https://doi.org/10.18485/analiff.2026.38.1.7
81'255:004.89
004.738.5:004.89

Comparative analysis of Serbian-to-English translation: a study of LLM-based system vs university students' performance

Emilija Z. Mitić*
Singidunum University, Department of English
https://orcid.org/0009-0008-9312-3849
Valentina B. Bošković Marković
Singidunum University, Department of English
https://orcid.org/0000-0001-8873-6706

Key words:
AI,
LLM,
translation quality index,
ChatGPT,
error,
style

Abstract

This paper examines the quality of Serbian-to-English translations produced by university students and by a large language model, represented by ChatGPT. The study is situated within the broader context of the increasing use of artificial intelligence and large language models in translation practice and language education. Although AI-assisted translation systems show considerable progress in grammatical accuracy and fluency, their ability to interpret cultural references, pragmatic meaning, idiomatic expressions, and stylistic nuance remains open to further investigation. The main aim of the study was to compare the translation performance of ChatGPT and second-year university students in an academic setting. More specifically, the research sought to identify the strengths and limitations of each translation approach across formal, linguistic, pragmatic, and stylistic dimensions. The study was based on identical Serbian-to-English translation tasks assigned to ChatGPT and to students from the Department of English Language at Singidunum University. A corpus of 50 student translations and one AI-generated translation was analyzed using a Translation Quality Index. This index combined four weighted components: Quantitative Score, Linguistic Score, Pragmatic Score, and Stylistic Score. The evaluation included both quantitative error-based analysis and qualitative assessment of meaning accuracy, contextual appropriateness, cultural localization, fluency, and readability. The results showed that overall translation quality was comparable, with students achieving a final TQI score of 85.08 and ChatGPT achieving 84.77. ChatGPT performed better in formal and linguistic accuracy, particularly in grammar, consistency, and sentence-level correctness. However, students outperformed ChatGPT in pragmatic and stylistic categories, especially in culturally appropriate choices, idiomatic expression, register, and naturalness. These findings suggest that AI systems can effectively support translation practice, but they should not be viewed as replacements for human translators. The study supports a hybrid model in which AI contributes formal precision and efficiency, while human translators provide cultural awareness, interpretive judgment, and stylistic sensitivity. *(примљено: 5. фебруара 2026; прихваћено: 19. априла 2026)*

https://anali.fil.bg.ac.rs
АНАЛИ ФИЛОЛОШКОГ ФАКУЛТЕТА

* Univerzitet Singidunum
Departman za engleski jezik
Danijelova 32
11000 Beograd, Srbija
emarkovic@singidunum.ac.rs

1. Introduction

The increasing reliance on technology-mediated instruction in higher education has reshaped language teaching practices and teacher–student interaction, particularly in online and hybrid learning environments (Ilić/Bošković Marković, 2020; Veljković Michos/Bošković Marković, 2021). The translation industry, a cornerstone of global communication, has undergone profound transformations with the advent of neural machine translation (NMT), artificial intelligence (AI), and large language models (LLMs). All these technologies have significantly altered traditional translation practices, raising fundamental questions about the evolving roles of human translators and machine-driven systems. Neural machine translation (NMT) has demonstrated remarkable improvements in speed, efficiency, and grammatical accuracy over time. In contrast, the capabilities of AI-driven translation systems, particularly large language models such as ChatGPT, represent a significant advancement over previous NMT approaches. By employing deep learning and vast linguistic datasets, these systems can produce translations with even higher lexical accuracy and syntactic consistency. However, despite these advancements, AI translations often exhibit deficiencies in handling cultural nuances, idiomatic expressions, and pragmatic meaning; factors that human translators instinctively navigate. Conversely, human translation may demonstrate greater cultural and contextual awareness, but can be susceptible to inconsistencies in grammatical precision, terminological accuracy, and structural coherence.

This dynamic has sparked discourse on the comparative strengths and weaknesses of AI-generated versus human translations. Our study focuses on a linguistically complex pair such as Serbian and English. The complexity stems from significant differences in grammatical structure, syntax, morphology, and cultural nuances. While English relies on a fixed word order and auxiliary verbs to convey grammatical relationships and meaning, Serbian employs a rich case system, flexible syntax, and aspect-based verb distinctions, making translation challenging. Additionally, disparities in idiomatic expressions and pragmatic conventions further complicate the translation process. These linguistic discrepancies are examined in depth in the subsequent sections, serving as a basis for identifying and addressing the challenges encountered in both AI-generated and human translations.

By employing a Translation Quality Index (TQI), this study seeks to objectively compare the translation performance of AI and university students in the following translation direction: Serbian to English. TQI integrates both quantitative and qualitative metrics, allowing for a systematic comparison of these translations. The development and application of this index are discussed in detail in the Methodology section, where the specific criteria and scoring mechanisms will be thoroughly explained. This structured approach strives to ensure an objective and, most importantly, replicable assessment of translation quality, and to identify complementary strengths and recurring limitations in both translation approaches. In doing so, it offers insights into the evolving interplay between human expertise and AI-assisted translation, thus contributing to academic translation training and shaping future pedagogical approaches to translation education.

2. Theoretical background

2.1. Machine Translation (MT) and AI

The field of Artificial Intelligence (AI) is considered to have been formally established during the Dartmouth Summer Research Project on Artificial Intelligence in 1956, organized at the prestigious Dartmouth College (USA). This event is widely recognized as the birth of AI as a distinct academic discipline, with its organizers and leading scholars stating the following on the reasons and the purpose of such technology:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. (McCarthy et al., 1955: 1)

Machine Translation (MT), a subfield of AI, has evolved through several distinct phases:

1. Early Developments (1950s–1960s): The pioneering endeavor of MT was marked by a joint experiment of Georgetown University and IBM conducted in 1954, where computational linguistics scholars and IBM staff managed to successfully translate over sixty Russian sentences into English, thus laying the foundations of the idea that machines might be able to process and interpret foreign languages. The demonstration fostered optimism about the possibility of rapid, effortless international communication in a global context, increasingly marked by political tension and mutual misunderstanding (Hutchins, 2004). This and subsequent early experiments predominantly used rule-based approaches, where linguistic rules were manually crafted to enable translation.

2. Challenges and Critiques (1960s): Although these machine translation systems looked promising, their further application would reveal limitations in their ability to capture the complexities of human language, and specifically so when it came to processing languages with complex syntactic features, which led to their overall reduced accuracy and limited practical adoption (Shahmerdanova, 2025). Despite early optimism, the Automatic Language Processing Advisory Committee (ALPAC, 1966) concluded that machine translation systems failed to meet expectations, resulting in a significant reduction in funding for the following two decades, and thereby contributed to a decline of interest in MT in the US. ALPAC (1966) findings emphasized the linguistic inaccuracy and inconsistency of MT, assessing its output as mediocre compared to that of human translators, and highlighted it as cost-ineffective. Moreover, ALPAC (1966) recognised the need for more fundamental linguistic research before machine translation systems were to be granted further funding. Chomsky (1975, as cited in Lee/Liao, 2011: 111) shared this stance and argued that MT "seemed [to me] pointless as well as probably quite hopeless".

3. Statistical Approaches (1980s–1990s): The idea underlying this approach was to learn translation patterns from previously translated examples, because translation *per se* involved numerous decisions that are difficult to formalize. This principle employed translation memory systems used by human translators, which stored and retrieved past translations to assist with new input texts. Building on this idea, example-based machine translation systems emerged in Japan in the 1980s, aiming to locate similar sentences in parallel corpora and adapt their existing translations accordingly. In the late 1980s, this paradigm evolved into statistical machine translation at IBM Research (Koehn, 2009). However, Hardmeier (2015) observed that SMT systems were still constrained by their reliance on statistical probabilities based on simplified assumptions of direct translational equivalence rather than understanding context at a deeper level, thus failing to capture the complexities of the translation process, especially in the realm of pragmatics and semantics.

4. Neural Machine Translation (2010s–Present): The adoption of neural networks and deep learning marked a turning point in the development of machine translation. By leveraging artificial neural architectures, deep learning enables translation systems to learn complex patterns from large-scale bilingual data and to refine their performance over time with minimal manual intervention (Gajiwala, 2025). In the context of machine translation, Naveen and Trojovský (2024) observe that neural machine translation systems, particularly those based on transformer architectures, have achieved substantial improvements over earlier systems by modeling entire sentences as integrated sequences rather than as isolated units. The incorporation of attention mechanisms marks a groundbreaking innovation in this respect, allowing the system to dynamically focus on relevant parts of the input during translation and capture dependencies between words regardless of their location within the sentence. This capability has significantly enhanced translation quality, predominantly in terms of context, addressing what had long been one of the most persistent weaknesses of machine translation (Naveen/Trojovský, 2024).

2.2. Large Language Models (LLMs)

Having analyzed the virtues and flaws of all the previously mentioned models, scientists have created their improved version known as Large Language Models (LLMs), which are a type of deep neural network in the field of artificial intelligence. They are trained by processing massive amounts of data, which enables them to recognize, understand, and generate content that closely mimics human, natural language styles, as well as to solve a wide range of tasks. They function based on the principles of probability and prediction. During training, we provide them with prompts; each time we feed the model text for training, it analyzes it word by word. During this processing, the model predicts the next word in a sentence and then evaluates whether its predicted sequence matches the original text. The models memorize sequences and patterns. Consequently, when independently generating text based on a prompt, they again predict the most likely next word in a sentence

using probability. Once trained, these models can be applied and adapted to various needs, such as translation or debugging code. LLMs can also be trained as AI agents, allowing them to independently solve various tasks that would otherwise be handled by humans. Large Language Models emerged as a result of decades of research and progress in the field of Natural Language Processing (NLP) and research in the field of machine learning. The development of artificial intelligence, and consequently large language models, reached its peak in the period from 2010 to the present day. Some of the most famous artificial intelligence models – specifically large language models that are easily accessible to the public – include ChatGPT, Perplexity AI, DeepSeek, Gemini, and Copilot.

3. Methodology

While existing research has primarily focused on students' perceptions of AI chatbots in academic contexts (Pantelić et al., 2023), the present study shifts the focus toward an empirical evaluation of translation performance produced by students and an AI system under comparable conditions. This chapter provides a comprehensive explanation of the Translation Quality Index (TQI), the methodology employed in this research, which serves as the primary tool for comparing translations produced by university students and ChatGPT. TQI combines both quantitative and qualitative metrics to assess translation quality from multiple dimensions. The pedagogical orientation of the Translation Quality Index (TQI) reflects broader trends in the integration of digital tools into university-level language instruction, where technology functions as a complement to, rather than a replacement for, human expertise (Veljković Michos/Bošković Marković, 2021). Components of TQI are presented in detail within the following subchapters, with a thorough explanation of the calculation process and how each factor contributes to the overall evaluation. The proposed Translation Quality Index is not intended as a universal QA model, but as a pedagogically motivated, replicable framework suitable for comparative human–AI translation analysis.

3.1. Overview of the Translation Quality Index (TQI)

Translation Quality Index (TQI) consists of four components, the sum of which results in a comprehensive measure of translation quality:

1. Quantitative Score (Q)
2. Linguistic Score (L)
3. Pragmatic Score (P)
4. Stylistic Score (S)

The Translation Quality Index (TQI) proposed in this study is grounded in established approaches to Translation Quality Assessment (TQA) and operationalized through the methodological framework described in Section 3. Rather than aiming to replace existing models, TQI integrates insights from functional, pragmatic, and error-based approaches into a composite evaluation scheme tailored to comparative human–AI translation analysis in an academic context. From a functional–pragmatic

perspective, the Pragmatic Score (P), as defined in Section 3.4, reflects key principles of House's (1997) model of translation quality assessment, particularly the emphasis on register, genre, and pragmatic equivalence. By explicitly evaluating register appropriateness, cultural localization, and the preservation of pragmatic intent, this component operationalizes the assumption that translation quality must be assessed in relation to communicative function rather than formal accuracy alone.

Similarly, the Linguistic Score (L) and Pragmatic Score (P), described in Sections 3.3 and 3.4, respectively, are conceptually aligned with Nord's (1997) functionalist approach and Skopos theory. Although the present study does not manipulate Skopos variables experimentally, the evaluation criteria assume a stable communicative purpose of the target text and assess whether this purpose is successfully fulfilled. This functional orientation is particularly relevant in the comparison between student and AI-generated translations, where divergences often arise in the degree of functional and contextual adaptation rather than in surface correctness.

The distinction between the Quantitative Score (Q) and the Linguistic Score (L), detailed in Sections 3.2 and 3.3, further reflects Pym's (2010) discussion of equivalence and communicative risk in translation. By separating form-based accuracy from meaning-oriented adequacy, TQI makes it possible to identify translation choices that are formally correct yet pragmatically risky, as well as those that prioritize contextual appropriateness over surface precision. This analytical separation is central to the study's aim of revealing complementary strengths and weaknesses in human and AI translation performance.

Finally, the Quantitative Score (Q), outlined in Section 3.2, draws on principles underlying error-based evaluation frameworks such as the ATA error typology and the Multidimensional Quality Metrics (MQM). While these frameworks typically employ extensive and highly granular taxonomies for professional or industrial assessment, TQI adapts its core logic to a simplified scoring system that maintains transparency, proportionality, and replicability across a larger dataset. The deliberate limitation of error categories ensures consistency in evaluation while avoiding overlap with meaning- and pragmatics-based parameters.

Taken together, the four components of TQI, as implemented in the Methodology section, form a coherent evaluative framework that bridges theoretical approaches in translation studies with empirical assessment procedures. This explicit alignment between theory and methodology supports the use of TQI as both an analytical and diagnostic tool in comparative studies of human and AI translation, particularly within the domain of translation pedagogy. Each component is assigned a weight coefficient that reflects its importance within the context of the translation task. The sum of the weights equals 1.

The final TQI is calculated using the following formula:

$$TQI = 0.20\,Q + 0.30\,L + 0.30\,P + 0.20\,S$$

Where:

- Q represents the Quantitative Score (form-based accuracy)
- L denotes the Linguistic Score (meaning accuracy)

• P refers to the Pragmatic Score (appropriateness and localization)
• S stands for the Stylistic Score (fluency and readability)

This particular weight distribution has been chosen with a view to balancing accuracy with higher-level communicative qualities. The heaviest weights (30% each) are assigned to L and P. These two dimensions represent the core of communicative translation quality, where the former relates to the faithful transfer of meaning (L) and the latter refers to successful adaptation of the text to the target cultural and contextual environment (P). By assigning the heaviest weights to these categories, the model ensures that both semantic precision and cultural appropriateness are recognized as the most critical indicators of translation quality, whether produced by students or by the machine system. On the other hand, Q and S are both assigned 20% each. This distribution acknowledges that, while accuracy of formal elements (e.g., orthography) and stylistic refinement are vital to professional translation, they are not as decisive for communicative adequacy as L and P. Nevertheless, by including them as weighted but not dominant factors, the formula seeks to retain sensitivity to surface precision and literary quality, the areas where machine and human translations often diverge most clearly.

Ultimately, this weighting scheme allows the TQI to function not only as a performance index but also as a diagnostic tool. The weighting system ensures that the TQI framework captures both core adequacy (L and P) and supporting qualities (Q and S). This approach aligns with the research objective of the paper, which is not simply to assign a numerical grade to translations, but to reveal how different translators (students vs AI) perform across varied dimensions of translation quality, thereby underscoring their relative strengths and weaknesses.

3.2. Quantitative Score (Q)

The Quantitative Score (Q) measures the formal accuracy of each translation, focusing strictly on errors of *form* rather than meaning, context, or style. It evaluates whether the translation conforms to the expected linguistic forms of English such as orthography, morphology, grammar, and surface accuracy, without taking into consideration deeper semantic adequacy. This separation ensures that Q does not overlap with the other three parameters.

Within Q, the following categories of form-based errors are penalized:

- Orthographic errors: spelling mistakes, missing capitalization (e.g., *boulevard* instead of *Boulevard*), unnecessary bold/italics, or typos.
- Morphological errors: wrong tense or aspect (*he go* instead of *he goes*), incorrect plural or possessive forms, misused comparative/superlative forms, or incorrect derivations (e.g., *untidied* when no such word exists).
- Syntactic errors of form: wrong word order (*His eyes dark brown* instead of *His eyes were dark brown*), missing auxiliary verbs, missing subject/verb, etc.).
- Function word errors: wrong or missing articles, incorrect prepositions (*on the corner* instead of *at the corner*), or missing infinitive markers (*to* in *want to go*).

- Omissions of form: missing words or entire sentences/paragraphs are counted as major errors.

Crucially, Q focuses only on whether the linguistic form is acceptable in English. Errors that change meaning or appropriateness (e.g., false friends like *eventually* instead of *finally*) are not scored here, but rather under L.

The score is calculated proportionally to the total word count of each translation, which ensures comparability across works of different lengths. The formula is:

$$Q = \max \left(0.100 - \alpha \cdot \frac{1m+2d+4g}{w} \cdot 100\right)\left(0.100 - \alpha \cdot \frac{1m+2d+4g}{w} \cdot 100\right)$$

Where:

- *W* = total word count of the translation
- m = count of minor errors (weight 1): typographical, punctuation, or capitalization errors with no effect on meaning
- *d* = count of moderate errors (weight 2): local grammar or consistency issues with no effect on meaning
- *g* = count of major-form errors (weight 4): grammar/syntax errors affecting readability but not meaning
- *á* = penalty factor, set by default at 1 (points deducted per weighted error per 100 words)

By tying the penalty to the text length and scaling error severity, Q provides a transparent and replicable measure of form-based accuracy. Importantly, it does not evaluate the meaning *per se*, contextual appropriateness, or stylistic choices, thereby avoiding an overlap with the other three scoring parameters. Nevertheless, it is important to point out that if a form error also distorts the meaning, it is additionally penalized in L. This reflects the dual impact of the mistake on both the accuracy of form and accuracy of meaning.

3.3. Linguistic Score (L)

The Linguistic Score (L) measures whether the translation successfully conveys the intended *meaning* of the source text in a semantically accurate way. Unlike Q, which focuses on the mechanical form, L evaluates whether the structures and choices used preserve the informational content, logical relationships, and overall comprehensibility of the text.

Within L, the following meaning-based errors are penalized: mistranslations (changing the truth of the source segment), omissions, additions, wrong numbers/names/dates, wrong polarity/modality, misinterpreted idioms/phrases.

$$L = \left(1 - \frac{\Sigma deductions across all segments}{Number of segments}\right) \times 100\left(1 - \frac{\Sigma deductions across all segments}{Number of segments}\right) \times 100$$

The formula for L score is as follows:

For the purpose of evaluating Linguistic Score (L), the source text was segmented at the sentence level (i.e., sentences as taken from the source text, making that a fixed number across all translations), and each sentence is rated as:

- Fully correct (0 deduction) – Proposition and nuances fully preserved.
- Micro-deviation (0.25 deduction) – Tiny lexical nuance is off, but the proposition stays intact.
  e.g., painting → picture (when the medium is not crucial)
- Partially correct (0.5 deduction) – Part of the proposition is off or ambiguous; readers can still recover the main idea.
- Incorrect (1 deduction) – Core meaning wrong/missing; reader misled/ cannot recover.
  e.g., eventually for finally; missing sentence/paragraph; wrong conditional type flipping timeframe, etc.

Sentences were chosen as the unit of analysis to ensure that evaluators can judge whether the intended meaning has been fully conveyed or distorted without having to weigh ambiguous units. Using sentences rather than clauses or larger text chunks also makes the assessment more transparent and replicable. Every translation of the same source has the same number of units, which helps maintain comparability across student translations and the ChatGPT output. While clause-level segmentation could capture finer nuances, sentence-level segmentation strikes a balance between granularity and practicality, avoiding unnecessary complexity while still identifying distortions of meaning.

3.4. Pragmatic Score (P)

The Pragmatic Score (P) evaluates whether the translation is appropriate within its contextual and cultural setting. A simple checklist method is used as a calculation formula for the Pragmatic Score. For each text, evaluators check whether 5 predefined items are achieved and the formula is as follows:

$$P = \frac{T}{A} \times 100$$

Where:
T = total number of checklist items (default = 5)
A= number of items achieving a "Yes" response
Default checklist (1 point each):

1. Register/genre appropriate (formal/informal/academic as required)
2. Idioms and culture-specific references localized appropriately
3. Conventions localized (dates, units, currency, decimal separators)
4. Names, numbers, and titles accurate
5. Pragmatic intent preserved (politeness, hedging, tone)

3.5. Stylistic Score (S)

The Stylistic Score (S) captures qualities of style, readability, and fluency. It recognizes whether the translation reads naturally in the target language and whether it maintains stylistic cohesion with the source genre. S is assessed holistically on a 0–5 scale, then translated to a percentage:

$$S = \frac{r}{Rmax} \times 100$$

Where:

- r = holistic style rating assigned (0–5 scale)
- $Rmax$ = 5 = maximum rating

Holistic style scale:

- 5 = Excellent (natural, cohesive, varied, target-language-like)
- 4 = Good (minor awkwardness, overall smooth)
- 3 = Adequate (readable but clunky or repetitive)
- 2 = Weak (frequent awkwardness, poor cohesion)
- 1 = Poor (hard to follow, unnatural)
- 0 = Unacceptable (not fluent or incoherent)

3.6. AI system setup and generation conditions

The AI-generated translation analyzed in this study was produced by ChatGPT, a large language model developed by OpenAI, using a GPT-4–based version available via the publicly accessible interface at the time of data collection (March 2025). The system was provided with the same source text as the student participants and prompted to translate the text from Serbian into English, without additional contextual guidance or stylistic constraints. Default generation parameters were used, and no iterative prompting or prompt refinement was applied. Only a single AI output was generated and analyzed, rather than averaging multiple outputs, as the aim of the study was to compare student translation performance with a representative AI translation produced under typical academic use conditions. No post-editing, correction, or human intervention was applied to the AI-generated translation prior to evaluation. This design choice ensures methodological transparency and allows for a direct comparison between unedited human and AI translations within the same evaluation framework.

3.7. Evaluation procedure and reliability

All translations included in the corpus were evaluated using the same Translation Quality Index (TQI) framework and scoring rubric described in Sections 3.2–3.5. To ensure consistent categorization of errors across evaluators, explicit decision rules were applied to determine whether an error affected form (Q), meaning (L), or both; errors were penalized in both categories only when they demonstrably compromised grammatical well-formedness and semantic accuracy. The evaluation was conducted by two evaluators with expertise in the English language and translation studies. Both evaluators independently assessed the translations across

all four parameters (Q, L, P, and S), applying identical criteria and scoring procedures to both student translations and the AI-generated output. The evaluation was carried out blindly, in the sense that evaluators assessed the translations without being informed of whether a given text was produced by a student or by the AI system. Following the independent scoring phase, the evaluators compared their assessments and discussed cases of divergence to reach a consensus score. Overall agreement between evaluators was high, with discrepancies occurring primarily in borderline cases involving stylistic and pragmatic judgments. These were resolved through discussion, based on the predefined evaluation criteria. This procedure was adopted to ensure consistency, transparency, and reliability of the evaluation process while maintaining feasibility within the scope of the study.

3.8. Methodological limitations

Despite its systematic design, the present study is subject to several methodological limitations that should be acknowledged. First, the corpus consists of fifty student translations and a single AI-generated output, which limits the statistical generalizability of the findings and reflects the behavior of the AI system under the specific conditions of the task rather than its full range of potential variability. Second, the analysis is restricted to a single translation direction (Serbian into English), and as translation quality and error distribution may vary depending on language pair and directionality, the results cannot be automatically extended to other translation contexts. In addition, sentence-level segmentation may occasionally overlook extrasentential errors, such as misused tenses or connectors within complex sentences. However, this approach was adopted deliberately for reasons of consistency, feasibility, and reliability, ensuring that the evaluation remains both accurate and manageable across all translations. Finally, some errors that directly alter grammatical form (e.g., malformed conditionals or missing verbs) are counted both under the Quantitative Score (Q) and the Linguistic Score (L), as they simultaneously affect sentence well-formedness and semantic faithfulness. This double penalization is intentional: since form and meaning, while closely related, are treated as analytically distinct parameters within the TQI framework, counting such errors in both categories avoids underestimating their severity and allows the scoring model to indicate precisely where breakdowns in both grammaticality and meaning occur. These limitations do not undermine the internal validity of the study; rather, they suggest that the findings should be interpreted as indicative rather than exhaustive. Future research may address these constraints by expanding the corpus, including multiple AI outputs, and examining additional language pairs and translation directions.

3.9. Research hypotheses

This study is based on two hypotheses:

- Students outperform ChatGPT when it comes to pragmatics and stylistics, as this LLM tool focuses on literal translation, rather than on cultural values and pragmatic meaning.

- ChatGPT provides a consistently accurate translation across the realms of formal and linguistic accuracy.

4. Results and interpretation

The aggregated results for 50 student translations and one ChatGPT output reveal distinct patterns of strengths and weaknesses across the four evaluation parameters (Q, L, P, S) and in the overall Translation Quality Index (TQI). The scores for students were as follows**:** Q = 93.18, L = 82.16, P= 85.20, S = 81.20, with a final TQI = 85.08. By contrast, the ChatGPT output achieved Q = 97.70, L = 84.09, P = 80.00, S = 80.00, and a final TQI = 84.77.

As shown in Table 1, overall TQI scores for student translations and the AI-generated translation are comparable, while notable differences emerge across individual evaluation categories.

Translator	QQ	LL	PP	SS	TQI
Students (N=50)	93.18	82.16	85.20	81.20	85.08
ChatGPT (GPT-4)	97.70	84.09	80.00	80.00	84.77

Table 1. Mean scores by evaluation category

These results indicate that overall performance is relatively close (students = 85.08 vs ChatGPT = 84.77 in TQI), yet the distribution across the four parameters demonstrates complementary strengths. ChatGPT outperformed students in the category of the Quantitative Score (Q) and Linguistic Score (L), whereas students outperformed ChatGPT in the two remaining categories, Pragmatic Score (P) and Stylistic Score (S). Moreover, ChatGPT clearly outperformed students in the Quantitative Score (Q), achieving an average of 97.7 compared to 93.18 for student translators. ChatGPT's advantage in Q arises from its consistency in producing grammatically well-formed English sentences without the small but frequent slips that students exhibited. It consistently supplied correct articles, maintained subject–verb agreement, and handled auxiliary verbs without omission. Unlike student translations, it avoided errors such as:

- Wrong article usage (e.g. *a exterior* instead of *the exterior*, or inserting an article before a demonstrative pronoun such as *the this picture*).
- Word order errors (e.g., inversion errors like *how would it look like* instead of *how it would look like* in a declarative statement).
- Verb form errors, including wrong tense selection, particularly in terms of the sequence of the tenses.
- Morphological slips such as creation of non-existent words (e.g., *untidied* instead of *unkempt* to refer to one's appearance), or confusion of families (e.g., *effect* vs *affect*).
- Missing objects where students occasionally omitted obligatory objects after transitive verbs, leaving sentences incomplete or ungrammatical (e.g., *It*

serves as an arts museum and tourists often visit instead of *It serves as an arts museum and tourists often visit it*).

- Prepositional misuse in form: *passion towards painting* instead of *passion for painting*.
- Omissions of larger units where some student translations excluded entire words, sentences, or even a paragraph, which is heavily penalised under Q.
- Spelling errors in words such as *turquoise, exhibit,* and missing hyphens in compound adjectives such as *three-story, dark-haired, light-blue,* etc.

On the other hand, in terms of form, a notable pattern emerged from the analysis of the student output. While ChatGPT consistently outperformed students in surface-level form accuracy, it still struggled when translation demanded critical thinking and a deeper command of both languages. For instance, it rendered the one-word Serbian sentence *Visok.* too literally as *Tall.* Although this output is grammatically valid in Serbian, English requires a subject–verb structure (*He is tall*) for the sentence to be well-formed. This example illustrates that ChatGPT's advantage in form accuracy holds primarily at the level of orthography, morphology, and word order, but when the task required structural reconstruction across languages, i.e. moving beyond a literal surface rendering, it failed to provide a fully grammatical and natural solution. Conversely, more than half of the students successfully identified the structural difference in syntax and inserted a subject and a verb, thus creating a natural English sentence. The pattern suggests that ChatGPT excels at form when form is local and mechanical but remains weaker when form is entangled with meaning and context, requiring the kind of bilingual interpretive judgment that advanced human translators provide.

When it comes to the Linguistic Score (L), students achieved a score of 82.16, while ChatGPT performed slightly better with a score of 84.09. This reflects its consistent ability to produce meaning-preserving and grammatically correct sentences across the text, particularly in constructions where advanced learners often falter. ChatGPT has displayed strength in handling complex verb structures, such as the sequence of tenses and conditional sentences, the change of which leads to a change in meaning. For example, while some students produced sentences such as *I would have stayed if I had enough time* (incorrect omission of the past perfect), ChatGPT reliably produced the correct structure *I would have stayed if I had had enough time*. Similarly, in cases involving the infinitive versus gerund (*stopped to by* vs *stopped buying*) or modal auxiliaries (*could*, *would*, *might*), ChatGPT produced the expected meaning with fewer deviations than students. ChatGPT's advantage is also showcased in its rendering of prepositional meaning in the phrase *in the painting* versus *on the painting*. The distinction is critical: *in the painting* refers to something depicted within the artwork, whereas *on the painting* would suggest something physically lying on its surface, which is an altogether different meaning. Several students failed to capture this nuance and produced the less accurate phrase *on the painting*, while ChatGPT consistently selected the correct idiomatic form *in the painting*. This demonstrates that, despite its tendency toward literalism,

the system sometimes achieved greater semantic precision in collocations than human translators. Although ChatGPT's score is evidently higher here, students also demonstrated competence in capturing the intended meaning of the source text in nuanced constructions. Their innate human interpretive ability helped preserve idiomatic equivalence and avoid certain mistranslations of false friends. One illustrative example is its handling of the Serbian adverb *eventualno*. ChatGPT rendered it as *eventually* in English, a false friend, instead of the correct equivalents, such as *finally* or *perhaps* depending on context. This literal transfer distorted the meaning and demonstrated that, despite producing structurally correct sentences, the system struggled with semantic nuance and pragmatic appropriateness. By contrast, most student translators avoided such false-friend errors and demonstrated a stronger awareness of lexical nuance. Overall, while ChatGPT's higher numerical score reflects its generally reliable meaning-related accuracy, human translators showed superior judgment in capturing subtle meanings. This further suggests that while ChatGPT did outperform students in the domain of Linguistic Score, students' interpretive judgment often produced more semantically accurate results where ChatGPT could not deliver.

Regarding the Pragmatic Score (P), students scored an average of 85.20, clearly surpassing ChatGPT, which reached 80.00, marking this as the area where students excelled most clearly. They successfully adapted cultural references, localized names (e.g., street names), and maintained registers more consistently. ChatGPT, while accurate at the sentence level, often rendered culture-specific items too literally, reducing naturalness and appropriateness in the target language. While ChatGPT generally produced grammatically correct sentences, it often failed to adjust cultural references, stylistic conventions, and pragmatic tone in a way that would be natural to an English-speaking audience. For instance, a recurring issue was the translation of proper names and culturally bound references. ChatGPT tended to render place and street names in overly literal or hybrid forms, such as *King Alexander* instead of *Kralja Aleksandra Street*. While formally correct, such translations deviate from the established conventions. Students, on the other hand, more often recognized the need for context-specific localization, providing idiomatic and culturally appropriate forms. Finally, ChatGPT sometimes overlooked pragmatic markers such as hedging, modality, or politeness strategies, resulting in translations that, while correct at the surface level, lacked the subtlety necessary to convey the author's intent. Students generally handled these features better, reflecting both their training and their awareness of cross-cultural communication norms. Taken together, these findings explain why students outperformed ChatGPT in the pragmatics category. The system's reliance on statistical patterning and literal rendering makes it strong in sentence-level accuracy but weaker in adapting to the socio-cultural and pragmatic environment of the target language. By contrast, human translators naturally employ interpretive judgment and cultural awareness, which allow them to produce translations that not only make sense but also make sense in context.

As for the Stylistic Score (S), students achieved an average score of 81.20, maintaining a small but noteworthy advantage over ChatGPT, which scored 80.00. Students were slightly better at preserving cohesion, coherence, and natural flow. Human translators often achieved smoother paragraph transitions and more idiomatic phrasing, whereas ChatGPT occasionally produced stylistically flat or overly literal segments, though mechanically accurate. While both achieved relatively close scores, the nature of their stylistic performance differed slightly. Consistency and fluency at the sentence level could be highlighted as ChatGPT's strong points. It rarely produced awkward or ungrammatical renderings, and it ensured smooth word-to-word transitions. This was especially visible in paragraphs where students occasionally introduced stylistically vague renderings (*things* instead of precise, more localized nouns) or slightly clumsy collocations (*draw a painting* instead of *paint a picture*). ChatGPT avoided such imprecisions and maintained a uniform register throughout its output. Its style was stable and error-free, making it easy to read. However, ChatGPT's weakness lies in its tendency toward literalism and stylistic flatness. For example, its occasional literal treatment of expressions (e.g. *the exterior of it* instead of *its exterior*) weakened stylistic performance. Students, on the other hand, demonstrated more stylistically aware choices. They more often supplied the missing elements that English requires for natural phrasing, avoided one-word "sentences," and achieved smoother cohesion between paragraphs. They also showed greater flexibility in choosing literary synonyms (e.g., *painting* instead of the more generic *picture*), which added stylistic depth. In short, ChatGPT produced consistently fluent but stylistically flat translations, while students produced more idiomatic and nuanced texts that were occasionally disrupted by imprecise lexical choices or structural slips. The narrow margin between the two scores reflects this balance: ChatGPT is strong in surface fluency, while students excel in stylistic richness and naturalness, even if their execution is not entirely error-free.

5. Conclusion

The integration of digital technologies into higher education has reshaped language teaching practices and instructional design. Studies conducted at the university level indicate that both teachers and institutions recognize the pedagogical value of ICT in foreign language teaching, particularly when technology is integrated as a complement to established instructional practices rather than as a substitute for human expertise (Veljković Michos/Bošković Marković, 2020; Veljković Michos/Bošković Marković, 2021). Building on this foundation, recent research has begun to explore the role of more advanced AI-based tools, including chatbots and large language models, in academic contexts.

Inferring from the aggregated results obtained from a blend of quantitative and qualitative scores and their subsequent analysis, it can be stated that the initial hypothesis of this research has evidently been reinforced and confirmed. Students did outperform ChatGPT when it came to pragmatics and stylistics – fields that inherently require a comprehensive, multidimensional, non-literal approach

that customarily needs to surpass the sentential level to yield a corresponding translation result. Even though ChatGPT employs a deep learning artificial neural network, which constitutes the most sophisticated model that has provided impressive results, especially when compared to all other previous models, it still occasionally falters in producing a satisfactory result in the aforementioned realms, underscoring its tendencies towards literalism. Furthermore, as it has been initially suggested by our hypothesis, ChatGPT did provide a consistently accurate translation across the realms of formal and linguistic accuracy. This underscored its strength in generating an almost faultless result when it came to surface-level tasks and when it came to producing a sentence that was both grammatically and semantically correct, i.e. a sentence that adhered to all grammar norms and simultaneously preserved the intended meaning.

On this basis, the main conclusion that can be drawn is that the highest quality translation output can be obtained when human expertise is combined with machine-aided translation. The research successfully highlighted the respective areas of both human translators' and ChatGPT's strengths, thus suggesting a hybrid model in which human expertise and artificial intelligence operate in a complementary manner. While AI systems excel in formal precision and structural consistency, human translators contribute with interpretive, pragmatic judgement, which are the competences that remain beyond the reach of current computational models. In this perspective, we strongly suggest that ChatGPT should not be regarded as a replacement for the human translator, but rather as a powerful assistive tool, the full potential of which is realized when combined with human-generated translation. Such a collaborative model offers the most promising path toward achieving both efficiency and quality in contemporary translation practice and holds important implications for future translator training and the evolving role of AI in professional translation.

At the same time, the findings of the present study should be interpreted in light of certain limitations. The analysis was conducted within a single institutional context and focused on student translators at a specific level of training, which limits the broader generalizability of the results. In addition, the study examined a single translation direction (Serbian into English) and a limited range of texts, with the AI system represented by a single unedited output. These constraints do not undermine the internal validity of the study, but they indicate that the conclusions are context-specific. Future research could extend this line of inquiry by including other text genres, additional language pairs and translation directions, and professional translators, as well as by analyzing multiple AI outputs, in order to further explore the evolving relationship between human and AI-assisted translation across diverse translation settings.

References

American Translators Association. (n.d.). *ATA certification program: Error marking framework.* American Translators Association. https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/

Automatic Language Processing Advisory Committee. (1966). *Language and machines: Computers in translation and linguistics* (Publication No. 1416). National Academy of Sciences–National Research Council. https://www.mt-archive.net/50/ALPAC-1966.pdf

Gajiwala, C. (2025). The rise of deep learning and neural networks: Revolutionizing artificial intelligence. *European Journal of Computer Science and Information Technology, 13*(17), 88–99.

Hardmeier, C. (2015). On statistical machine translation and translation theory. In B. Webber, M. Carpuat, A. Popescu-Belis, C. Hardmeier (Eds.), *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)* (pp. 168–172). Lisbon: Association for Computational Linguistics.

House, J. (1997). *Translation quality assessment: A model revisited*. Tübingen: Gunter Narr Verlag.

Hutchins, W. J. (2004). The Georgetown–IBM experiment demonstrated in January 1954. *Machine translation: From real users to research (Lecture Notes in Computer Science), 3265*, 102–114. https://doi.org/10.1007/978-3-540-30194-3_12

Ilić, D., Bošković Marković, V. (2020). Teaching in the online classroom from the perspective of English language teachers. *Primenjena lingvistika, 21,* 25–44.

Koehn, P. (2009). *Statistical machine translation*. Cambridge: Cambridge University Press.

Lee, J., Liao, P. (2011). A comparative study of human translation and machine translation with post-editing. *Compilation and Translation Review, 4*(2), 105–149.

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

McCarthy, J., Minsky, M., Rochester, N., Shannon, C. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Hanover: Dartmouth College.

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

Nord, C. (1997). *Translating as a purposeful activity: Functionalist approaches explained*. Manchester: St. Jerome Publishing.

Pantelić, N., Milošević, M., Bošković Marković, V. (2023). Using AI chatbots in academia: The opinions of university students. In M. Stanišić (Ed.), *Sinteza 2023 – International Scientific Conference on Information Technology and Data Related Research* (pp. 306–311). Belgrade: Singidunum University.

Pym, A. (2010). *Exploring translation theories.* London: Routledge.

Shahmerdanova, R. (2025). Artificial intelligence in translation: Challenges and opportunities. *Acta Globalis Humanitatis et Linguarum, 2*(1), 62–70. https://doi.org/10.69760/aghel.02500108

Veljković Michos, M., Bošković Marković, V. (2020). Teachers' perception of the use of ICT in foreign language teaching at a higher education institution. In M. Stanišić (Ed.), *Sinteza 2020 – International Scientific Conference on Information Technology and Data Related Research* (pp. 93–98). Belgrade: Singidunum University.

Veljković Michos, M., Bošković Marković, V. (2021). Informaciono-komunikacione tehnologije u nastavi stranih jezika pre i tokom pandemije virusa COVID-19: Pregled istraživanja na univerzitetskom nivou. *Anali Filološkog fakulteta, 33*(2), 135–151.

Emilija Z. Mitić
Valentina B. Bošković Marković

Sažetak

KOMPARATIVNA ANALIZA PREVODA SA SRPSKOG NA ENGLESKI: POREĐENJE SISTEMA ZASNOVANOG NA VELIKOM JEZIČKOM MODELU I STUDENTSKIH ZADATAKA

Ovaj rad se bavi komparativnom analizom kvaliteta prevoda sa srpskog na engleski jezik koje su izradili studenti osnovnih studija i sistem zasnovan na velikom jezičkom modelu, predstavljen kroz ChatGPT. Istraživanje je sprovedeno u kontekstu sve izraženije upotrebe veštačke inteligencije, neuronskog mašinskog prevođenja i velikih jezičkih modela u savremenoj prevodilačkoj praksi i nastavi stranih jezika. Osnovni cilj rada jeste da se ispita u kojoj meri se prevodi koje proizvodi ChatGPT razlikuju od studentskih prevoda u pogledu formalne tačnosti, očuvanja značenja, pragmatičke primerenosti i stilskog kvaliteta. Polazište istraživanja čini pretpostavka da ChatGPT ostvaruje visok stepen formalne i jezičke tačnosti, ali da studenti mogu pokazati veću uspešnost u oblastima koje zahtevaju kulturnu interpretaciju, pragmatičko rasuđivanje i stilsku prilagodljivost. U tom smislu, rad ne posmatra veštačku inteligenciju isključivo kao konkurenciju ljudskom prevodiocu, već kao potencijalno korisno pomoćno sredstvo čija se vrednost najbolje sagledava u odnosu na ljudsku stručnost. Podaci su prikupljeni tako što su isti prevodilački zadaci dodeljeni grupi studenata druge godine Departmana za engleski jezik Univerziteta Singidunum i sistemu ChatGPT. Analiziran je korpus od 50 studentskih prevoda i jednog prevoda generisanog korišćenjem veštačke inteligencije. Za procenu kvaliteta korišćen je *Translation Quality Index*, odnosno indeks kvaliteta prevoda, koji obuhvata četiri komponente: kvantitativni skor, jezički

skor, pragmatički skor i stilski skor. Kvantitativni aspekt analize obuhvatio je greške u formi, gramatici, pravopisu i strukturi, dok su kvalitativni aspekti uključili procenu tačnosti značenja, idiomatskih izraza, kulturnih referenci, registra, kontekstualne primerenosti, fluentnosti i čitljivosti. Rezultati istraživanja pokazuju da su ukupni skorovi studentskih i prevoda generisanog uz pomoć veštačke inteligencije veoma bliski. Studenti su ostvarili ukupni TQI skor od 85,08, dok je ChatGPT ostvario skor od 84,77. Ipak, razlike između pojedinačnih kategorija ukazuju na komplementarne prednosti. ChatGPT je bio uspešniji u domenu formalne i jezičke tačnosti, naročito u pogledu gramatičke pravilnosti, doslednosti i rečenične strukture. Nasuprot tome, studenti su ostvarili bolje rezultate u pragmatičkoj i stilskoj dimenziji, posebno u prepoznavanju kulturnospecifičnih elemenata, idiomatskih značenja, registra i prirodnijeg izraza na ciljnom jeziku. Zaključuje se da najkvalitetniji prevodilački rezultat može nastati kombinovanjem prednosti mašinski potpomognutog prevođenja i ljudske prevodilačke kompetencije. ChatGPT može doprineti efikasnosti, formalnoj preciznosti i jezičkoj konzistentnosti, dok ljudi kao prevodioci ostaju nezamenljivi u oblastima koje zahtevaju interpretativno rasuđivanje, kulturnu osetljivost i stilsku nijansu. Rezultati istraživanja imaju značajne implikacije za nastavu prevođenja, razvoj prevodilačkih kompetencija i buduće promišljanje uloge veštačke inteligencije u akademskom i profesionalnom prevođenju.

Ključne reči:
veštačka inteligencija, LLM, indeks kvaliteta prevoda, ChatGPT, greška, stil

Appendices

Appendix A. Translation text

Na uglu ulice Kralja Aleksandra i Bulevara Miloša Obrenovića stajala je stara, trokatna zgrada. Baš taj deo Beograda pruža divan pogled, a sama ulica nosi ime jedne od najpoznatijih srpskih istorijskih ličnosti. Zgrada je stara stotinu godina i spoljni izgled joj se znatno promenio tokom vremena. Na fasadi zgrade se mogu primetiti tragovi vremenskih uticaja. Međutim, unutrašnjost je bila preuređena i modernizovana. Sada se koristi kao muzej umetnosti i turisti je često posećuju.

U muzeju se čuvaju vredne slike i skulpture, a mnoge od njih pripadaju jednom od najpoznatijih srpskih slikara. Sava Šumanović je proveo veći deo svog života slikajući prelepe pejzaže i portrete. Njegova najpoznatija slika prikazuje ženu sa smeškom na licu. Na slici, žena nosi dugu, prozračnu, svetloplavu haljinu, a u pozadini se pruža peskovita plaža i tirkizno more.

Sam slikar je bio muškarac kratke, crne kose. Visok. Brada mu je bila uvek dobro podšišana i uredna, kao da nikada nije želeo da dozvoli da ga vide neurednog. Oči su mu bile tamno smeđe, gotovo crne i zračile su strašću prema slikanju. Promućuran

i oštar pogled je odavao utisak osobenjaka. Međutim, on je bio topao čovek sa istančanim okom za detalje, zapravo, za njega, život je umetničko delo samo po sebi, a svaki trenutak je smatrao vrednim da se proslavi i podeli sa drugima.

Pored slika i skulptura, u muzeju se takođe čuvaju dragoceni predmeti koji su nekada pripadali vlasnicima ove stare zgrade. Neki od njih su uključivali porodične fotografije, pisma i lične dnevnike, to su stari dokumenti koji svedoče o životu ljudi koji su nekada živeli ovde.

Posetivši ovaj muzej, ostao sam fasciniran lepotom izloženih predmeta. Ostao bih unutra čitavu večnost, da sam imao dovoljno vremena. Gledao sam u te stare slike i zamišljao kako bi bilo živeti u to vreme. Eventualno, ako bih imao priliku da se vratim u prošlost, verovatno bih voleo da istražujem ove ulice i da se upoznam sa ljudima koji su nekada živeli ovde.