UDC 004.932:81’42:81’37

HYBRID TRANSLATION VERIFICATION WITH FIVE-LEVEL ERROR CLASSIFICATION

I. M. Liakh[1], R. Ya. Zhovtani[2], A. Yu. Tsipino[3], A. M. Roman[4]

1. Uzhhorod National University, Narodna Square, 3, Uzhhorod, 88000, Ukraine https://orcid.org/0000-0001-5417-9403 e-mail:igor.lyah@uzhnu.edu.ua
2. Uzhhorod National University, Narodna Square, 3, Uzhhorod, 88000, Ukraine https://orcid.org/0000-0002-7421-148X e-mail:ruslana.zhovtani@uzhnu.edu.ua
3. Uzhhorod National University, Narodna Square, 3, Uzhhorod, 88000, Ukraine https://orcid.org/0009-0007-1704-2732 e-mail:tsipino.artemii@student.uzhnu.edu.ua
4. Uzhhorod National University, Narodna Square, 3, Uzhhorod, 88000, Ukraine https://orcid.org/0009-0005-5774-8946 e-mail:roman.anhelina@student.uzhnu.edu.ua

The problem of systemic errors in modern machine translation systems based on neural network and transformer architectures is researched in the article. The aim of the work is to formalize a unified error classification model, develop an integrated indicator for evaluating translation quality, and create a conceptual scheme for a hybrid algorithm to minimize errors. For experimental analysis, a corpus of about 200 sentences was formed, covering technical, legal, literary, and colloquial texts, idiomatic expressions, complex syntactic constructions, passive forms, cases of homonymy, and polysemy. A comparative study was conducted based on the analysis of five machine translation systems: Google Translate, DeepL, Microsoft Translator, ChatGPT (GPT API), and MarianMT.

A five-level classification of errors is proposed, including lexical, syntactic, semantic, pragmatic deviations, and hallucinations. It has been established that traditional automatic metrics (BLEU, TER, METEOR) do not fully reflect deep semantic distortions, and therefore the COMET neural network metric and expert human evaluation have been integrated into the evaluation procedure. Based on a multifactorial approach, the Unified Translation Error Index (UTEI) has been developed, which combines normalized values of automatic metrics with a weighted sum of classified errors according to their criticality, ensuring multidimensionality and interpretability of results.

To minimize the detected deviations, a hybrid algorithm for universal translation verification is proposed, combining formalized linguistic verification with semantic validation based on neural network models. The algorithm uses embedding similarity analysis, syntactic and morphological correctness checks, and a back-translation mechanism to detect and correct deep semantic distortions. The proposed approach ensures the adaptability and scalability of the algorithm for integration into various NMT and LLM systems, increasing the structural stability, semantic consistency, and overall reliability of machine translation.

Keywords: *machine translation; neural network models; error classification; quality assessment; UTEI; hybrid algorithm.*

Problem statement. The issues of machine translation quality, in particular the typology of errors, their impact on the content adequacy and perception of the text, as well as the presence of linguistic and cultural biases, are actively researched within the framework of modern natural language processing. Scientific works cover both the algorithmic aspects of neural network systems and applied issues of machine translation use in educational, technical, legal, and multimedia environments, with a particular focus on automatic and expert evaluation methods. The growing role of large language models based on transformer architectures expands the possibilities for semantic adaptation, but at the same time raises issues of interpretability, stability of results, and the risk of hallucinations. This necessitates unified approaches to error classification and the development of integrated mechanisms for translation quality control.

Analysis of recent research and publications. In particular, the issue of automated detection and correction of errors in text data, relevant to machine translation tasks, is discussed in the work of R. B. Fedchuk and V. A. Vysotska. The authors analyze the combination of rule-based approaches, statistical models, and machine learning methods for detecting and correcting deviations, justifying the effectiveness of hybrid systems in improving the quality of text information processing. The principles they propose for integrating formalized linguistic rules with intelligent correction algorithms are conceptually consistent with the approach developed in this article for verifying translations and minimizing errors in machine translation systems [1].

A significant contribution to the study of modern natural language processing methods was made in the work of Iosifov I., and Sokolov V., where the relevance of the application of NLP is considered in connection with the rapid growth of text data in social networks, e-commerce and online media. The authors analyze the use of deep neural networks and transformer architectures, in particular in cybersecurity tasks, and outline the key problems of modern NLP systems, including multilingualism, high computational complexity, interpretability of models and the presence of language biases. These aspects confirm the relevance of further research aimed at increasing the reliability and neutrality of language technologies [2].

Muftah's research on the impact of artificial intelligence on translation processes and translator training is discussed in an article that analyzes the use of large language models, particularly ChatGPT, in translation practice. The authors investigated how students and professionals evaluate AI translations (ChatGPT-T) compared to professional human translations (HT) of technical and literary texts from German into Arabic and English. Using a combination of qualitative analysis (observation and annotation) and quantitative metrics (BLEU), it was found that the level of linguistic competence and professional experience significantly affect the ability to distinguish AI translation from human. In technical texts, ChatGPT often demonstrated results close to human ones, while in stylistically complex literary texts the differences were more noticeable. The study emphasizes that purely linguistic analysis is not enough to fully assess the quality of translation; cultural and stylistic aspects must be taken into account. Overall, the work shows that AI acts as a powerful but auxiliary tool that can increase the effectiveness of translation training and contribute to the development of key competencies of future specialists [3].

The study by Abdelhalim et al. analyzed the use of artificial intelligence tools, specifically ChatGPT and Google Translate, in literary translation by students of English as a foreign language (EFL). The authors compared the assessments of 63 students (27 beginners and 36 experienced) in terms of accuracy, fluency, and cultural adequacy of translations. The results showed that students noted the advantages of ChatGPT in preserving the stylistic and cultural features of texts, while Google Translate demonstrated significant limitations in this regard. The study emphasizes the importance of using AI in translator training and the need to integrate such technologies to improve translation quality and develop student's competencies [4].

In a study by Chang et al. on the use of Google Translate (GT) in teaching English as a foreign language (EFL), the impact of the system on error correction, accuracy, and complexity of students' written work was evaluated. The experimental group, which worked with recursive editing using GT throughout the semester, showed significant improvement in text accuracy and syntactic complexity compared to the control group. At the same time, the control group demonstrated better writing fluency. Overall, students rated the use of GT for learning positively, noting reduced anxiety, increased motivation, and support in completing tasks. The study highlights the potential of MT as a learning tool that promotes accuracy and linguistic competence but does not always increase overall writing speed [5].

In a study by Qiu and Pym on machine translation errors in subtitling, they analyzed the reactions of viewers with varying levels of knowledge of the source language. Nine participants did not know the original language, and seven were learning it; screen recordings, think-aloud protocols, comprehension tests, and interviews were used for evaluation. The authors identified which errors were most noticeable and how they affected trust in the subtitles and immersion in the viewing experience. Errors that cause significant misunderstanding are considered "fatal," but viewers show considerable tolerance, adjusting their perception using context or additional sources. The study emphasizes that the perception of errors in MT depends on knowledge of the source language and the balance between translation accuracy and enjoyment of the content [6].

Purpose of the article. The purpose of the article is to develop a unified model for classifying machine translation errors and to form an integrated quality assessment indicator that combines automatic metrics with an analysis of the criticality of detected deviations. An additional goal is to create a conceptual scheme for a hybrid translation verification algorithm that ensures structural correctness based on formalized rules and deep semantic consistency through neural network validation.

Presentation of the main research material. Machine translation systems, despite the rapid development of neural network architectures and transformer models, often make mistakes related to incorrect interpretation of context, ambiguity of lexical units, violation of syntactic structure, and distortion of semantic relationships between elements of a statement. Particularly critical are cases of loss of pragmatic meaning, incorrect transmission of specialized terminology, and the appearance of hallucinations—the addition of information that is absent in the original text. Despite the high values of automatic evaluation metrics, such errors can significantly affect the accuracy of content transmission, which necessitates their systematic analysis and the development of unified

approaches to classification and minimization. Five machine translation systems were selected for the study, representing different approaches to neural translation and covering both commercial and open-source solutions. The list includes systems such as Google Translate, DeepL, Microsoft Translator, ChatGPT (GPT API), and MarianMT [7].

Google Translate was chosen as one of the most widely used and technologically advanced systems, utilizing multilingual transformer models and large-scale pre-training [8]. Its inclusion allows for the evaluation of errors in the context of the widest possible application and diversity of language pairs.

DeepL is characterized by high-quality translation of European languages and is positioned as a system with improved contextual processing [9]. The inclusion of DeepL allows us to compare the behavior of models optimized for stylistic and semantic quality. *Microsoft Translator* represents an integrated translation ecosystem focused on corporate use and cloud services [10]. Its analysis allows us to evaluate the stability of translation in technical and specialized texts.

ChatGPT (GPT API) was included as an example of a large general-purpose language model that is not a specialized translator but demonstrates competitive translation quality thanks to instruction-tuning and contextual modeling [11]. This makes it possible to compare classic NMT systems with the LLM approach.

MarianMT was chosen as an open-source solution, allowing the model's behavior to be studied without the influence of closed commercial optimization [12]. Its inclusion ensures transparency of the experiment and the possibility of reproducibility of the results.

Thus, the formed set of systems covers various architectural approaches, scaling levels, degrees of commercial optimization, and source code availability, which ensures the representativeness and objectivity of further error analysis. To ensure the representativeness of the experiment, a test corpus of 200 sentences was formed, covering various thematic domains, stylistic registers, and structural features of speech. This volume allows for statistical reliability of the analysis and, at the same time, enables a detailed manual expert evaluation.

The corpus was formed on the basis of:

– hand-selected examples;

– open parallel corpora (Tatoeba, Europarl, WMT);

– specially constructed sentences for testing specific linguistic phenomena.

The inclusion of various topics and types of constructions is due to the need to test machine translation systems in terms of:

1. Terminological accuracy (technical and legal texts).

2. Contextual ambiguity (homonymy, polysemy).

3. Complex syntactic organization (subordinate constructions, inversions, passive forms).

4. Pragmatic adaptation (colloquial style, idiomatic expressions).

5. Semantic resistance to ambiguity (context-dependent meanings).

This approach allows us to identify not only superficial lexical errors, but also deep semantic, pragmatic and discursive errors.

Table 1

Corpus Structure

Text type	Purpose of inclusion	Example sentence
Technical	Checking the accuracy of terminology and structural formalization	The neural network was trained using backpropagation with adaptive learning rates.
Legal	Analysis of complex formulations and normative vocabulary	The contract shall enter into force upon signature by both parties.
Literary	Checking stylistic and figurative expression	The wind whispered secrets through the ancient trees.
Informal	Pragmatic adaptation assessment	Are you kidding me right now?
Idioms	Identifying literal translations of fixed expressions	He kicked the bucket yesterday.
Complex syntactic constructions	Analysis of embedded subordinating and logical connections	The system that was designed by researchers who previously worked on distributed models failed unexpectedly.
Passive forms	Checking the correctness of grammatical transformation	The data were processed and validated before publication.
Homonymy	Testing contextual value selection	The bank raised the interest rate.
Polysemy	Analysis of polysemic units	She saw the man with the telescope.

The quantitative distribution of the test corpus was formed taking into account the need for uniform representation of different functional styles and linguistic phenomena. Technical and scientific texts account for the largest share (25%), as it is in these texts that the accuracy of terminology, the correctness of formalized structures and the preservation of logical connections are critical. Such texts allow us to assess the ability of systems to adequately reproduce specialized vocabulary and complex subject concepts.

Legal texts make up 20% of the corpus. Their inclusion is due to their high level of syntactic complexity, the presence of normative formulations and specific vocabulary, where even a slight deviation can lead to distortion of the meaning. Analysis of the translation of such sentences allows us to identify errors related to modality, conditional constructions, and formal business style.

Literary texts (15%) are integrated to assess the ability of systems to convey imagery, metaphorical language and stylistic nuances. Semantic flexibility and contextual sensitivity of models are particularly important in this segment. Conversational sentences (15%) are designed to test the pragmatic adaptation of the translation, including informal expressions, abbreviations and emotionally charged vocabulary.

Idiomatic constructions account for 10% of the corpus, as they are one of the most vulnerable segments for machine translation systems. Their analysis allows us to determine the tendency of models to translate literally and to assess the level of contextual generalization. Separately, 15% are sentences with complex syntactic and grammatical structures, including nested subordinate constructions, passive forms, homonymy, and polysemy. This segment is aimed at studying the impact of structural complexity on translation stability.

These categories overlap to some extent, reflecting the real nature of speech, in which thematic, stylistic and grammatical characteristics often coexist within a single utterance.

The proposed corpus structure creates the conditions for multidimensional analysis of machine translation errors. This approach allows not only to record the frequency of individual types of errors, but also to investigate their dependence on stylistic register, thematic domain, and syntactic complexity.

Comparative analysis of translations in different functional styles makes it possible to determine how stable the models are in formal and informal contexts. Studying sentences with increasing structural complexity allows us to assess the limits of the generalization ability of transformer architectures and their resistance to long dependencies. In addition, the inclusion of examples with homonymy and polysemy creates conditions for testing deep contextual understanding, which is a key indicator of the semantic adequacy of a translation.

Therefore, the formed corpus not only ensures the representativeness of the experiment, but also creates a methodological basis for the further development of algorithms for detecting and minimizing machine translation errors.

In order to systematize the identified deviations, a unified error classification model was developed, which integrates linguistic, semantic and pragmatic levels of analysis. The proposed model allows not only to record individual cases of incorrect translation, but also to determine their nature, depth and potential impact on the content of the statement.

The classification is based on a hierarchical principle and covers five main categories: lexical, syntactic, semantic, pragmatic errors, and hallucinations [13, 14].

Lexical errors are associated with the incorrect selection or reproduction of individual language units. This category includes cases of incorrect selection of the meaning of a polysemous word, omission of a specialized term or its replacement with a generalized equivalent, as well as excessive generalization, which leads to a loss of accuracy. Such errors are particularly critical in technical and legal texts, where terminological correctness determines the accuracy of the content.

Syntactic errors occur at the level of sentence structure. These include word order violations, subject-predicate agreement errors, incorrect subordinate clause construction, or incorrect transformation of passive forms. Such deviations may not always distort the meaning, but they significantly reduce the grammatical acceptability of the translation and affect its formal quality.

Semantic errors are characterised by a profound distortion of the meaning of a statement. They manifest themselves in changes in logical relationships, the inversion of cause-and-effect relationships, the loss of modality, or the distortion of key concepts. Unlike lexical errors, semantic deviations affect the integrity of the message and can lead to a fundamentally incorrect interpretation of the text.

Pragmatic errors are related to the inconsistency of the translation with the communicative situation. These include incorrect style, register violation, incorrect transmission of emotional coloring or culturally determined features. Such errors most often occur in informal and literary texts and indicate limited contextual adaptation of the model.

Hallucinations constitute a separate category – cases of adding information that is absent in the original text or generating logically unfounded fragments. This type of error is typical of large language models and poses an increased risk in technical or regulatory documents, where even a minor addition can change the meaning.

The proposed five-level taxonomy of machine translation errors was systematically validated to confirm its conceptual and empirical reliability. Experts classified 200 sentences, 154 of which contained errors of various types. Each sentence was assigned only one dominant error category, which ensured the purity of the methodological approach.

Table 2

Frequency of Errors

Error category	Subtype	Number of cases	Percentage of the total number (%)
Lexical (Lex)	Incorrect choice of the meaning	32	21%
Lexical (Lex)	Omission of a term	18	12%
Syntactic (Syn)	Word order violation	27	18%
Semantic (Sem)	Distortion of the meaning	22	14%
Pragmatic (Prag)	Style violation	19	12%
Hallucinations (Hall)	Addition of the information	15	9%
Other	Combined errors	21	14%

The results of the inter-expert analysis showed a high level of agreement. Cohen's kappa coefficient, calculated based on the agreement matrix, made 0.829, which corresponds to the "almost perfect agreement" category on the Landis & Koch scale, demonstrating almost complete congruence between the experts' assessments. This indicator confirms that the distinction between lexical, syntactic, semantic, pragmatic errors and hallucinations has clear boundaries, and the proposed taxonomy is reproducible and methodologically sound. The recognition of semantic errors, and hallucinations, which in modern large language models (LLMs) pose the most significant risk of content distortion, proved to be remarkably stable.

Table 3

Matrix of Inter-Expert Agreement

	Lex	Syn	Sem	Prag	Hall	Other	Σ
Lex	50	2	1	1	0	6	60
Syn	3	27	2	1	0	5	38
Sem	1	2	22	0	0	4	29
Prag	1	1	0	19	0	3	24
Hall	0	0	1	0	15	2	18
Other	4	3	3	2	1	8	21
Σ	59	35	29	23	16	28	*190*

The sum of the values of the main diagonal (correct category matches), as well as "Others" are not included in the comparison between experts (standard practice), then the calculation is performed according to five core categories:

$$P_o = \frac{50+27+22+19+15}{154} \approx 0.864 \quad (1)$$

The expected value is defined as:

$$P_e = \sum_{i=1}^{k} \left(\frac{row_i}{N} \cdot \frac{col_i}{N} \right) \quad (2)$$

Since this is a symmetric matrix, it has equal marginal sums (single marking), then:

$$P_e = \left(\frac{50}{154}\right)^2 + \left(\frac{27}{154}\right)^2 + \left(\frac{22}{154}\right)^2 + \left(\frac{19}{154}\right)^2 + \left(\frac{15}{154}\right)^2 + \left(\frac{8}{154}\right)^2 \approx 0.184$$

The Kappa coefficient is calculated using the classic formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0.864 - 0.184}{1 - 0.184} \approx 0.8333 \quad (3)$$

To provide the additional methodological verification, the proposed taxonomy was compared with international standards for assessing translation quality, such as MQM (Multidimensional Quality Metrics), DQF (Dynamic Quality Framework), and SAE J2450 [15-17]. The analysis showed structural compatibility among the categories: lexical and semantic errors correlate with the Accuracy subcategories in MQM and DQF, syntactic deviations correspond to the Fluency component, pragmatic errors correlate with the Style and Register categories, and hallucinations conceptually correspond to critical errors such as Addition or Non-translation. A distinctive feature of the proposed model is the identification of hallucinations as an independent category, justified by the specifics of LLM systems, where the generation of additional information non-available in the original constitutes a separate class of risk.

Table 4

Correlation between the Categories of the Proposed Error Taxonomy and International Standards for Translation Quality Assessment

Suggested category	MQM	DQF	SAE J2450
Lexical	Accuracy → Terminology	Accuracy	Terminology
Syntactic	Fluency → Grammar	Fluency	Grammar
Semantic	Accuracy → Mistranslation	Accuracy	Meaning
Pragmatic	Style / Register	Style	–
Hallucinations	Non-translation / Addition	Critical Error	Addition

Thus, the proposed model not only aligns with the existing international standards, but also extends them by adapting to the specific characteristics of new-generation neural network architectures. The held statistical validation and comparative analysis acknowledge that the taxonomy has a high level of inter-expert reproducibility, is structurally compatible with MQM, DQF, and SAE J2450, accounts for the specific characteristics of modern LLM systems, and is suitable for quantitative formalization, particularly, within the UTEI integral indicator.

This allows moving from a descriptive typology of errors to a formalized tool for assessing translation quality, creating a foundation for further analysis automation and minimization of translation deviations. The identification of hallucinations as a separate category is methodologically justified and relevant for working with modern language models, and the high coefficient of agreement ($\kappa = 0.8333$) indicates an almost complete inter-expert agreement. The proposed model integrates the critical aspects of international standards, while simultaneously adapting them to modern technological conditions, which makes it relevant both for scientific research and practical application in translation quality assessment.

The proposed scheme enables the transition from a subjective description of errors to a formalized analysis, establishing a basis for the quantitative assessment of translation quality. In addition, the unified taxonomy can be used to develop algorithms for automatic deviation detection, since each category correlates with a specific level of language structure, ranging from the lexical to the discourse level.

The machine translation quality assessment was carried out using a combined approach that integrates automatic metrics and expert human evaluation. Using a single metric, such as BLEU, does not fully reflect the real quality of a translation, as such indicators are focused mainly on surface n-gram correspondence and do not take into account deep semantic or pragmatic aspects [18]. In the study, four basic approaches to evaluation were applied: BLEU, TER, METEOR, COMET, and an expert human assessment.

BLEU (*Bilingual Evaluation Understudy*) was used as a baseline metric to ensure comparability of the results with previous studies. Despite its limitations, this metric allows for determining the degree of formal closeness of the translation to the reference version.

TER (*Translation Edit Rate*) is employed to estimate the number of edits required to align a machine translation with the reference translation. Unlike BLEU, TER allows estimating the actual “edit distance”, and is more sensitive to structural variations.

METEOR integrates lexical variation, synonymy, and morphological congruence, making it more flexible than BLEU. The use of this metric allows for partial consideration of semantic equivalence, rather than literal correspondence.

COMET was chosen as a modern neural network metric based on transformational models and which takes into account contextual representations of the text. It demonstrates a high correlation with human evaluation and enables evaluation of the translation at the level of semantic adequacy.

The automatic metrics were supplemented by expert evaluation conducted by independent annotators based on the criteria of adequacy, accuracy, and stylistic conformity. To enhance the reliability of the results, the use of an inter-expert agreement coefficient (e.g., Cohen’s kappa) is planned to be applied, which allows minimizing the subjective influence of the assessment.

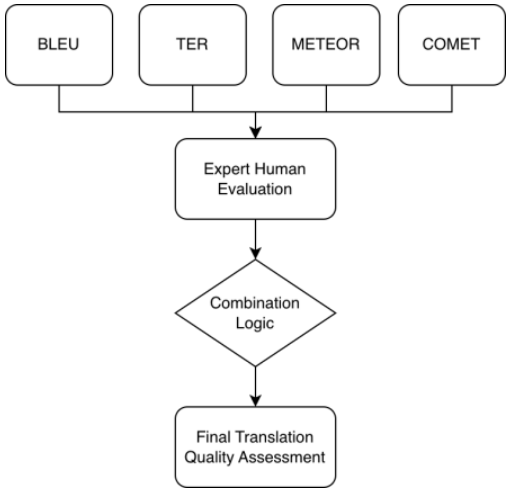

Fig 1. Scheme for using basic approaches to evaluation

The proposed combined evaluation approach has enabled the obtaining of generalized quantitative results for every machine translation system under study. The comparative analysis has shown that the highest level of semantic consistency according to the COMET metric was demonstrated by the ChatGPT system (0.76), which correlates with the maximum value of the UTEI integral index (0.72). DeepL demonstrated consistently high results according to the traditional metrics (BLEU - 0.47; METEOR - 0.52; TER - 0.44), however, its integral index (0.69) turned out to be slightly lower due to a higher frequency of semantic deviations in complex contexts. Google Translate and Microsoft Translator demonstrated average scores across both automatic metrics and UTEI, indicating sufficient formal relevance of the translation despite the structural and lexical errors. The lowest values of the integral index were recorded for MarianMT (0.53), which may be connected with its lower adaptability to pragmatically complex constructions. The obtained results endorse the feasibility of using UTEI as a multidimensional indicator that is fully able to reflect the translation quality, taking into account the criticality of the identified errors.

Table 5

Comparative Results of Machine Translation Systems Evaluation

System	BLEU ↑	TER ↓	METEOR ↑	COMET ↑	UTEI ↑
Google Translate	0.42	0.51	0.46	0.68	0.61
DeepL	0.47	0.44	0.52	0.74	0.69
Microsoft Translator	0.39	0.55	0.43	0.64	0.57
ChatGPT (GPT API)	0.45	0.48	0.50	0.76	0.72
MarianMT	0.36	0.59	0.40	0.61	0.53

Human evaluation is a critically essential component of the study, as it enables the identification of semantic distortions, pragmatic inconsistencies, and hallucinations that often remain outside the scope of automatic metrics.

In order to comprehensively evaluate the results, we propose the Unified Translation Error Index (UTEI), which combines the automatic metrics and error classification results. The UTEI is based on a weighted combination of:

– normalized values of BLEU, TER, METEOR, and COMET;

– a frequency of each error category;

– a semantic criticality coefficient (higher weights for semantic and hallucinatory errors) BLEU, TER, METEOR, and COMET.

In general, the integral indicator can be represented as follows:

$$UTEI = \alpha_1 \cdot BLEU_{norm} - \alpha_2 \cdot TER_{norm} + \\ +\alpha_3 \cdot METEOR_{norm} + \alpha_4 \cdot COMET_{norm} - \beta \cdot E_{weighted} \quad (4)$$

where $E_{weighted}$ – a weighted sum of classified errors according to their criticality, α_i and β – weighting factors.

Weighting factors α_1, α_2, α_3, α_4 and β determine the relative importance of each component of the integral index and ensure the index's adaptability to different research scenarios. Their introduction allows modulating the impact of formal automatic metrics and errors on the final result.

The coefficients α_1-α_4 are responsible for the contribution of the respective automatic metrics to the overall score. In particular, α_1 determines the weight of the BLEU, which characterizes the surface n-gram correspondence; α_2 regulates the influence of TER, which reflects the editing distance; α_3 controls the contribution of the METEOR, which focuses on lexical and morphological flexibility; α_4 sets the weight of the COMET neural network metric, which correlates with human evaluation and takes into account the semantic context.

The coefficient β determines the impact of the weighted sum of the classified errors $E_{weighted}$ on the integral index. Its value is critical as this parameter ensures that deep semantic and pragmatic deviations that may not be adequately reflected in automatic metrics are taken into account.

Normalization of the coefficients can be carried out under the condition:

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \beta = 1 \quad (5)$$

This ensures that the index is interpretable and stable within a fixed range. Weights can be determined in several ways:

– Expert-based, relying on the prioritization of criteria for a particular indicator (for example, in legal texts, the weight of β and α_4).

– Empirically – by optimizing the coefficients to maximize the correlation with human assessment.

– Through a regression model, where the coefficients are the parameters of a function that minimizes the error in predicting the translation quality.

It is also appropriate to note that the $E_{weighted}$ component itself has an internal weighting system:

$$E_{weighted} = \sum_{i=1}^{n} w_i \cdot e_i \quad (6)$$

where e_i – a number of errors of the i-ro type, w_i – a criticality factor of the respective category.

For example, semantic errors and hallucinations may carry greater weight than syntactic or stylistic deviations, as they directly affect the adequacy of the meaning.

Thus, the weighting coefficient system transforms the UTEI integral indicator from a simple aggregated formula into an adaptive multi-criteria analytical tool that can be tailored to a specific text type, language pair, or application area. The proposed metric system creates a multi-level evaluation model in which automatic indicators ensure objectivity and scalability, while an expert evaluation ensures semantic validity. The integral UTEI indicator, alternatively, forms the basis for a comparative analysis of different machine translation architectures and can be applied for further optimisation of error minimisation algorithms.

To minimise the types of errors identified, a hybrid machine translation verification algorithm has been offered, combining a multi-level post-editing mechanism with neurosemantic validation. The algorithm is focused not on generating translations, but on its intelligent verification and adaptive correction. The suggested approach integrates two complementary mechanisms: a formalised linguistic control (rule-based validation) and contextual-semantic evaluation based on large language models.

At the first level, a structural check of the translation is performed after it is generated by the baseline system. Morphological Analysis involves checking the agreement of grammatical categories (number, gender, case, tense, voice) and detecting syntactic anomalies. This stage enables the localization of surface-level grammatical deviations.

Semantic verification via embeddings is conducted by calculating the cosine similarity between the vector representations of the original text and its translation. A low score in the Syntax Constraints Check may indicate potential distortions in the content.

Embedding Similarity Evaluation is performed by analysing inter-sentential relationships and coreference patterns. This procedure is particularly important when dealing with long or syntactically complex constructions.

Back-Translation Comparison involves retranslating the output and comparing it with the original text. Significant divergence between the original and the retranslated text indicates potential semantic loss. The second level of the algorithm implements a hybrid correction model. Syntactic restrictions and rules are applied at the formal level, including templates of acceptable structures, verb valency checks, and compliance of terminology dictionaries and domain glossaries. The Rule-based Adjustments module effectively detects lexical and syntactic errors.

Collaterally, LLM Semantic Validation is used to assess the adequacy of the translation concerning communicative intent, stylistic appropriateness, and the absence of hallucinations. The model receives the source text and the translation as input and either evaluates the semantic error risk or suggests an alternative correction.

Contextual Re-ranking involves reranking several translation candidates based on their semantic, structural, and terminological consistency with the source text. The final version is selected based on an integrated assessment that considers embedding similarity, Back-Translation Comparison results, and classified error types, minimising the risk of semantic distortions and hallucinations. Thus, formal rules ensure structural stability, whereas neural network validation is responsible for deep semantic consistency. The conceptual diagram of a universal verification algorithm is presented below:

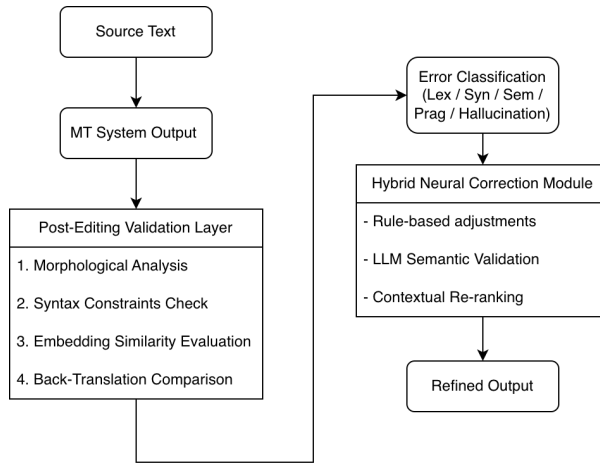

Fig 2. A conceptual diagram of a universal text verification algorithm

Conclusions. A comprehensive analysis of the quality of modern machine translation systems based on neural network architectures was carried out within the scope of the study, which enabled the detection of systemic patterns in the occurrence of lexical, syntactic, semantic, pragmatic errors, and hallucinations. A unified five-level taxonomy of errors has been developed, which demonstrated a high level of inter-expert agreement ($\kappa = 0.8333$), confirming its methodological reproducibility and clear differentiation of categories. A comparison with the international standards MQM, DQF, and SAE J2450 confirmed the structural compatibility of the suggested model and, at the same time, its extension by distinguishing hallucinations as a separate class of deviations relevant to LLM systems.

An integrated indicator, the Unified Translation Error Index (UTEI) has been suggested, which combines automatic metrics (BLEU, TER, METEOR, COMET) with a weighted sum of classified errors according to their criticality. This approach ensures a multidimensional evaluation and considers not only formal translation accuracy, but also deep semantic adequacy. The empirical results have shown that UTEI correlates with the COMET neural network metric and expert evaluation, confirming its suitability to be used as a generalised quality criterion.

Prospects for further research include expanding the experimental corpus, resulting in multilingual data, specialised domains (medical, financial, military-technical translation), and texts with a higher level of pragmatic and cultural specificity. It is also advisable to improve the UTEI integral indicator by adaptively adjusting the error-criticality weights based on the text type and communicative purpose of the translation. A single area of focus is the automation of error classification using LLM as a meta-analysis tool, which will reduce the proportion of manual expert annotation and increase the scalability of the methodology. In a broader context, further research may be directed towards the development of a standardised protocol for comprehensive machine translation quality assessment, combining quantitative metrics, interpretability of the models, and risk analysis in critical areas of application.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Федчук Р., Висоцька В. Автоматизація процесу виявлення та виправлення помилок в текстах. *Проблеми енергоефективності та автоматизації в промисловості та сільському господарстві*, 2024, С. 139-141. URL: https://kntu.kr.ua/file/content/13789/zbirnyk-tez.pdf#page=139

2. Іосіфов Є., Соколов В. Ю. Методи аналізу природної мови та застосування нейронних мереж в кібербезпеці. *Кібербезпека: освіта, наука, техніка*, 2024, 4(24), С. 398-414. DOI: 10.28925/2663-4023.2024.24.398414

3. Muftah M. The impact of artificial intelligence (AI) on translation students' training practices: a case study of ChatGPT translation (ChatGPT-T) output. *Computer Assisted Language Learning*, 2025, P. 1-37. DOI: 10.1080/09588221.2025.2599148

4. Abdelhalim S. M., Alsahil A. A., Alsuhaibani Z. A. Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. *Cogent Arts and Humanities*, 2025, 12(1). DOI: 10.1080/23311983.2025.2508031

5. Chang P., Chen P. J., Lai L. L. Recursive editing with Google Translate: the impact on writing and error correction. *Computer Assisted Language Learning*, 2024, 37(7), P. 2116-2141. DOI: 10.1080/09588221.2022.2147192

6. Qiu J., Pym A. Fatal flaws? Investigating the effects of machine translation errors on audience reception in the audiovisual context. *Perspectives*, 2025, 33(5), P. 913-929. DOI: 10.1080/0907676X.2024.2328757

7. Fitria T. N. Performance of Google Translate, Microsoft Translator, and DeepL Translator: Error analysis of translation result. *Al-Lisan: Jurnal Bahasa (e-Journal)*, 2023, 8(2), P. 115-138. DOI: 10.30603/al.v8i2.3442

8. Ismailia T. The Analysis of Errors on Translating Informative Texts by Google Translate. JETLEE: *Journal of English Language Teaching, Linguistics, and Literature*, 2022, 2(2), P. 123-132. DOI: 10.47766/jetlee.v2i2.498

9. Rodríguez-Rubio Santiago. An analysis of Google Translate and DeepL translation of source text typographical errors in the economic and legal fields. *Revista de Llengua i Dret, Journal of Language and Law*, 2024, 81, P. 88-105. DOI: 10.58992/rld.i81.2024.4188

10. Qasemi P., Ferdowsi S., Bahrami N. On Microsoft Translator's Performance in English-Persian speech-to-text Translation: Recognizing Translation Errors and Identifying the Source of Errors. *Journal of Foreign Language Research*, 2023, 13(3), P. 439-456. DOI: 10.22059/jflr.2023.357547.1028

11. Lu Q., Qiu B., Ding L., Xie L., Tao D. Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT. *Preprints*, 2023. DOI: 10.20944/preprints202303.0255.v1

12. Raju R., Pati P. B., Gandheesh S. A., Sannala G. S., Suriya K. S. Grammatical versus spelling error correction: An investigation into the responsiveness of transformer-based language models using BART and MarianMT. *Journal of Information and Knowledge Management*, 2024, 23(03), 2450037. DOI: 10.1142/S0219649224500370

13. Бушуєв Д. Перекладацька діяльність як лінгвокогнітивний процес (до проблеми машинного перекладу) / Д. Бушуєв // Філологічні студії : зб. студент. наук. пр. / ОНУ ім. І. І. Мечникова, філол. ф-т; [ред. кол.: Н. Г. Арефьєва, О. А. Войцева, М. Л. Дружинець та ін.; відп. за вип. В. Б. Мусій]. Одеса: Одес. нац. ун-т ім. І. І. Мечникова, 2024. Випуск 15. С. 24-33. URL: https://dspace.onu.edu.ua/handle/123456789/42366

14. Шокуров, О. В. Типологія помилок у перекладі наукових термінів із використанням онлайн-ресурсів / О. В. Шокуров, С. М. Чернявська, О. Є. Немерцова // *Закарпатські*

філологічні студії / редкол.: І. М. Зимомря (голов. ред.), М. М. Палінчак, Ю. М. Бідзіля та ін. – Ужгород: Видавничий дім "Гельветика", 2023. Т. 3, вип. 27. С. 155–161. Бібліогр.: с. 161 (12 назв); рез. укр., англ. DOI: 10.32782/tps2663-4880/2022.27.3.29

15. Li Y., Suzuki J., Morishita M., Abe K., Inui K. MQM-chat: Multidimensional quality metrics for chat translation. *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, P. 3283-3299. URL: https://aclanthology.org/2025.coling-main.221/

16. Strandvik I. Translation quality and the role of specifications – How standards can help the translation sector today. *Across Languages and Cultures*, 2025, 26(S), P. 5-24. DOI: 10.1556/084.2025.01057

17. Yameng P., Zhongchi Z., Jiaxin L. The Translation Quality Assessment of Mainstream Neural Machine Translation Tools on Multidimensional Quality Metrics. *5th International Conference on Artificial Intelligence and Education (ICAIE)*, 2025, P. 273-276. IEEE. DOI: 10.1109/ICAIE64856.2025.11158658

18. Лях І. М., Кляп М. П., Ціпіньо А. Ю., Роман А. М. Гібридна лінгвістично орієнтована модель машинного перекладу англо-українських текстів. *Поліграфія і Видавнича Справа*, 2025, 2 (90), С. 128-145. DOI: 10.32403/0554-4866-2025-2-90-128-145

REFERENCES

1. Fedchuk, R., & Vysotska, V. (2024). Avtomatyzatsiia protsesu vyiavlennia ta vypravlennia pomylok v tekstakh. In Problemy enerhoefektyvnosti ta avtomatyzatsii v promyslovosti ta silskomu hospodarstvi. 139-141. https://kntu.kr.ua/file/content/13789/zbirnyk-tez.pdf#page=139

2. Iosifov, I., & Sokolov, V. Y. (2024). Metody analizu pryrodnoi movy ta zastosuvannia neironnykh merezh v kiberbezpetsi. In Kiberbezpeka: osvita, nauka, tekhnika, 4(24), 398-414. https://doi.org/10.28925/2663-4023.2024.24.398414

3. Muftah, M. (2025). The impact of artificial intelligence (AI) on translation students' training practices: a case study of ChatGPT translation (ChatGPT-T) output. Computer Assisted Language Learning, 1-37. https://doi.org/10.1080/09588221.2025.2599148

4. Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. Cogent Arts & Humanities, 12(1). https://doi.org/10.1080/23311983.2025.2508031

5. Chang, P., Chen, P. J., & Lai, L. L. (2024). Recursive editing with Google Translate: the impact on writing and error correction. Computer Assisted Language Learning, 37(7), 2116-2141. https://doi.org/10.1080/09588221.2022.2147192

6. Qiu, J., & Pym, A. (2025). Fatal flaws? Investigating the effects of machine translation errors on audience reception in the audiovisual context. Perspectives, 33(5), 913-929. https://doi.org/10.1080/0907676X.2024.2328757

7. Fitria, T. N. (2023). Performance of google translate, microsoft translator, and deepL translator: Error analysis of translation result. Al-Lisan: Jurnal Bahasa (e-Journal), 8(2), 115-138. https://doi.org/10.30603/al.v8i2.3442

8. Ismailia, T. (2022). The Analysis of Errors on Translating Informative Texts by Google Translate. JETLEE: Journal of English Language Teaching, Linguistics, and Literature, 2(2), 123-132. https://doi.org/10.47766/jetlee.v2i2.498

9. Rodríguez-Rubio, Santiago. (2024). An analysis of Google Translate and DeepL translation of source text typographical errors in the economic and legal fields. Revista de Llengua i Dret, Journal of Language and Law, 81, 88-105. https://doi.org/10.58992/rld.i81.2024.4188

10. Qasemi, P. Ferdowsi, Ṡ and Bahrami, N. (2023). On Microsoft Translator's Performance in English-Persian speech-to-text Translation: Recognizing Translation Errors and Identifying

the Source of Errors. Journal of Foreign Language Research, 13(3), 439-456. https://doi.org/10.22059/jflr.2023.357547.1028

11. Lu, Q., Qiu, B., Ding, L., Xie, L., & Tao, D. (2023). Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT. Preprints. https://doi.org/10.20944/preprints202303.0255.v1

12. Raju, R., Pati, P. B., Gandheesh, S. A., Sannala, G. S., & Suriya, K. S. (2024). Grammatical versus spelling error correction: An investigation into the responsiveness of transformer-based language models using BART and MarianMT. Journal of Information & Knowledge Management, 23(03), 2450037. https://doi.org/10.1142/S0219649224500370

13. Bushuiev, D. (2024). Perekladatska diialnist yak linhvokohnityvnyi protses (do problemy mashynnoho perekladu). Filolohichni studii, 15, 24-33. https://dspace.onu.edu.ua/handle/123456789/42366

14. Shokurov, O. V., Cherniavska, S. M., & Nemertsova, O. Ye. (2023). Typolohiia pomylok u perekladi naukovykh terminiv iz vykorystanniam onlain-resursiv. Zakarpatski filolohichni studii, 3(27), 155-161. https://doi.org/10.32782/tps2663-4880/2022.27.3.29

15. Li, Y., Suzuki, J., Morishita, M., Abe, K., & Inui, K. (2025, January). MQM-chat: Multidimensional quality metrics for chat translation. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 3283-3299). https://aclanthology.org/2025.coling-main.221/

16. Strandvik, I. (2025). Translation quality and the role of specifications – How standards can help the translation sector today. Across Languages and Cultures, 26(S), 5-24. https://doi.org/10.1556/084.2025.01057

17. Yameng, P., Zhongchi, Z., & Jiaxin, L. (2025, May). The Translation Quality Assessment of Mainstream Neural Machine Translation Tools on Multidimensional Quality Metrics. In 2025 5th International Conference on Artificial Intelligence and Education (ICAIE) (pp. 273-276). IEEE. https://doi.org/10.1109/ICAIE64856.2025.11158658

18. Liakh, I. M., Klyap, M. P., Tsipino, A. Y., & Roman, A. M. (2025). Hybrid linguistically oriented model of english-ukrainian machine translation. Printing and Publishing, 2(90), 128-145. https://doi.org/10.32403/0554-4866-2025-2-90-128-145

doi: 10.32403/0554-4866-2026-1-91-60-76

ГІБРИДНА ПЕРЕВІРКА ПЕРЕКЛАДУ З П’ЯТИРІВНЕВОЮ КЛАСИФІКАЦІЄЮ ПОМИЛОК

І. М. Лях[1], Р. Я. Жовтані[2], А. Ю. Ціпіньо[3], А. М. Роман[4]

1. Ужгородський національний університет, вул. Народна, 3, Ужгород, 88000, Україна https://orcid.org/0000-0001-5417-9403 e-mail: igor.lyah@uzhnu.edu.ua

2. Ужгородський національний університет, вул. Народна, 3, Ужгород, 88000, Україна https://orcid.org/0000-0002-7421-148X e-mail: ruslana.zhovtani@uzhnu.edu.ua

3. Ужгородський національний університет, вул. Народна, 3, Ужгород, 88000, Україна https://orcid.org/0009-0007-1704-2732 e-mail: tsipino.artemii@student.uzhnu.edu.ua

4. Ужгородський національний університет, вул. Народна, 3, Ужгород, 88000, Україна https://orcid.org/0009-0005-5774-8946 e-mail: roman.anhelina@student.uzhnu.edu.ua

У статті досліджено проблему системних помилок у сучасних системах машинного перекладу, побудованих на основі нейромережевих і трансформерних архітектур. Метою роботи є формалізація уніфікованої моделі класифікації помилок, розроблення інтегрального показника оцінювання якості перекладу та створення концептуальної схеми гібридного алгоритму їх мінімізації. Для експериментального аналізу сформовано корпус обсягом близько 200 речень, що охоплює технічні, юридичні, художні та розмовні тексти, ідіоматичні вирази, складні синтаксичні конструкції, пасивні форми, випадки омонімії та багатозначності. Порівняльне дослідження виконано на основі аналізу п'яти систем машинного перекладу: Google Translate, DeepL, Microsoft Translator, ChatGPT (GPT API) та MarianMT.

Запропоновано п'ятирівневу класифікацію помилок, що включає лексичні, синтаксичні, семантичні, прагматичні відхилення та галюцинації. Встановлено, що традиційні автоматичні метрики (BLEU, TER, METEOR) не повною мірою відображають глибинні змістові спотворення, у зв'язку з чим до процедури оцінювання інтегровано нейромережну метрику COMET та експертну людську оцінку. На основі багатофакторного підходу розроблено інтегральний показник Unified Translation Error Index (UTEI), який поєднує нормалізовані значення автоматичних метрик зі зваженою сумою класифікованих помилок відповідно до їх критичності, забезпечуючи багатовимірність та інтерпретованість результатів.

Для мінімізації виявлених відхилень запропоновано гібридний алгоритм універсальної перевірки перекладу, що поєднує формалізовану лінгвістичну перевірку зі семантичною валідацією на основі нейромережевих моделей. Алгоритм використовує аналіз embedding-подібності, контроль синтаксичної та морфологічної коректності, а також механізм зворотного перекладу для детекції та корекції глибинних змістових спотворень. Запропонований підхід забезпечує адаптивність і масштабованість алгоритму для інтеграції в різні NMT та LLM-системи, підвищуючи структурну стабільність, семантичну узгодженість і загальну надійність машинного перекладу.

Ключові слова: *машинний переклад; нейромережеві моделі; класифікація помилок; оцінювання якості; UTEI; гібридний алгоритм.*

Стаття надійшла до редакції 16.02.2026.
Received 16.02.2026.
Стаття прийнята 20.05.2026
Accepted 20.05.2026
Опубліковано 30.05.2026
Published 30.05.2026