjectives, and their corresponding nouns was widespread. This issue was especially prevalent in Microsoft Translator and NLLB-600M, with the latter displaying nearly three times as many occurrences. Microsoft's system misclassified typically feminine nouns ending in a, such as *книга (kniga / book)* and *хартија (hartija / paper)*, as masculine. Meanwhile, NLLB-600M tended to treat most nouns and adjectives as neuter, regardless of their actual grammatical gender.

The wrong definiteness of Macedonian nouns and adjectives was the second most frequent linguistic inaccuracy MT issue with 5.45% presence among all errors. However, it is a frequent problem even to native speakers [58]. They often overlook or misuse definiteness, largely due to the influence of English, which handles definiteness differently and more simply. This tendency is reinforced by frequent exposure to English through media and education, code-switching, and informal digital communication, leading to structural simplification and reduced grammatical awareness in native usage [59].

Incorrect voice selection (active vs. passive) did not pose significant problem to most MT systems, except in the case of GPT-5, which mistakenly used the active voice instead of the passive on two occasions. We note that a change in verb voice does not necessarily result in an incorrect translation, as active–passive alternations may preserve propositional meaning. In our annotation scheme, however, incorrect voice selection was marked only in cases where the choice of voice affected discourse structure, information focus, or stylistic appropriateness in the target language. In Macedonian, passive constructions, such as the reflexive passive and, less frequently, the periphrastic passive is commonly used in colloquial and narrative contexts to background the agent or maintain textual cohesion. In the annotated instances, the source sentence employed a passive construction with no explicit

agent, while the MT output rendered it as an active clause with an overt or implied agent, thereby altering the information structure and narrative perspective. Such cases were therefore classified as incorrect voice selection, despite the semantic content remaining largely intact.

Word order errors are diverse. For example, both GPT versions translated the sentence There were no windows in it at all. as *Во него немаше прозорци воопшто (Vo nego nemashe prozorci voopshto)* instead of the more natural *Во него воопшто немаше прозорци. (Vo nego voopshto nemashe prozorci)*. This reflects a literal translation of the phrase at all, which in English is typically placed at the end of a clause to add emphasis in negative statements, conditional clauses, and questions [9]. In Macedonian, however, adverbs such as *воопшто* should be placed next to the verb they modify [60]. Since *воопшто* complements *нема - немаше (nema – nemashe / there is not – there was not)*, it should immediately precede or follow the verb, not the object. Another error, found in both GPT versions and NLLB-600M, is the use of *било кој* instead of the correct *кој било (koj bilo / whichever)*. This reversed order of the pronoun *кој* and the particle *било* is typical of Serbian and frequently appears in colloquial Macedonian, but it is not standard.

A closer examination of the observed linguistic inaccuracies reveals that weaknesses in gender recognition and tense selection manifest differently across MT systems. Gender-related errors are primarily linked to morphological agreement and long-distance dependency resolution, reflecting persistent difficulties in correctly propagating gender features in morphologically rich languages such as Macedonian. In contrast, tense-related inaccuracies, often involving the choice between aorist, imperfect, and L-form constructions appear to stem more from stylistic normalization and system-level preferences rather than from a failure to identify temporal reference per se. While both types of errors affect grammatical correctness, gender mismatches tend to be more disruptive to local syntactic well-formedness, whereas tense-related errors more subtly undermine narrative coherence and stylistic authenticity.

Stylistic Errors While linguistic inaccuracies affect grammatical structure and clarity, stylistic errors undermine the tone, register, and expressive intent of the original text. In this subsection, such issues are analyzed in terms of literalism, punctuation and formatting, and vocabulary inaccuracies (Table 2.).

Stylistic errors represent the core weakness of the MT systems examined, accounting for 48.32% of all annotated errors. More than half of them involve the use of distant synonyms, primarily for nouns, where the intended meaning becomes difficult to infer without knowing the original novel. For example, the noun landing from the sentence “On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.” is translated as *меѓукатие (megjukatie / mezzanine)* by both GPT translators, as *под (pod / floor)* and *слетување (sletuvanje / flight landing)* by all other MT systems. The translation of lift-shaft is even more confusing: *лифт-шахта (lift-shahta / lift manhole)*, *шахта на лифтот (shahta na liftot / manhole for the lift)*, *вратило за лифт (vratilo za lift / lift axle)* all appear. Two translators simply skipped to translate it, while the correct translation *отвор за лифт (otvor za lift / elevator shaft)* was produced only by Microsoft Translator. Even the verb gazed was mistranslated using the close synonym *гледа (gleda / watch, see)*,

[9] www.collinsdictionary.com/dictionary/english/at-all

Table 2. Distribution of stylistic errors

Stylistic errors		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Literalism	Calque	0	0	1	2	2	3
	Foreign language	2	0	0	0	4	33
	Pleonasm	0	0	1	1	2	0
	Too literal translation	1	0	2	2	1	0
Punctuation and formatting	Missing punctuation	2	2	0	1	4	0
	Quotation marks	0	0	2	3	1	0
	Lowercase letters in formal expressions	0	1	0	0	1	0
	Wrong punctuation	4	2	0	0	3	0
Vocabulary inaccuracies	Spelling error	0	1	1	4	1	24
	Misused synonyms	23	21	33	45	23	40
	Incorrect phrasing	7	7	8	7	2	16
Total errors		39	34	48	65	44	116

although the verb *зјана (zjapa)* would more accurately reflect Orwell's intention to convey the act of staring.

Nearly 70% of Google Translate's stylistic errors were due to misused synonyms (45 out of 65 errors in total), indicating that this type of mistake is the primary contributor to its stylistic inaccuracies. Even the most accurate systems among the six studied GPT-4o and GPT-5 averaged more than one stylistic error per five sentences. Integrating a high-quality interpretive bilingual resource, such as the Digital Dictionary of the Macedonian Language [10] could significantly reduce these errors. Unfortunately, due to copyright restrictions, it is currently unavailable for implementation in MT systems.

Incorrect phrasings were also frequently observed, particularly in NLLB-600M. For example, in half of the MT systems, the phrase black-mustachio'd face was mistranslated as a face (made) of a black moustache, rather than a face with a black moustache. Macedonian offers several dedicated adjectives for this expression, including *црномустаќест (crnomustaќjest)*, *црномустак (crnomustak)*, and *црномустаклест (crnomustaklest)*. Notably, both GPT versions produced *црномустак (crnomustak)*, demonstrating that MT systems can generate this less frequent but legitimate Macedonian word correctly, even if it is not the most common form.

Foreign language errors were significantly recorded only with NLLB-600M. This system is directly integrated with Meta's social network, which, according to the company, supports 25 billion translations per day across more than 200 languages This extensive multilingual reach was clearly reflected in our experiment. Specifically, 39 foreign words were detected in the output: 28 in Bulgarian, 9 in Serbian, 1 in Russian, and 1 in Old Church Slavonic. Many of these intrusions were immediately spotted due to the use of Cyrillic characters, which do not exist in the Macedonian alphabet. They include, for example, Bulgarian *одвън (odvŭn / outside)*, *гладък (gladŭk / smooth)*, or *жълта (zhŭlta / yellow)*. Most other Slavic words, for example the Bulgarian *бутилка (butilka / bottle)*, *етажи (etazhi / flights)*, *парцални (parcalni / rag)*, the Russian *слева (sleva / to the left)* and Old Slavonic *ввирено (vvireno / corrugated)* do not exist even in the dialects, thus the Macedonian reader cannot recognize them. The Serbian words: *бели (beli / white)*

[10] http://drmj.eu/

in *бели бетон (beli beton / white concrete)* or *тупи (tupi / blunt)* are acceptable, mainly because the Macedonian adverbs are almost identical: *бел (bel)* and *тап / tap)*. The word *њушкајќи (njushkajkji)*, which, judging by the suffix *-јќи* could be interpreted as a verbal adverb formed from the Serbian verb *њушкати (njushkati, "to snoop")* combined with the Macedonian suffix *-јќи*, is nevertheless unclear. Specifically, the word *њушка (njushka)* exists in several dialects only as a noun, meaning snout or trunk, and is predominantly used in reference to animals.

Calques appeared only rarely. However, without exception, all MT systems that are not LLMs mistranslated the phrase indoor display, each in a different way. The accurate translation: *затворен простор (zatvoren prostor)* is a well-established expression in Macedonian, yet it was not used by any of the systems.

Punctuation errors, particularly missing or misused commas, were also recorded, with 9 instances of each type. This issue is not unique to MT systems; it is also prevalent across many Macedonian news portals, especially those that do not employ official proofreaders. Once legislation enforcing mandatory editing of published texts is fully implemented, the frequency of such errors is expected to decline.

Incorrect phrasing accounts for 13.50% of all stylistic errors. Except for Microsoft Translator, whose performance was almost perfect, such errors occurred nearly ten times in each of the other systems, highlighting a serious and recurring issue. A potential solution would be the integration of a bilingual English – Macedonian phraseological dictionary [11], although it is currently available only in a searchable format and remains under copyright protection. Interestingly, pleonasms, overly literal translations, wrongly used quotation marks and lowercase letters in formal expressions were rare and therefore fall outside the scope of our proposed strategies for improving MT accuracy.

Stylistic errors, ranging from imprecise synonym choices to incorrect phrasing, represent a major obstacle for producing natural, fluent translations. Addressing these challenges will require both better linguistic modelling and access to high-quality bilingual lexical resources.

Lexical and Semantic Errors Lexical and semantic errors arise when the system fails to convey the basic meaning of the source text, in our study due to terminological errors, untranslatability, and context misinterpretation (Table 3.). These errors go beyond grammar or word choice. Namely, they reflect a fundamental misunderstanding of the original sentence and sometimes significantly distort the intended message.

These errors, accounting for just 4.46% of the total, are significantly fewer than those in other categories, particularly since a quarter of them stem not from mistranslation, but from the transliteration of Newspeak terms. During manual annotation, we encountered five pairs of inconsistently translated words and phrases using the NLLB-600M, an inconsistency we did not expect to be on such a large scale.

GPT-4o proved to be a highly competent translator, avoiding the typical errors associated with this category. Its upgraded version, GPT-5, had less consistent handling of these structural elements. For example, the named entity (NE) Victory Mansions at the beginning of the novel was translated as *Победничките Палати (Pobednichkite Palati / Victory Palace)*, while later it appeared as *Победничките Згради (Pobednichkite Zgradi / Victory Buildings)*. According to standard English – Macedonian bilingual dictionaries,

[11] https://zoze.mk/en-mk/

Table 3. Distribution of lexical and semantic errors

Lexical and semantic errors		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Terminological errors	Wrong adverb	0	0	0	0	1	1
	Wrong medical term	0	0	0	1	0	1
	Wrong Newspeak translation	0	0	1	1	0	0
Untranslatability	Inconsistent translation of same word or phrase	0	1	1	0	0	5
	Not translated word which affects the meaning	0	0	1	0	0	0
	Transliteration instead of translation	0	0	0	4	1	3
Context misinterpretation	Wrong POS	0	0	1	2	0	1
	Wrong translation due to lack of informationin the previous context	0	2	1	1	1	1
Total errors		0	3	5	9	3	12

the only translation of mansion is *палата (palata / palace)*. The conjunction *дека (deka / that)* is a better choice than *да (da / that, to)* in the translation of the sentence "It was even conceivable that they watched everybody all the time." This mistake is relatively subtle.

On the other hand, the omission of the preposition *со (so / with)* in the phrase *небото беше со јасно сина боја (neboto beshe so jasno sina boja / the sky a harsh blue)* is a serious issue, as it significantly affects the meaning and clarity of the sentence.

Gemini's inconsistency involves the abbreviation of the named entity *Министерство за изобилство (Ministerstvo za izobilstvo / Ministry of Plenty)*, which should be rendered as *Минизоб (Minizob)* or *Миниизоб (Miniizob)*, rather than *Миниобил (Miniobil)*. Still, the abbreviation used is arguably acceptable, as the adjectives *изобилен (izobilen / abundant)* and *обилен (obilen / abundant)* are often used interchangeably. In its first occurrence, NLLB correctly translated the word caption as *натпис (natpis)*, but in the second instance, it used the Bulgarian form *надпис (nadpis)*. The term Thought Police appears both as *Мислова Полиција (Mislova Policija / Thinking Police)* and *Полиција на мисла (Policija na misla / The Ministry of Thought)*. The term telescreen is translated as both *телевизор (televizor / TV)* and *телескрин (teleskrin / Macedonian pronunciation of telescreen)*. The word alcove was either translated into the non-existent term *алкова (alkova)* or left untranslated, indicating a lack of familiarity with its meaning.

A common issue across most observed systems, except for both versions of GPT, was the misinterpretation of the part of speech (POS) in the source language at the beginning of the fourth sentence: It depicted simply an enormous face. Specifically, the adverb simply was confused with the adjective simple. In Macedonian, both the adverb simply and the neuter singular form of the adjective simple are rendered as *едноставно (ednostavno)*. Incorrect translations placed *едноставно* in front of the noun *лице (lice / face)*, resulting in the incorrect phrase *прикажуваше едноставно огромно лице (prikazhuvashe ednostavno ogromno lice / depicted a simple enormous face)*, instead of the correct translation *едноставно прикажуваше огромно лице (ednostavno prikazhuvashe ogromno lice / depicted simply an enormous face)*.

The case of mistranslation due to a lack of contextual understanding involves the named entity Airstrip One, which was predominantly rendered as *Писта Еден (Pista Eden)* or *Авионска писта Еден (Avionska pista Eden)*, both of which mean Runway One. Even a Macedonian human translator interpreted it as *Воздушниот коридор Еден (Vozdushniot*

koridor Eden / Air Corridor One). In reality, according to Orwell, Airstrip One refers to one of the provinces of Oceania in Orwell's 1984.

Overall, structural and consistency errors were relatively infrequent, with NLLB-600M showing the lowest level of internal coherence among the evaluated models. While some inconsistencies could be explained by lexical ambiguity or lack of context, others highlight the need for improved handling of named entities and domain-specific terminology in MT systems.

Fluency and Naturalness Fluency and naturalness errors occur when the translated text, although grammatically correct, sounds awkward, stilted, or unidiomatic in the target language. These issues often result from meaningless translations, omitted words, non-existent or untranslated units, and unnatural expressions, ultimately reducing the readability and stylistic quality of the output (Table 4).

Table 4. Distribution of fluency and naturalness errors

Fluency and naturalness		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Omissions	Missing conjunction	0	0	1	1	0	1
	Missing phrase	0	0	0	0	0	1
	Missing preposition	1	1	0	2	1	2
	Missing verb	0	0	0	0	1	1
Non-existent or untranslated units	Non-existent word	0	1	2	2	0	21
	Not translated phrase	0	1	0	0	0	2
	Not translated words	0	0	2	2	9	4
	Meaningless translation	1	0	2	2	6	10
Unnatural expressions	Too descriptive	0	0	2	2	2	1
	Misuse of conjunctions	1	1	0	0	0	0
Total errors		3	4	9	11	19	43

GPT-4o produced only one example of a meaningless translation. The phrase: *Лицето му го обликуваше во израз на тивок оптимизам (Liceto ti go oblikuvashe vo izraz na tivok optimizam)* can be back translated as His face was shaped into an expression of quiet optimism, which is lexically different from the original phrase: He had set his features into the expression of quiet optimism. Their deeper analysis shows that they describe the same expression, but they do not express the same meaning.

It seems that GPT-5 on two occasions responded with overconfidence, either by inventing the non-existent and nonsensical word *бљаболеше (bljaboleshe)*, or by omitting words from larger expressions, for example, translating *во овој час (vo ovoj chas / at this time)* instead of the full phrase *во овој час од денот (vo ovoj chas od denot / at this time of day)*.

The quarto-sized term, which refers to a book format created by folding a single sheet of paper into four leaves, is entirely obscure. GPT-4o translated it as *големина кварто (golemina kvarto / size quarto)*; GPT-5 rendered it as *квартов формат (kvartov format / quarto format)*; Gemini used *со големина на четвртина (so golemina na chetvrtina / with a size of a quarter)*; and Google Translate offered *големина на четврт парче (golemina na chetvrt parche / size of a quarter piece)*. Each translation is partially true in its own way, and the last two can be easily visualized by the reader.

Microsoft Translator and NLLB-600M chose not to translate the term at all. This approach is arguably better than inventing incorrect or non-existent words. In relation to this phenomenon, it is worth noting that both Gemini and Google Translate generated two non-existent words each; the most amusing among them is *холиоза (holioza)*, used as a translation of the noun hallway. In contrast, four systems translated hallway correctly as *ходник (hodnik)*, and one as *коридор (koridor)*.

Microsoft Translator decided not to translate some of the Newspeak-named entities, for example Ministry of Truth and Victory Gin and all the abbreviations: Minitrue, Mini-pax, Miniluv, and Miniplenty. This system generated six meaningless translations, the most interesting are the translations of the phrases picked out on its white face, which is *избрани на неговото бело лице (izbrani na negovoto belo lice / selected on his white face)* and He had set his features into the expression, which was transformed to *Тој ги постави своите карактеристики во израз (Toj gi postavi svoite karakteristiki vo izraz / He set his characteristics in the expressions)*. While pick up on is a commonly used phrasal verb meaning to notice or respond to something subtle, selected on is only correct when followed by a specific criterion, such as selected on the basis of merit. In the second example, although back translation of the phrase is almost identical to the original, the Macedonian phrase has no logical interpretation.

Overly descriptive translations are exemplified by the adjective black-mustachio'd, which has already been discussed as an example of incorrect phrasing. Another example is the phrase day in April, which can be translated literally as *ден во април (den vo april)*, a solution used by Gemini, Microsoft Translator, and NLLB-600M, although the more natural expression is *априлски ден (aprilski den)*, as rendered by both GPT models and Google Translate.

NLLB-600M has proven to be a brilliant inventor of non-existent words, most of which are completely incomprehensible, such as *долборот (dolborot)*, *патрула (patrula)*; *шпикувајќи (shpikuvajkji)* or *контејната (kontejnata)*. They contributed to the creation of meaningless translations.

In addition to non-existent words, this MT also invented completely incomprehensible translations. For example, the sentence Winston made for the stairs. was translated as *Уинстон го направи тоа за скалите. (Uinston go napravi toa za skalite. / Winston did it for the stairs.)*. However, the most notable example is the phrase *тревата од вила се држеше над купки од рушеви (trevata od vila se drzeshe and kupki od rushevi)*. Although willow-herb is correctly translated as *врбовка (vrbovka)*, the use of *тревата (trevata / the grass)*, even *тревата од вила (trevata od vila / the fairy grass)*, at least retains some sense. The aorist form of the reflexive verb *се држи (se drzhi)* is *се држеше (se drzheshe)* and in the context of herbs, the English verbs hold, support, keep, stick, and maintain can be used, none of which corresponds closely to the source verb straggle. The phrase *купки од рушеви* contains the non-existent word *рушеви (rushevi)*, which should probably be *рушевини (rushevini / rubble)*, a perfect match. However, the plural noun *купки (kupki / bath)* has nothing in common with the original word heaps. If you cannot guess the source English phrase, here it is: the willow-herb straggled over the heaps of rubble.

Fluency and naturalness remain challenging aspects for machine translation systems, with notable variability in performance among different models. While some LLMs, such as GPT-4o, demonstrate strong capabilities with only minor issues, others like NLLB-

600M frequently produce nonsensical or invented words that undermine overall translation quality.

The nuanced handling of idiomatic expressions, specialized terms, and stylistic subtleties often reveals the limits of current technology. Nevertheless, the relatively low overall error rates highlight encouraging progress toward producing translations that are both accurate and readable.

Comparative Analysis of Manual Error Annotations To assess similarities in error behavior across systems, this section conducts a correlation-based comparison of manual error annotations. The analysis aims to reveal whether different MT systems exhibit shared or distinct error profiles when translating from English to Macedonian.

Table 5 presents the distribution of translation errors across four error categories for three LLM-based systems (GPT-4o, GPT-5, and Gemini 2.5 Flash) and three conventional NMT systems (Google Translate, Microsoft Translator, and NLLB-600M). In total, 526 errors were identified across all systems. The share of errors of LMS- based systems is only 26.74%. Their results are incomparably better than the NMT system.

Table 5. Distribution of translation errors

Linguistic evaluation of MT systems	GPT-4o	GPT-5	Gemini 2.5-Flash	All LLMs	Google Translate	Microsoft Translator	NLLB-600M	All NMT systems
Linguistic inaccuracy	19	14	11	44	23	90	91	204
Stylistic errors	40	35	48	123	65	44	116	225
Lexical and semantic errors	0	3	6	9	9	3	12	24
Fluency and naturalness	3	4	9	16	11	19	43	73
Total inaccuracies	62	56	74	192	108	156	262	526

Across all systems, stylistic errors constitute the most frequent error type (348 instances overall), affecting both LLMs and NMT systems. However, stylistic issues are particularly prominent in NMT output, with 225 errors compared to 123 errors for LLMs. A similar pattern is observed for linguistic inaccuracies, where NMT systems collectively produced 204 errors, compared to only 44 errors for all LLMs combined. This suggests that LLM-based systems handle grammatical and lexical correctness more robustly in English–Macedonian translation.

In contrast, Lexical and semantic errors are relatively rare across all systems, though they appear slightly more often in NMT systems than in LLMs. Errors related to fluency and naturalness show the clearest divergence between paradigms: NMT systems exhibit 73 such errors, whereas LLMs register only 16, reinforcing the observation that LLM-generated translations tend to be more fluent and natural. Overall, the table highlights a consistent trend in which LLM-based systems outperform conventional NMT systems across all evaluated error categories, both in total error count and in qualitative dimensions related to fluency and stylistic adequacy.

The correlation matrix (Table 6.) reveals very strong internal consistency within the LLM group. GPT-4o, GPT-5, and Gemini 2.5 Flash show near-perfect correlations with one another, ranging from 0.93 between GPT-4o and Gemini 2.5Flash to 0.99 between both

Table 6. Correlation matrix

Mutual correlation between all MT systems	GPT-4o	GPT-5	Gemini 2.5-Flash	All LLMs	Google Translate	Microsoft Translator	NLLB-600M	All NMT systems
GPT-4o	1.00	0.99	0.93	0.98	0.97	0.53	0.95	0.91
GPT-5	0.99	1.00	0.97	1.00	0.99	0.41	0.90	0.85
Gemini 2.5Flash	0.93	0.97	1.00	0.98	0.99	0.18	0.79	0.71
All LLMs	0.98	1.00	0.98	1.00	1.00	0.37	0.89	0.83
Google Translate	0.97	0.99	0.99	1.00	1.00	0.32	0.86	0.80
Microsoft Translator	0.53	0.41	0.18	0.37	0.32	1.00	0.74	0.82
NLLB-600M	0.95	0.90	0.79	0.89	0.86	0.74	1.00	0.99
ALL NMT systems	0.91	0.85	0.71	0.83	0.80	0.82	0.99	1.00

GPT versions. The aggregated "All LLMs" score, computed by summing the identified errors across all LLM-based MT systems, is even higher, ranging from r = 0.98 to r = 1.00. This indicates that, despite architectural and training differences, LLM-based systems exhibit highly similar error distributions across the four error categories, suggesting a shared translation behavior and error profile when translating from English to Macedonian.

A similarly strong pattern is observed within the "All NMT" group, defined as the aggregate of manually identified errors across NMT systems only, particularly between NLLB-600M and the resulting aggregated score (r = 0.99). NLLB-600M also correlates strongly with Google Translate (r = 0.86) and Microsoft Translator (r = 0.74), indicating that conventional NMT systems likewise share a broadly comparable error distribution. However, the correlations among NMT systems are generally lower and more variable than those observed among LLMs, pointing to greater heterogeneity within the NMT paradigm.

Cross-paradigm correlations between LLMs and NMT systems are mixed. Google Translate shows very high correlations with LLMs (r = 0.97–0.99), suggesting that its error profile is closer to that of LLM-based systems than to other NMT systems. In contrast, Microsoft Translator exhibits consistently weak correlations with LLMs, ranging from r= 0.18 with Gemini 2.5 Flash to r=0.53 with GPT=40. This divergence reflects systematic differences in how error types are distributed rather than overall quality alone.

4.2. Automatic Evaluation of Observed Machine Translation Systems

According to the manual annotation, Microsoft Translator and NLLB-600M performed noticeably worse than the other systems. Google Translate achieved average results, whereas LLMs were more successful, with GPT-5 outperforming all others. Similarly, the automatic metrics confirmed that the two lowest-performing MT systems also produced the weakest results (see Table 7). Some versions of COMET also identified Google Translate as weaker than the LLM-based systems, a distinction not reflected in the lexical metrics. Additionally, TER rated Gemini and Google Translate as more divergent from the reference translation compared to the GPT models.

These scores are broadly consistent with the manual evaluation, which found that Microsoft Translator and NLLB produced the highest number of errors, with Google Translate trailing slightly behind the remaining systems. However, automatic metrics do not fully capture the qualitative differences within these groups. We also calculated COMETkiwi scores for the human reference translation. Its mean score was 0.7889, lower than most of

Table 7. Mean metric scores for each system.

MT systems / Evaluation metrics	BLEU	chrF	METEOR	TER	BERT score (F1)	COMET	COMET kiwi	XCOMET-XL
GPT-4o	27.28	56.30	0.4953	53.36	0.8669	0.8503	0.8168	0.8611
GPT-5	27.47	57.65	0.5006	51.65	0.8693	0.8488	0.8179	0.8670
Gemini2.5Flash	27.30	56.74	0.5060	54.04	0.869	0.8448	0.8154	0.8501
Google Translate	27.58	56.41	0.5004	54.01	0.8678	0.8322	0.8154	0.8363
Microsoft Translator	22.72	52.34	0.4429	57.64	0.8551	0.8207	0.8054	0.8379
NLLB-600M	17.83	47.54	0.3844	66.44	0.8376	0.7602	0.7722	0.7288

the machine translation outputs, highlighting that automated metrics do not always rank translations with more human-like features at the top.

The distribution of per-sentence scores is shown in Figures 1 and 2. The weaker systems exhibit particularly wide score ranges when evaluated with the COMET models, whereas their ranges remain relatively narrow when assessed using lexical metrics. In contrast, the lexical metrics tend to assign a broader range of scores to translations produced by Google Translate and the LLM systems. Notably, these metrics also assigned very low scores to some systems that were identified through human evaluation as having the fewest errors. This suggests that lexical metrics may underestimate the quality of more fluent or freer translations that, while correct, do not closely resemble the reference.

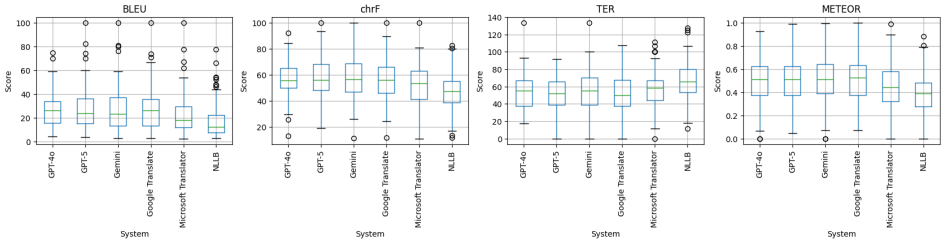


Fig. 1. Distribution of scores for lexical metrics

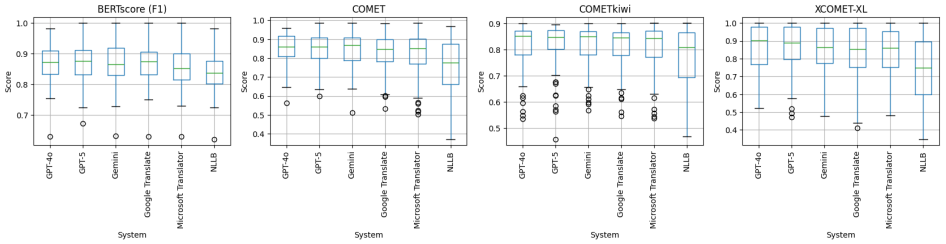


Fig. 2. Distribution of scores for embedding-based and learned metrics

To get a better picture of how each metric compares to the manual evaluation, we computed the correlation between them. We quantified the manual evaluation with a formula inspired by the MQM and ESA scoring systems. Errors were given weights depending on their severity, -1 for minor errors and -5 for major errors. Accuracy errors that change the meaning of the sentence were considered more major than grammatical, punctuation, or stylistic errors. For example, errors from the categories Literalism and Morphological disagreement were given a weight of -1, while missing words, transliteration instead of translation and including non-existent words were given a weight of -5. Correct translations were assigned a weight of 0. The score for each sentence was computed as the weighted sum of errors in that sentence.

The correlation between the manual evaluation and automated metrics is given in Table 8. The first two columns show the sentence-level Spearman correlation between the manual evaluation and each metric, and the second two columns show the system-level Spearman correlation. Note that TER is negatively correlated because a lower TER score indicates better translation quality.

Table 8. Correlation with manual evaluation and pairwise accuracy

	Sentence-level		System-level	
	Correlation with manual evaluation	Pairwise accuracy	Correlation with manual evaluation	Pairwise accuracy
Manual evaluation	1	1	1	1
COMET	0.41	0.528	0.943	0.93
chrF	0.407	0.53	1	1
BERTscore	0.4	0.515	0.943	0.93
METEOR	0.372	0.485	0.486	0.67
XCOMET-XL	0.372	0.52	0.943	0.93
BLEU	0.347	0.41	0.829	0.87
TER	-0.341	0.448	-0.943	0.93
COMETkiwi	0.233	0.45	1	1

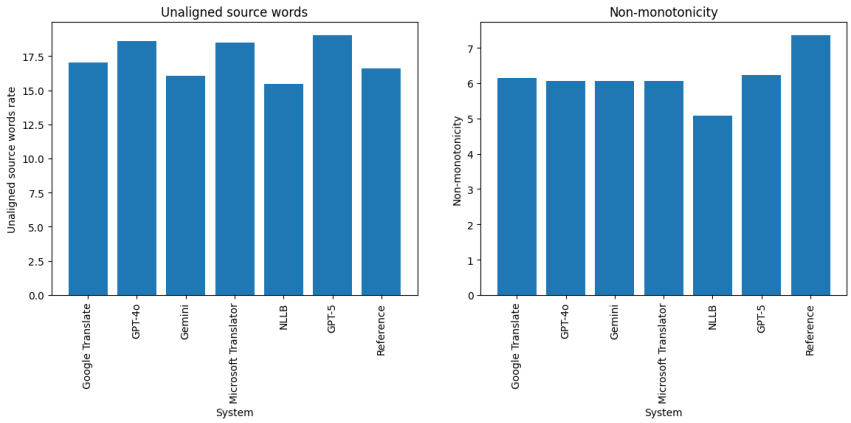
While the system-level correlations are strong, the sentence-level correlations are low to moderate. This is expected, since the human evaluation metric can be prone to fluctuations and is more sensitive to errors that might not be captured by the automated metrics. Among the metrics, COMET and chrF show the high correlations on both the sentence and system level, while BLEU and METEOR lag behind. COMETkiwi and XCOMET-XL, as well as TER (in terms of magnitude) show strong correlations on the system level but are ranked in the middle of the pack or lower on the sentence level. Overall, these results show that the included metrics, especially the highly correlated ones, can adequately rank the systems' translation quality, but highlight the fact that a gap between them and human judgment exists on the sentence level.

We also computed the pairwise accuracy using the human evaluation as the gold standard. A pair of systems is considered to be accurately ranked by a metric if that metric ranks them in the same order as the human evaluator, using the same scoring scheme as above. The pairwise accuracy results follow the trends of the Spearman correlation, with

the same set of metrics generally achieving the best results. One exception is XCOMET-XL, which is ranked low by the Spearman correlation on the sentence level, but is among the best when evaluated through pairwise accuracy.

Further evaluation including multiple human evaluators, which might stabilize the fluctuations in the human evaluation on the sentence level is left for future work.

Inspired by Raunak et al. (2023) [28], we also computed the number of unaligned source words and the degree of non-monotonicity for all MT systems and compared these to the corresponding scores for the reference translation. Additionally, we examined the number of unaligned source words that were not stopwords as defined in the nltk module. To evaluate the aligner model, we used a sample of 60 sentences, 10 from each MT system. While the model made occasional errors, these were rare and consistent across systems. Most errors involved multi-word expressions, where each word in the expression was aligned with all words in the target language. There were almost no instances where a word that should have been aligned was missed.



(a) Unaligned source word rate **(b)** Non-monotonicity score

Fig. 3. Comparison between machine and human translation performance

Figure 3a presents the unaligned source word rate for all MT systems alongside the human reference translation. In the sample analyzed, this metric appears to be unrelated to overall translation quality. Human reference contains fewer unaligned source words than most machine translations. While the relative ordering of systems remains largely the same when stopwords are excluded, the top-ranked system changes, ranking first in this case, compared to second when stopwords are included as potential unaligned words. The non-monotonicity metric, however, reveals that the reference translation has a significantly freer word order compared to the machine translations (Figure 3b). Among the MT systems, there were no substantial differences in scores, except for NLLB-600M, which had by far the lowest non-monotonicity score.

4.3. Comparison of the Similarity of Machine Translations

The last experiment examined sentence-level similarities between the different MT systems. For this analysis, we used cosine similarity based on multilingual MPNet embeddings [56], Jaccard similarity, and Levenshtein distance. We calculated the mean similarity between all pairs of systems and analyzed the distribution of each similarity metric for pairs grouped by translation correctness, as determined by the manual evaluation. Figure 4. presents the mean cosine similarity. As expected, cosine similarity is high across all systems, given that they are translations of the same source text. Nonetheless, certain systems, most notably NLLB-600M, exhibit considerably lower similarity to the translations produced by the other systems.

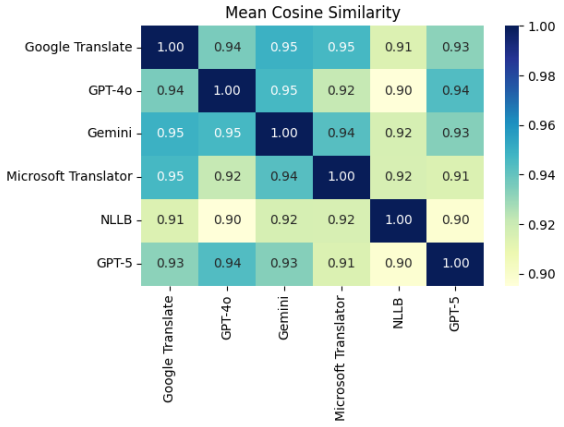
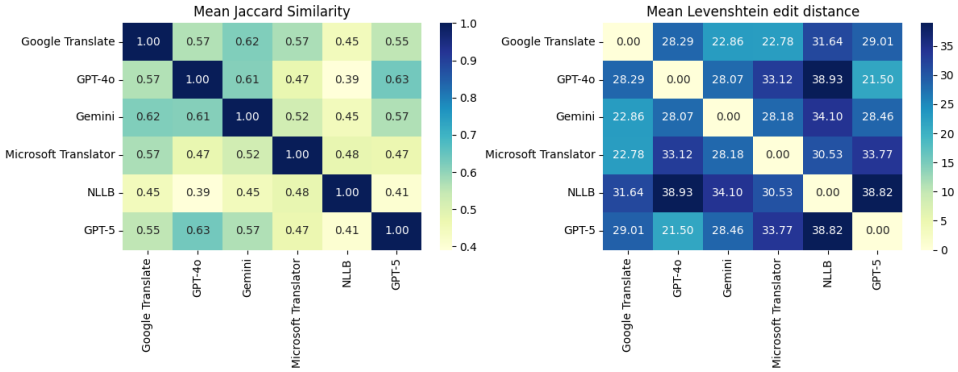


Fig. 4. Mean cosine similarity

The Jaccard similarity and Levenshtein distance, which focus more on the exact words used in each translated sentence, may provide a clearer indication of which systems tend to produce similar translations. The mean scores for these metrics are shown in Figure 5. Both metrics highlight that NLLB-600M stands out as particularly dissimilar from the other systems. At the same time, they reveal a strong similarity between the GPT systems, as expected. Interestingly, the Google-developed systems Google Translate and Gemini also produce notably similar translations, despite their differing architectures.

To determine whether correctly translated sentences tend to be more similar to each other and how different incorrect translations are, we analyzed the distribution of sentence-level similarities between all sentence pairs, regardless of the system, based on their correctness. Sentences with no identified errors in the manual evaluation were labeled as correct, while all others were marked as incorrect. The results are shown in Figure 6.



(a) Jaccard similarity **(b)** Levenshtein distance

Fig. 5. Comparison of exact wording

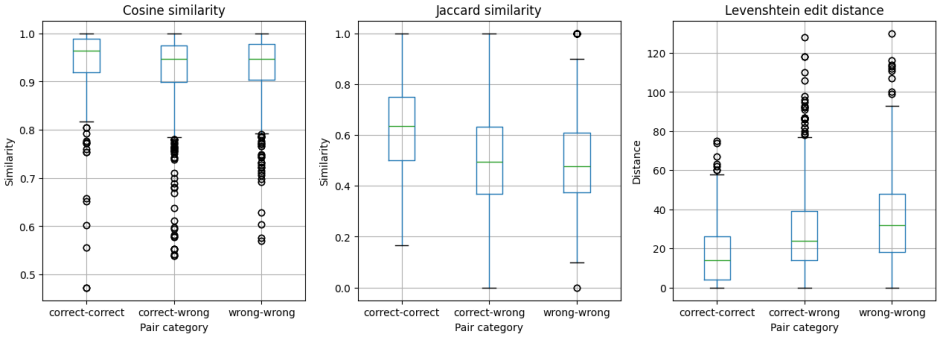


Fig. 6. Distribution of scores for embedding-based and learned metrics

As expected, pairs of correctly translated sentences exhibit higher mean similarity across all metrics, with generally smaller score ranges. For cosine and Jaccard similarity, there is no significant difference between the scores of correct–wrong and wrong–wrong sentence pairs. Both categories have similar means, though correct–wrong pairs tend to show slightly higher variance. A notable difference appears in the Levenshtein distance: wrong–wrong pairs receive higher scores, indicating they are less similar to each other than correct–wrong pairs. Additionally, the variance in Levenshtein scores is considerably higher for wrong–wrong pairs compared to correct–wrong pairs.

We preformed paired Mann-Whitney U test on these results to check their statistical significance. They shows that the difference between the correct-correct and the other groups is significant ($p \ll 0.0001$), while the differences between the correct-wrong and wrong-wrong groups sometimes aren't significant, particularly in the case of cosine similarity and Jaccard distance.

Overall, these findings highlight that sentence-level similarity metrics can reflect patterns in translation quality and system behavior, though their sensitivity varies depending

on the type of metric used. They show that wrong translations tend to diverge more than correct ones. While embedding-based and lexical similarity measures confirm expected trends, such as greater consistency among accurately translated sentences, only certain metrics, like Levenshtein distance, capture more nuanced differences between error types. These insights further underscore the importance of combining multiple evaluation approaches when analysing machine translation output.

5. Conclusions

This study provides a comprehensive linguistic evaluation of six prominent machine translation systems applied to the translation of George Orwell's 1984 into Macedonian. The three LLMs (GPT-4o, GPT-5, and Gemini 2.5 Flash) consistently outperformed traditional MT systems (Google Translate, Microsoft Translator) and especially the social media NLLB-600M, which produced the weakest output in both human and automatic evaluations.

Manual annotation revealed an average of 1.2 errors per sentence, with stylistic (48.47%) and linguistic (34.54%) errors being the most frequent. While LLMs demonstrated superior fluency and syntactic control, evaluation metrics, particularly lexical ones, sometimes penalized these systems for producing accurate but freer translations.

Echoing our initial premise inspired by Tolstoy, while accurate translations often converge, the diversity in error patterns reveals each system's unique translation behavior. Our findings show that accurate translations tend to converge across systems, exhibiting higher consistency and lower variance, especially when assessed using Levenshtein distance, while inaccurate translations display far more variation in both form and structure.

Sentence similarity metrics confirmed these patterns, reinforcing the idea that accurate outputs share common traits, whereas errors are more system-specific. Notably, this discrepancy highlights a key limitation of existing automatic metrics, which still tend to favor surface-level similarity over deeper linguistic accuracy or stylistic nuance. While metrics such as cosine similarity and Jaccard index revealed overall alignment trends, particularly among systems with similar architectures, only Levenshtein distance was sensitive enough to capture nuanced differences between error types.

What we have recognized within our research is that the performance of machine translation systems when applied to Orwell's 1984 is relatively satisfactory, especially when compared with the high-quality translations these systems produced in professional and technical domains. As teachers of several courses delivered in parallel in Macedonian and English, we frequently rely on such systems to translate slides, examination tasks and advertisements. Apart from errors arising from the absence of standardized terminological dictionaries, the translations are almost impeccable. The development of such dictionaries, once common practice but now largely abandoned, would enable experts to rely more confidently on machine translation, particularly for translations from major languages. By contrast, machine-generated translations found on Wikipedia are often of questionable quality, largely because they have not undergone editorial revision. What is particularly unfortunate is that these unedited and often unreliable translations are used as training data for artificial intelligence systems, a practice that risks entrenching errors, amplifying stylistic and semantic distortions, and giving rise to serious long-term problems in the development and deployment of machine translation technologies.

To improve the performance of machine translation systems, particularly those targeting low-resource languages, such as Macedonian, several key steps are necessary.

1. Expand and enrich training data
   There is a pressing need for high-quality, annotated training data in Macedonian and similar low-resource languages. Data collection efforts should prioritize stylistically diverse and domain-rich corpora, including literary texts, formal documents, and colloquial speech. The inclusion of varied linguistic registers ensures broader system generalizability and cultural relevance. The primary problem for the Macedonian language is the lack of a national corpus, which has been initiated for years by several official institutions, but unfortunately, it has never been created. There are collections of texts on the web, most of them are copyright protected, so they are unusable for training LLMs. A coordinated, state funded initiative to establish an open, legally compliant national corpus would therefore be a crucial step toward enabling effective and equitable development of language technologies for Macedonian.
2. Improve access to bilingual and interpretative resources
   Greater access to paper-based bilingual and interpretative dictionaries is essential. In the case of Macedonian, many of these valuable resources remain protected by copyright and are underutilized in digital contexts. Digitizing and licensing these materials for research and development could significantly enhance language visibility and resource accessibility. For Macedonian, this particularly applies to dictionaries by private publishers that do not allow the digitization of their dictionaries. Interestingly, there are two digital dictionaries of the Macedonian language: an interpretative dictionary published by the government (`https://makedonski.gov.mk/`) and a private one (`http://drmj.eu/`). The second one illegally digitized the existing bilingual dictionaries from English, Albanian and Turkish. Unfortunately, both dictionaries are not accessible for automated data extraction, so they have no impact on LLMs. Some digital repositories already exist, most of them are not in a fully machine-readable format, thus they cannot be used as training data for machine learning models.
3. Modernize and digitize phraseological dictionaries
   Existing phraseological dictionaries should be updated and digitized to reflect contemporary usage. This includes the incorporation of newly coined expressions and frequently used idioms, which are often underrepresented in traditional MT training data yet crucial for fluent, culturally accurate translations. The two official phraseological dictionaries of the Macedonian language were published in 2003 [61] and in 2008 [62]. Both were sold out a long time ago and they are not available in electronic format. In the meantime, the language has evolved and thousands of words and phrases have entered it, which are visible in the interpretive dictionary.
4. Advance evaluation methods
   MT evaluation must go beyond surface-level similarity. A more diverse set of evaluation metrics is needed to capture the nuanced behavior of translation systems. Refining current methods or incorporating reference-free approaches, such as quality estimation models, and human-in-the-loop evaluation protocols will lead to more accurate and reliable assessments of translation quality. This applies to all languages, not only to low-resourced ones. Our proposal to introduce mutual similarity can be the initial basis for establishing some new metric. For now, it is primarily assessed by existing

metrics that do not relate to mass translation. In our next research, we are thinking about establishing some combined metric that would unite both directions.

5. Link error patterns to linguistic phenomena
   Future research should investigate whether observed translation divergences correspond to specific linguistic phenomena or error taxonomies. Establishing such links can offer deeper insights into model limitations and guide more linguistically informed improvements. In our work, we have already pointed out four types and causes of translation errors: linguistic inaccuracies, stylistic errors, translation errors, and fluency and naturalness. They are widely accepted and are applied for manual translation evaluation. What is specific is that most of these errors are related to the grammar and spelling of the Macedonian language, so it is not possible to map them into other languages.
   The joint effort of researchers of mutually similar languages would contribute to the establishment of common patterns of errors and the discovery of interesting linguistic phenomena. For example, in the Macedonian language, but probably in other languages as well, there are translation errors that are calques from English, although there are original words for the same concepts that, unfortunately, under the pressure of the mass use of new words, become extinct over time.
6. Enhance low-resource support in open-source models
   Developers of open-source models, such as NLLB-600M, should prioritize the robust handling of low-resource languages. This involves fine-tuning on carefully curated datasets and minimizing common issues like hallucinations or the generation of non-existent words. Special attention should be given to ensuring linguistic fidelity, especially in morphologically rich and syntactically complex languages. Unfortunately, this requires synchronized action between linguists, who often lack technical training, and computer scientists, who may have limited expertise in the grammatical structure of the language. A true symbiosis between these two expert profiles would yield excellent results, provided they develop a shared conceptual and terminological framework. The lack of common framework is probably the main reason why the initiative for interdisciplinary studies in computational linguistics, which was started several times, was never established in our country.
   Achieving equal translation quality in all languages requires a coordinated, multidisciplinary effort. Stakeholders, primarily researchers, developers, linguists, and policy makers, should work together to: support data collection and annotation initiatives; supported community-led and open access projects; enabled responsible digitization and licensing of key language resources; promoted evaluation frameworks that are inclusive of low-resource contexts, balancing automated metrics with human expertise; and encouraged language diversity in machine translation benchmarks and collaborative tasks, thereby encouraging innovation beyond high-resource language pairs.

6. Limitations and Future Work

Despite the depth of the manual analysis and the expertise involved, this study has the following three limitations.

The error annotation was performed by a single primary annotator, and no formal inter-annotator agreement metrics were computed. Future studies will explore simplifying

or consolidating low-frequency error categories and expanding the corpus to other genres and language pairs, thereby enabling more robust quantitative analysis and facilitating broader comparison with established evaluation frameworks.

A further limitation concerns the size and scope of the evaluation dataset. The manual analysis is based on 100 sentences, which reflects a common trade-off in fine-grained MT error analysis between analytical depth and dataset scale. While this sample cannot be claimed to be fully representative of all English–Macedonian translation phenomena, it was selected to include a range of syntactic structures, lexical choices, and stylistic features characteristic of literary prose. Consequently, the findings should be interpreted as indicative of systematic tendencies in system behavior rather than as exhaustive performance measurements. Future work will aim to expand the dataset and validate the observed trends across additional texts and genres.

An additional consideration relates to the choice of Orwell's "1984" as the source text for evaluation and the potential risk of prior exposure by MT systems, particularly LLM-based models. Although the Macedonian translation has existed for many years, its electronic version is not publicly available and, to our knowledge, was not included in MT training corpora. The digital version used here is part of MULTEXT-East and was created for linguistic research. While an alternative translation is now partially accessible online, it differs substantially from our reference and was not used in this study. Moreover, it is subject to copyright protection by both the publisher and the hosting platform, which makes its inclusion in large-scale training datasets less likely. Future work will further enhance robustness by including additional control texts of similar literary and stylistic complexity, as well as newly produced or unpublished reference translations.

Acknowledgement. This work is partially financed by the Ministry of Education and Science of the Republic of North Macedonia through the project "Utilising AI and National Large Language Models to Advance Macedonian Language Capabilities".

References

1. Al-Awawdeh, N.: Translation between creativity and reproducing an equivalent original text. Psychology and Education Journal 58(1), 2559–2564 (2021)
2. Levỳ, J.: The art of translation. John Benjamins Publishing Company (2011)
3. House, J.: Translation quality assessment: Past and present. Routledge (2014)
4. Lefevere, A.: Translation/history/culture: A sourcebook. Routledge (2002)
5. Koby, G.S., Fields, P., Hague, D.R., Lommel, A., Melby, A.: Defining translation quality. Tradumàtica (12), 0413–420 (2014)
6. Scott, C.: Literary translation and the rediscovery of reading. Cambridge University Press (2012)
7. Singh, S.P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., Jain, S.: Machine translation using deep learning: An overview. In: 2017 international conference on computer, communications and electronics (comptelix). pp. 162–167. IEEE (2017)
8. Ide, N., Véronis, J.: MULTEXT: Multilingual text tools and corpora. In: COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics (1994)
9. Erjavec, T.: Multext-east. In: Handbook of linguistic annotation, pp. 441–462. Springer (2017)
10. Bonchanoski, M., Zdravkova, K.: Machine Learning-based approach to automatic POS tagging of Macedonian language. In: Proceedings of the 8th Balkan Conference in Informatics. pp. 1–8 (2017)

11. Bonchanoski, M., Zdravkova, K.: Learning syntactic tagging of Macedonian language. Computer Science and Information Systems 15(3), 799–820 (2018)
12. Ljubešić, N., Zdravkova, K., Stojanoska, S., Erjavec, T., Krsnik, L.: The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Macedonian 1.1 (2021)
13. Newmark, P.: A textbook of translation, vol. 66. Prentice Hall New York (1988)
14. Stapleton, P., Kin, B.L.K.: Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. English for Specific Purposes (2019), https://api.semanticscholar.org/CorpusID:201394427
15. Reinhart, C.M., Rogoff, K.S.: Is the 2007 US sub-prime financial crisis so different? An international historical comparison. American Economic Review 98(2), 339–344 (2008)
16. Marie, B.: An automatic evaluation of the WMT22 general machine translation task. arXiv preprint arXiv:2209.14172 (2022)
17. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation. pp. 392–395 (2015)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
19. Rei, R., De Souza, J.G., Alves, D., Zerva, C., Farinha, A.C., Glushkova, T., Lavie, A., Coheur, L., Martins, A.F.: COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 578–585 (2022)
20. Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 388–395. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-3250/
21. Kocmi, T., Zouhar, V., Federmann, C., Post, M.: Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1999–2014. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://aclanthology.org/2024.acl-long.110/
22. Jiao, W., Wang, W., Huang, J.t., Wang, X., Shi, S., Tu, Z.: Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745 (2023)
23. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: 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. pp. 223–231 (2006)
24. Van Egdom, G.W., Kosters, O., Declercq, C.: The riddle of (literary) machine translation quality. Revista Tradumàtica: Traduccio i Tecnologies de la Informacio i la Comunicacio (21), 129–159 (2023)
25. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How good are gpt models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023)
26. Sizov, F., España-Bonet, C., Van Genabith, J., Xie, R., Dutta Chowdhury, K.: Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1183–1199. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.116/
27. Luo, J., Cherry, C., Foster, G.: To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. Transactions of the Association for Computational Linguistics 12, 355–371 (04 2024), https://doi.org/10.1162/tacl_a_00645
28. Raunak, V., Menezes, A., Post, M., Hassan, H.: Do GPTs Produce Less Literal Translations? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1041–

1050. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-short.90/
29. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., et al.: Preliminary WMT24 ranking of general MT systems and LLMs. arXiv preprint arXiv:2407.19884 (2024)
30. Lommel, A.R., Burchardt, A., Uszkoreit, H.: Multidimensional quality metrics: a flexible system for assessing translation quality. In: Proceedings of Translating and the Computer 35. Aslib, London, UK (Nov 28-29 2013), https://aclanthology.org/2013.tc-1.6/
31. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W.: Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics 9, 1460–1474 (12 2021), https://doi.org/10.1162/tacl_a_00437
32. Klubička, F., Toral, A., Sánchez-Cartagena, V.M.: Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. Machine Translation 32(3), 195–215 (2018)
33. Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., Shmatova, M.: Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1440–1453. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.131/
34. Karpinska, M., Iyyer, M.: Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation. pp. 419–451. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.wmt-1.41/
35. Manakhimova, S., Macketanz, V., Avramidis, E., Lapshinova-Koltunski, E., Bagdasarov, S., Möller, S.: Investigating the Linguistic Performance of Large Language Models in Machine Translation. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 355–371. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.28/
36. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Shmatova, M., Steingrímsson, S., Zouhar, V.: Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1–46. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.1/
37. Costa-Jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
38. Polio, C.G.: Measures of linguistic accuracy in second language writing research. Language learning 47(1), 101–143 (1997)
39. Corbett, G.G.: Morphology and agreement. The handbook of morphology pp. 191–205 (2017)
40. Benson, S., DeKeyser, R.: Effects of written corrective feedback and language aptitude on verb tense accuracy. Language Teaching Research 23(6), 702–726 (2019)
41. Friedman, V.A.: Macedonian. Lincom Europa Munich (2002)
42. Steedman, M.: The syntactic process. MIT press (2001)
43. Harris, R.A.: Writing with clarity and style: A guide to rhetorical devices for contemporary writers. Routledge (2017)

44. Chironova, I.I.: Literalism in translation: Evil to be avoided or unavoidable reality. Journal of Translation and Interpretation 7, 1–28 (2014)
45. Fitria, T.N.: Performance of google translate, microsoft translator, and deepL translator: Error analysis of translation result. Al-Lisan: Jurnal Bahasa (e-Journal) 8(2), 115–138 (2023)
46. Bruton, A.: Vocabulary learning from dictionary referencing and language feedback in EFL translational writing. Language Teaching Research 11(4), 413–431 (2007)
47. Popović, M.: Error classification and analysis for machine translation quality assessment. In: Translation quality assessment: From principles to practice, pp. 129–158. Springer (2018)
48. Rahmatillah, K.: Translation errors in the process of translation. Journal of English and Education (JEE) (2013)
49. Ping, K.: Translatability vs. untranslatability: A sociosemiotic perspective. Babel 45(4), 289–300 (1999)
50. Linlin, L.: Artificial intelligence translator DeepL translation quality control. Procedia Computer Science 247, 710–717 (2024)
51. Post, M.: A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771 (2018)
52. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
53. Guerreiro, N.M., Rei, R., Stigt, D.v., Coheur, L., Colombo, P., Martins, A.F.: XCOMET: Transparent machine translation evaluation through fine-grained error detection. Transactions of the Association for Computational Linguistics 12, 979–995 (2024)
54. Rei, R., Treviso, M., Guerreiro, N.M., Zerva, C., Farinha, A.C., Maroti, C., De Souza, J.G., Glushkova, T., Alves, D.M., Lavie, A., et al.: CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. arXiv preprint arXiv:2209.06243 (2022)
55. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
56. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
57. Zdravkova, K., Kuzmanova, J.: Sonority Based Syllabification of Macedonian and Serbian. Technical editors p. 11 (2024)
58. Mirkulovska, M.: Definiteness in Macedonian with some parallels in Bulgarian. Z. Topolinjska & M. Mirkulovska, Balkan morpho-syntactic similarities a bridge for levelling differences among people. Retrieved June 20, 2018 (2005)
59. Dimova, S.: English in Macedonia. World Englishes 24(2), 187–202 (2005)
60. Tomić, O.M.: Macedonian as an Ausbau language. Pluricentric Languages. Different Norms in Different Countries. Ed. Michael Clyne, Berlin/New York. Mouton/de Gruyter pp. 437–454 (1992)
61. Shirilov, T., Dimitrovski, T.: Фразеолошки речник на македонскиот јазик [Phraseological Dictionary of the Macedonian Language]. 3 volumes, Огледало [Ogledalo], Skopje (2003–2009), iSBNs for volumes: 9989-686-22-X; 978-9989-686-02-3; 978-9989-686-13-9
62. Velkovska, S.: Македонска фразеологија со мал фразеолошки речник [Macedonian Phraseology with a Small Phraseological Dictionary]. БПТ принт [BPT Print], Skopje (2008)

Jana Kuzmanova completed her undergraduate studies at the Technical University of Munich and subsequently earned her master's degree from the University Ss. Cyril and Methodius in Skopje. Since the 2021/2022 academic year, she has been pursuing a PhD at the same institution. Her research interests include machine learning, data mining and its scientific applications, high-performance computing, and natural language processing.

Katerina Zdravkova is a full professor at the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje. Her research and teaching activities span areas such as artificial intelligence, machine translation, computer ethics, e-learning, and information systems. She has authored numerous scientific publications and has been actively involved in international conferences and editorial work in computer science.

Received: October 20, 2025; Accepted: April 15, 2026.