**UDC 81'322.4:004.912**
**DOI https://doi.org/10.24919/2308-4863/89-2-41**

***Anton SHPIGUNOV,***
*orcid.org/0009-0008-2294-9045*
*Ph. D student and Assistant at the Department of Theory and Practice of Translation from English*
*Educational and Research Institute of Philology of National Taras Shevchenko University of Kyiv*
*(Kyiv, Ukraine) shpigunov@knu.ua*

# USING AUTOMATED QUALITY METRICS TO IMPROVE MACHINE TRANSLATION INTO UKRAINIAN

*The article reviews the issue of evaluating the quality of machine translation into the Ukrainian language, taking into account the diversity of its morphological word forms and syntactic variants. The peculiarities of the Ukrainian language with its morphological inflection and variable word order make it difficult to apply traditional automated metrics that focus on counting matches of words or word sequences (n-grams). The evolution of automated machine translation quality metrics is considered, starting from approaches based on comparing word sequences (BLEU, NIST, ROUGE) to character-level methods (ChrF, BEER) and word vector embeddings (BLEURT, BERTScore, COMET, etc.).*

*In particular, it was found that metrics that calculate the match ratios for words and their sequences (in particular, the popular BLEU) appear to be overly sensitive to morphological variants and changes in word order, which are typical for Ukrainian. Instead, character-level metrics (ChrF) and metrics that measure the distance between vector representations of sentences (COMET, BLEURT, and others) demonstrate greater reliability, correlating better with human judgment due to their ability to account for semantic and morphological variations, as well as potential synonymy between words that are spelled completely differently.*

*Based on the conducted review, we recommend strategies for pre-processing the Ukrainian translation text in order to increase the informativity of metrics based on word and word sequence matches (stemming, i.e., truncation of non-stem morphemes, or lemmatization, i.e., reduction to dictionary form). It is also proposed to use automated metrics based on character matches (in particular, ChrF), as well as machine learning-based metrics that predict the human evaluation of machine translation based on vector representations of the source and target sentences (hypotheses), such as COMET and its variants.*

*Since automated quality metrics based on machine learning require specialized datasets in the form of annotated parallel corpora, the importance of creating such annotated datasets where the target language is Ukrainian is emphasized. This will significantly improve the quality of metrics for evaluating machine translation into Ukrainian. In turn, improved metrics will help machine translation systems generate better translated sentences using techniques such as metric-based ranking of candidate translations.*

***Key words:*** *machine translation, machine translation quality, automated machine translation quality metrics, BLEU, ChrF, COMET.*

***Антон ШПІГУНОВ,***
*orcid.org/0009-0008-2294-9045*
*аспірант, асистент кафедри теорії та практики перекладу з англійської мови*
*Науково-навчального інституту філології*
*Київського національного університету імені Тараса Шевченка*
*(Київ, Україна) shpigunov@knu.ua*

# ВИКОРИСТАННЯ АВТОМАТИЗОВАНИХ МЕТРИК ЯКОСТІ ПЕРЕКЛАДУ ДЛЯ ПОКРАЩЕННЯ МАШИННОГО ПЕРЕКЛАДУ НА УКРАЇНСЬКУ МОВУ

*У статті розглядається проблема оцінювання якості машинного перекладу на українську мову з урахуванням розмаїття її морфологічних словоформ і синтаксичних варіантів. Особливості української мови з її морфологічною флексією та змінним порядком слів ускладнюють застосування традиційних автоматизованих метрик, які орієнтовані на підрахунок збігів слів або їх послідовностей. Розглядається еволюція автоматизованих метрик якості машинного перекладу, починаючи від підходів, заснованих на зіставленні послідовностей (n-грам) слів (BLEU, NIST, ROUGE), до методів, які використовують символьний рівень (ChrF, BEER) та векторні вкладення слів (BLEURT, BERTScore, COMET тощо).*

*Зокрема, виявлено, що метрики, що рахують коефіцієнти збігів слів і їх послідовностей (зокрема, популярна BLEU) надмірно чутливі до морфологічних варіантів та зміни порядку слів, що є характерними для української. Натомість, метрики, що аналізують коефіцієнти збігів на рівні друкованих знаків (ChrF) та метрики, що вимірюють дистанцію між векторними представленнями речень (COMET, BLEURT та інші), демонструють більшу надійність, краще корелюючи з людською оцінкою завдяки їх здатності враховувати семантичні та морфологічні варіації, а також потенційну синонімію між словами, що мають різне написання.*

*На основі проведеного аналізу рекомендуються стратегії для попередньої обробки українського перекладного тексту з метою підвищення інформативності метрик на основі збігів слів і їх послідовностей («відсікання» не-кореневих морфем, також відоме як «стеммінг», або приведення до словникової форми (лематизація)). Також пропонується використовувати автоматизовані метрики на основі збігів друкованих знаків (зокрема, ChrF), а також метрики на основі машинного навчання, які прогнозують людську оцінку машинного перекладу на основі векторних представлень речень оригіналу та перекладу (гіпотези), такі як, наприклад, COMET і її варіанти.*

*Оскільки автоматизовані метрики якості на основі машинного навчання потребують спеціальних наборів даних у формі анотованих паралельних корпусів, наголошується на важливості створення таких анотованих наборів даних, де мовою перекладу є українська. Це суттєво покращить якість метрик у частині оцінки машинного перекладу на українську мову. В свою чергу, покращені метрики допоможуть системам машинного перекладу генерувати кращі перекладні речення за допомогою таких технік, як реранкінг кандидатів на основі метрики.*

***Ключові слова:*** *машинний переклад, якість машинного перекладу, автоматизовані метрики якості машинного перекладу, BLEU, ChrF, COMET.*

**Introduction and Problem Definition.** The linguistic profile of Ukrainian is marked by its rich morphology and syntactic flexibility – features that have significant implications for machine translation quality. As an East Slavic language, Ukrainian employs a highly inflected system and a flexible word order; rather than relying on a fixed syntactic structure, Ukrainian uses inflectional markers to indicate subject–verb–object relations, allowing syntactic variations that highlight focus or topicalization without loss of meaning. Such characteristics can challenge conventional evaluation metrics that depend on rigid word-level matches or fixed word order, as these methods may overlook acceptable and semantically equivalent variations inherent in Ukrainian translations. Consequently, tailoring machine translation evaluation for Ukrainian requires metrics that account for morphological variation and syntactic fluidity – approaches like character-level analysis or embedding-based models, which better capture semantic similarity across diverse structural expressions.

The field of machine translation has progressed steadily from rules-based and statistical systems to neural and large language model-based methods, achieving impressive translation quality. To gauge and measure this progress, standardized metrics are important. Although human evaluation remains the gold standard for evaluating any translation, including human and automated methods, automated metrics are crucial to ensure inexpensive and fast measurement which could enable rapid iteration on emerging machine translation methods.

An Overview of Machine Translation Quality Metrics

Like methods of machine translation proper, automated quality metrics had a similar evolution. The first wave of metrics was based on **matching n-grams** (sequences of *n* words) between the reference human translations and the MT hypothesis.

Metrics Based on n-Gram Matching

The first n-gram-based MT evaluation metric was **BLEU**, introduced in (Papineni et al., 2002). Specifically, BLEU is based on *n-gram precision*, which can be roughly explained as follows:

$$p_n = \frac{N_{matching\ n-grams}}{N_{total\ candidate\ n-grams}}$$

However, reliance on *n*-gram matching made BLEU sensitive to word order, synonyms and paraphrases in translation, and morphological differences between reference and translation in languages with a high degree of morphological variety like Ukrainian.

**NIST** is a later metric introduced in (Doddington, 2002), which augments BLEU's precision metric with weighing each n-gram's value based on how rare, and thus informative to the text, it is. This improvement, however, did not address the fundamental shortfalls of BLEU's approach.

**ROUGE**, proposed in (Lin, 2004), establishes a group of metrics based on *recall*, rather than precision used in BLEU:

$$r_n = \frac{N_{matching\ n-grams}}{N_{total\ number\ of\ n-grams\ in\ reference}}$$

Despite this change, in essence ROUGE remains sensitive to word order changes. Later versions based on longest common subsequence matching (**ROUGE-L**) and skip-grams (**ROUGE-S**) are somewhat more flexible with regard to word order.

Being a surface-level word overlap algorithm, ROUGE is fundamentally unable to account for synonyms and more complex translation transformations.

**METEOR**, presented in (Banerjee and Lavie, 2005), attempts to improve on its predecessors by using explicit word-for-word matching instead of matching n-length sequences of words (n-grams). Another improvement of METEOR metric is its optional use of word stemming and attempts to match by meaning, relying on WordNet to establish synonymy, thus making the metric more useful for morphologically complex languages like Ukrainian. Recent availability of WordNet-compatible dictionaries for Ukrainian

potentially makes meaning-based matching available (Siegel, 2024). A later iteration of the metric, called **METEOR-NEXT**, introduces improvements to the algorithm's sentence aligner and scoring scheme, as well as tunes the metric's parameters to optimize its correlation with the Human-targeted Translation Error Rate (Denkowski and Lavie, 2010).

Metrics Based on Edit Distance

A different approach to measuring MT quality envisaged **measuring the distance** or effort it would take an editor to edit the MT hypothesis so it would match the reference translation.

Translation Error Rate, or **TER** in short, proposed in (Snover et al., 2006), is the first metric to follow this approach. It is defined as the "minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references". Edits include insertions, deletions, substitutions of single words, and shifts of word sequences within the hypothesis, all scored at equal cost. The TER score is calculated by dividing the number of edits by the length of the reference translation if there is only one, or the average length of the reference translations if there are many presented.

To address some of the limitations of TER, a variant TER, called Human-targeted TER or **HTER** is also described. This variant makes use of human annotators to create targeted references. The HTER metric involves a human annotator editing a system-generated hypothesis to make it both fluent and semantically equivalent to the reference, while minimizing edits. The TER score is then calculated using this human-created targeted reference.

Snover et al.'s results show that "METEOR correlates with human judgments better than TER, when given the same number of references, but that HTER correlates with human judgments better than HMETEOR".

Still, TER requires an exact match between the hypothesis and the reference on the word level. This means that, because it only measures the number of edits between the best reference and the hypothesis, it does not take into account semantic equivalence or synonymy. For the same reason, TER will consider different morphological forms as completely different words.

**WER** (word error rate) is a metric similar to TER, but mostly used in evaluation of automatic speech recognition (Klakow and Peters, 2002). WER focuses on word substitutions, insertions, and deletions, but unlike TER, ignores word sequence shifts.

In general, it should be noted that word-level metrics based on either n-gram matching or edit distance estimation should not be expected to be accurate for target languages that are morphologically rich, have a high degree of inflection, and allow variation in word order, like Ukrainian. Word-level metrics can unjustly penalize differences between word forms by discounting them as full mismatches. They can also penalize word order shifts that are natural to the target language and do not in fact affect translation meaning.

If word-level metrics need to be used for languages with a high degree of morphological flexibility (e.g., in a large experiment with multiple languages for consistency), strategies like stemming and lemmatization can be useful to mitigate this over-penalization. For instance, (Servan et al., 2016) proposes an extension of the METEOR metric using a function called *Morphy-7WN1* which firstly checks special cases in an exception list and secondly uses rules to lemmatize words according to their syntactic class. For Ukrainian in specific, the Ukrainian version of *PyMorphy2* (Korobov, 2015) can be used for a similar purpose.

Character-Based Metrics

Sequence-matching MT quality metrics, however, are not limited to matching sequences of words. Instead of matching whole words, **ChrF** computes F-scores over character n-grams (Popović, 2015). This approach tends to be more forgiving of the morphological variations common in morphologically rich languages since it can capture similarities even when morphological word forms differ.

**BEER**, or Better Evaluation as Ranking, is a **learned metric** for evaluating machine translation (MT) quality that uses a regression model based on character n-grams and word bigrams (Stanojevic and Sima'an, 2014). It goes beyond n-gram matching by training a regression model to best correlate with human judgment. As such, it requires a human-annotated bilingual corpus for the language pair and domain in question to be representative.

Vector Embedding-Based Metrics

Word vector embeddings, such as those created by models like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have allowed linguists to capture the distributional semantics of words based on their co-occurrence patterns in large text corpora in the form of dense vectors of floating-point numbers. This idea is rooted in the distributional hypothesis (often summarized as "You shall know a word by the company it keeps"), as articulated in (Firth, 1957). These models map words to dense vectors, capturing relationships between words based on their context. Vector embeddings have been used as a proxy for meaning, because words with similar

meanings tend to have similar vector representations when compared using various metrics like cosine similarity or Euclidean distance, among others. This enables matching not by word – or character-level equivalence, but by semantic similarity, meaning that words or phrases don't have to be identical to be considered similar, which differs from word – or character-level matching.

In the context of MT quality metrics, vector embeddings have enabled comparisons between MT hypotheses and reference translations by the similarity of the former's vectors to those of the latter (Sun and Wang, 2024).

It's also worth distinguishing between static word embeddings (like Word2Vec) and contextualized word embeddings (generated by large models like BERT, ELMo, XLM). Contextual embeddings generate different vector representations for the same word depending on the surrounding words, which allows them to capture more nuanced semantic relationships.

It is also worth pointing out that vector-based metrics using vector embeddings generated by multilingual models like BERT and XLM appear to be effective for languages other than English (Zhang et al., 2020).

**METEOR-VECTOR** is a later version of the METEOR metric that uses vector embeddings to enhance its matching capabilities.

**YiSi** (Lo, 2019) and its variants, such as YiSi-2, are metrics that use word or contextual embeddings to compute semantic similarity between hypothesis and reference sentences. They work by aligning tokens or subword units in the candidate and reference based on their cosine similarity (the cosine of the angle between two vectors, or equivalently the dot product between their normalizations) in an embedding space, then aggregating these similarity scores (possibly with additional weighting schemes) to produce an overall quality score

**BERTSCORE** uses contextualized token embeddings from BERT to calculate the similarity between two sentences. It computes similarity as the sum of cosine similarities between the tokens' embeddings using a greedy matching (Corley and Mihalcea, 2005). It calculates precision, recall, and F1 scores based on these similarities.

**BLEURT** (Sellam et al., 2020) also utilizes BERT-based contextual embeddings to capture semantic similarities between candidate and reference sentences. The model undergoes a two-phase training process that includes pre-training, where the model is initialized with a pre-trained BERT and further trained on synthetic data derived from Wikipedia. This data is perturbed to simulate various linguistic variations, enhancing the model's robustness to domain and quality shifts; then BLEURT is fine-tuned on human-rated translations from datasets such as the WMT Metrics Shared Task, aligning its evaluations with human judgment. This training regimen enables BLEURT to effectively evaluate fluency and adequacy in translations, often achieving higher correlations with human assessments compared to metrics like BLEU. However, (Yan et al., 2023) has identified potential robustness issues, such as the presence of universal adversarial translations – inputs that receive high scores despite being inadequate.

**Prism** (Thompson and Post, 2020) is an automatic metric for machine translation evaluation that is based on a multilingual neural machine translation (NMT) model which functions as a paraphraser. It uses token-level probabilities from the paraphraser to evaluate the quality of translations. Due to its design, the metric can be used with or without reference translations.

**COMET** (Rei et al., 2020) is a vector-based, learned metric for evaluating machine translation quality. It focuses on optimizing for correlation with human judgments. COMET models can be trained with different objectives, including direct assessment (DA), human-targeted translation edit rate (HTER), and multi-dimensional quality metrics (MQM). COMET utilizes a highly multilingual encoder, which allows it to work well even when English is not the target language. COMET has been enhanced (Rei et al., 2022) with integration of OpenKIWI, a Pytorch-based open-source framework for translation quality estimation (Kepler et al., 2019). In (Rei et al., 2023), the model and training sizes for COMET were scaled up, among other optimizations.

**MetricX** (Juraska et al., 2023) is Google's recent translation quality metric. It is a learned regression-based metric that utilizes a pretrained language model. As with other regression-based metrics, the model is trained to infer a score that would have the best possible. correlation with human judgments, as expressed by DA and MQM ratings.

Selecting a Set of Metrics for MT System Development

With the multitude of metrics available, finding a metric or a combination of metrics is not a trivial task.

In our case, our end goal is to select metrics to for benchmarking, development and iteration of new machine translation systems. In our particular case, we are interested in language pairs where the target language is Ukrainian. Considering these constraints, we would identify the following criterial for metric selection:

a) Fitness for language pairs where Ukrainian is the target language;

b) High correlation with human evaluation;

c) Performance and compute cost permissive of continuous iteration of emerging MT models.

Fitness for Ukrainian-target Language Pairs

In Ukrainian, a language with rich morphological variety and more flexible word order, metrics based on matching sequences of words will excessively penalize word forms by scoring them as different words. Therefore, for languages like Ukrainian, character-level or embedding-based metrics are preferrable. As mentioned above, in the cases when word-level metrics still have to be used, preliminary stemming or lemmatization are recommended.

This means that n-gram based metrics like BLEU, ROUGE and NIST, and TER are not advisable to be used for Ukrainian-target language pairs in their naïve form. A word-level metrics that may be recommended is METEOR, provided that stemming is performed pre-evaluation, and that Ukrainian-language similarity dictionaries like `ukrajinet` are used.

Since ChrF is a character-level metric, it should be less sensitive to morphological variety of languages like Ukrainian and may therefore be recommended.

Metrics based on vector representations tend to more robust and consistent than n-gram based ones like BLEU, regardless of the language pair (Mathur et al., 2020).

Since different morphological forms of the same word would be relatively close to each other in vector representations, metrics based on those are not expected to penalize morphological variance of languages like Ukrainian.

Correlation with Human Evaluation

The correlation of a metric with human judgement is a key indicator of its effectiveness and reliability. Studies like (Kocmi et al., 2021) and (Mathur et al., 2020) evaluate the correlation between human-targeted scores like MQM, SQM and DA and various automated MT quality metrics using statistical correlation analysis methods including Pearson, Spearman correlation analysis, and Kendall's Tau.

Such correlation studies show that in general, learned metrics exhibit higher correlation with human judgment than rules-based ones. For instance, (Avramidis et al., 2023) points out that BLEU, SpBleu or Chrf correlate poorly with human ratings and perform less consistently than neural-based learned metrics. (Kocmi et al., 2021) suggest that learned metrics like COMET Prism, and BLEURT exhibit higher Pearson's and Spearman's correlation with human-targeted evaluations than BLEU. With this said, learned metrics may exhibit biases from their training data.

A key consideration for using DA and MQM for metric evaluation and training learned metrics for a language pair is the availability of annotated datasets for the said language pair. (Paniv et al., 2024) are assembling such a dataset, and although the primary purpose of their effort is primarily to train an MT system capable of translating to Ukrainian, these parallel sentences may potentially be used to train automated MT quality metric models as well.

Furthermore, creating an MQM dataset with Ukrainian as a target language would be an important contribution to machine translation quality assessment for Ukrainian-target language pairs, which means that future metrics could exhibit a stronger correlation to human evaluation for Ukrainian. In turn, metrics that better approximate human evaluation, especially in a reference-free context, could lead to gradual improvements in machine translation quality for Ukrainian targets.

Performance and compute cost

Metrics based on string and character matching are inherently less demanding in terms of compute than learned metrics. However, even the learned metrics based on comparatively large models like BERT or XLM-Roberta-Large, can still be run at reasonable scale on consumer hardware.

Emerging Improvements and Use Cases

(Kocmi et al., 2021) take MT quality metric evaluation beyond statistical correlation with human-targeted evaluation, introducing *accuracy*, the measure of how well an automatic metric aligns with human judgment in pairwise system comparisons. Using this approach, they confirm that learned metrics consistently out-perform sequence matching-based ones like BLEU, TER and ChrF.

Another advancement beyond statistical correlation with human judgments is the development of **explainable and interpretable metrics**. In addition to a numeric score, such metrics produce a "report" outlining detected translation errors and discrepancies, indicating the span where such errors have occurred (Guerreiro et al., 2024).

(Perrella et al., 2024) reference novel use cases for automated quality metrics. Namely, reference-less metrics like variations of COMET, can be used for translation re-ranking. A re-ranking pipeline involves an MT engine proposing several candidate translations, a reference-less metric algorithm ranking the candidates based on the evaluated score, and the system proposing the highest-ranking candidate to the user.

(Ramos et al., 2024) propose the use of reinforcement learning with human feedback, or RHLF, a recent technique to improve the quality of the text generated by a language model, making it closer to what humans would generate, to improve the quality of attained machine translations. In this

method, human annotations, or metrics trained on human annotations, can be used as a reward model for the algorithm's self-improvement. The findings of this study also demonstrate the effectiveness of combining RL training with reranking techniques described above, showcasing substantial improvements in translation quality.

**Conclusions**

– Automated MT quality metrics are an important tool for assessing the performance of MT methods, their development and continuous improvement.

– MT quality metrics have evolved since the early 2000's from matching sequences of words or characters, to algorithms based on large language models trained to infer how human annotators would evaluate a given translation.

– Word-level sequence matching metrics will not perform consistently with translations where the target language is Ukrainian, as they tend to over-penalize rich morphological variety and free word order inherent to the Ukrainian language. Sequence matching-based metrics can be used either on the character level (ChrF), or on the word level with mitigation strategies like lemmatization or stemming.

– As a rule, learned metrics tend to have a stronger correlation with human assessments, in terms of linear and pair-wise statistical metrics and other methods.

– Reference-less metrics can be used for continuous feedback and improvement of MT systems using techniques like candidate re-ranking and reinforcement learning with human feedback.

– For learned metrics, human-annotated assessment data like MQM and DA datasets constitute vital resources for training the metric models. Although efforts for collecting such datasets are underway, additional collection and annotation of such datasets will constitute an important contribution towards improving learned metrics for Ukrainian as a target language in a variety of settings and domains, and as a consequence, for improving the MT engines themselves.

**BIBLIOGRAPHY**

1. Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research -the second international conference*. (San Diego, California, 2002). San Diego, California: Association for Computational Linguistics, 2002. DOI:10.3115/1289189.1289273. C. 138.

2. Popović M. chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, 2015. DOI:10.18653/v1/W15-3049. C. 392–395.

3. Stanojevic M., Sima'an K. BEER: Better Evaluation as Ranking. *Proceedings of the Ninth Workshop on Statistical Machine Translation Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. DOI:10.3115/v1/W14-3354. C. 414–419.

4. Avramidis E., Manakhimova S., Macketanz V. et al. Challenging the State-of-the-art Machine Translation Metrics from a Linguistic Perspective. *Proceedings of the Eighth Conference on Machine Translation WMT 2023*. Singapore: Association for Computational Linguistics, 2023. DOI:10.18653/v1/2023.wmt-1.58. P. 713–729.

5. Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*(Ann Arbor, Michigan, 06.2005). Ann Arbor, Michigan : Association for Computational Linguistics, 2005. Also available online, URL: https://aclanthology.org/W05-0909/ (accessed 13/01/2025).P. 65–72.

6. Corley C., Mihalcea R. Measuring the Semantic Similarity of Texts. *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*(Ann Arbor, Michigan, 06.2005). Ann Arbor, Michigan : Association for Computational Linguistics, 2005. Also available online, URL: https://aclanthology.org/W05-1203/ (accessed 07/02/2025).P. 13–18.

7. Denkowski M., Lavie A. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational LinguisticsNAACL-HLT 2010*. (Los Angeles, California, 06.2010). Los Angeles, California : Association for Computational Linguistics, 2010. Also available online, URL: https://aclanthology.org/N10-1031/ (accessed 01/02/2025).P. 250–253.

8. Firth J. R. A synopsis of linguistic theory, 1930 – 1955. Studies in Linguistic Analysis. Special Volume of the Philological Society. *Selected Papers of JR Firth 1952-59*. (1957). Indiana University Press Bloomington & London, 1957. P. 168–205. 1957.

9. Guerreiro N. M., Rei R., Stigt D. van et al. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*. Vol. 12, 2024. P. 979–995. DOI:10.1162/tacl_a_00683.

10. Juraska J., Finkelstein M., Deutsch D. et al. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. *Proceedings of the Eighth Conference on Machine Translation WMT 2023*. Singapore : Association for Computational Linguistics, 2023. DOI:10.18653/v1/2023.wmt-1.63. P. 756–767.

11. Kepler F., Trénous J., Treviso M. et al. OpenKiwi: An Open-Source Framework for Quality Estimation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* Florence, Italy : Association for Computational Linguistics, 2019. DOI:10.18653/v1/P19-3020. P. 117–122.

12. Klakow D., Peters J. Testing the correlation of word error rate and perplexity. *Speech Communication*. Vol. 38, Issue 1. P. 19–28. DOI:10.1016/S0167-6393(01)00041-3.

13. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*(Barcelona, Spain, 07.2004). Barcelona, Spain : Association for Computational Linguistics, 2004. Also available online, URL: https://aclanthology.org/W04-1013/ (accessed 30/06/2025).P. 74–81.

14. Lo C. YiSi – a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1) WMT 2019*. Florence, Italy : Association for Computational Linguistics, 2019. DOI:10.18653/v1/W19-5358. P. 507–513.

15. Mathur N., Baldwin T., Cohn T. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics ACL 2020*. Online : Association for Computational Linguistics, 2020. DOI:10.18653/v1/2020.acl-main.448. P. 4984–4997.

16. Paniv Y., Chaplynskyi D., Trynus N. et al. Setting up the Data Printer with Improved English to Ukrainian Machine Translation. *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024UNLP 2024*. Torino, Italia : ELRA and ICCL, 2024. Also available online, URL: https://aclanthology.org/2024.unlp-1.6/ (accessed 30/06/2025).P. 41–50.

17. Papineni K., Roukos S., Ward T. et al. Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL 2002*. (Philadelphia, Pennsylvania, USA, 07.2002). Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, 2002. DOI:10.3115/1073083.1073135. P. 311–318.

18. Pennington J., Socher R., Manning C. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)EMNLP 2014*. (Doha, Qatar, 10.2014). Doha, Qatar : Association for Computational Linguistics, 2014. DOI:10.3115/v1/D14-1162. P. 1532–1543.

19. Ramos M., Fernandes P., Farinhas A. et al. Aligning Neural Machine Translation Models: Human Feedback in Training and Inference. *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1) EAMT 2024*. Sheffield, UK: European Association for Machine Translation (EAMT), 2024. Also available online, URL: https://aclanthology.org/2024.eamt-1.22/ (accessed 07/02/2025). p. 258–274.

20. Rei R., Stewart C., Farinha A. C. et al. COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)EMNLP 2020*. Online: Association for Computational Linguistics, 2020. DOI:10.18653/v1/2020.emnlp-main.213. P. 2685–2702.

21. Siegel M. hdaSprachtechnologie/ukrajinet. URL: https://github.com/hdaSprachtechnologie/ukrajinet (accessed 30/01/2025).2024.

22. Snover M., Dorr B., Schwartz R. et al. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers AMTA 2006*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 2006. Also available online, URL: https://aclanthology.org/2006.amta-papers.25/ (accessed 13/01/2025). p. 223–231.

23. Thompson B., Post M. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)EMNLP 2020*. Online: Association for Computational Linguistics, 2020. DOI:10.18653/v1/2020.emnlp-main.8. P. 90–121.

24. Yan Y., Wang T., Zhao C. et al. BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023. DOI:10.18653/v1/2023.acl-long.297. P. 5428–5443.

25. Kocmi T., Federmann C., Grundkiewicz R. et al. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. arXiv, 2021. DOI:10.48550/arXiv.2107.10821.

26. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. arXiv, 2015. DOI:10.48550/arXiv.1503.07283.

27. Mikolov T., Chen K., Corrado G. et al. Efficient Estimation of Word Representations in Vector Space. arXiv, 2013. DOI:10.48550/arXiv.1301.3781.

28. Perrella S., Proietti L., Cabot P.-L. H. et al. Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics. arXiv, 2024. DOI:10.48550/arXiv.2410.05183.

29. Rei R., Guerreiro N. M., Pombal J. et al. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. arXiv, 2023. DOI:10.48550/arXiv.2309.11925.

30. Rei R., Treviso M., Guerreiro N. M. et al. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. arXiv, 2022. DOI:10.48550/arXiv.2209.06243.

31. Sellam T., Das D., Parikh A. P. BLEURT: Learning Robust Metrics for Text Generation. arXiv, 2020. DOI:10.48550/arXiv.2004.04696.

32. Servan C., Berard A., Elloumi Z. et al. Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? arXiv, 2016. DOI:10.48550/arXiv.1610.01291.

33. un K., Wang R. Textual Similarity as a Key Metric in Machine Translation Quality Estimation. arXiv, 2024. DOI:10.48550/arXiv.2406.07440.

34. Zhang T., Kishore V., Wu F. et al. BERTScore: Evaluating Text Generation with BERT. arXiv, 2020. DOI:10.48550/arXiv.1904.09675.

**REFERENCES**

1. Doddington (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, Association for Computational Linguistics, San Diego, California, pp. 138 URL: http://portal.acm.org/citation.cfm?doid=1289189.1289273.

2. Popović (2015) chrF: character n-gram F-score for automatic MT evaluation, Association for Computational Linguistics, Lisbon, Portugal, pp. 392–395 URL: http://aclweb.org/anthology/W15-3049.

3. Stanojevic and Sima'an (2014) BEER: BEtter Evaluation as Ranking, Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 414–419 URL: http://aclweb.org/anthology/W14-3354.

4. Avramidis, Manakhimova, Macketanz and Möller (2023) Challenging the State-of-the-art Machine Translation Metrics from a Linguistic Perspective, Association for Computational Linguistics, Singapore, pp. 713–729 URL: https://aclanthology.org/2023.wmt-1.58/.

5. Banerjee and Lavie (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Association for Computational Linguistics, Ann Arbor, Michigan, pp. 65–72 URL: https://aclanthology.org/W05-0909/.

6. Corley and Mihalcea (2005) Measuring the Semantic Similarity of Texts, Association for Computational Linguistics, Ann Arbor, Michigan, pp. 13–18 URL: https://aclanthology.org/W05-1203/.

7. Denkowski and Lavie (2010) Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level, Association for Computational Linguistics, Los Angeles, California, pp. 250–253 URL: https://aclanthology.org/N10-1031/.

8. Firth (1957) A synopsis of linguistic theory, 1930 – 1955. Studies in Linguistic Analysis. Special Volume of the Philological Society, Indiana University Press Bloomington & London, pp. 168–205.

9. Guerreiro, Rei, Stigt, Coheur, Colombo and Martins (2024) xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection, Transactions of the Association for Computational Linguistics, MIT Press, Cambridge, MA, vol. 12, pp. 979–995.

10. Juraska, Finkelstein, Deutsch, Siddhant, Mirzazadeh and Freitag (2023) MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task, Association for Computational Linguistics, Singapore, pp. 756–767 URL: https://aclanthology.org/2023.wmt-1.63/.

11. Kepler, Trénous, Treviso, Vera and Martins (2019) OpenKiwi: An Open Source Framework for Quality Estimation, Association for Computational Linguistics, Florence, Italy, pp. 117–122 URL: https://aclanthology.org/P19-3020/.

12. Klakow and Peters (2002) Testing the correlation of word error rate and perplexity, Speech Communication, vol. 38, no. 1, pp. 19–28.

13. Lin (2004) ROUGE: A Package for Automatic Evaluation of Summaries, Association for Computational Linguistics, Barcelona, Spain, pp. 74–81 URL: https://aclanthology.org/W04-1013/.

14. Lo (2019) YiSi – a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources, Association for Computational Linguistics, Florence, Italy, pp. 507–513 URL: https://aclanthology.org/W19-5358/.

15. Mathur, Baldwin and Cohn (2020) Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics, Association for Computational Linguistics, Online, pp. 4984–4997 URL: https://aclanthology.org/2020.acl-main.448/.

16. Paniv, Chaplynskyi, Trynus and Kyrylov (2024) Setting up the Data Printer with Improved English to Ukrainian Machine Translation, ELRA and ICCL, Torino, Italia, pp. 41–50 URL: https://aclanthology.org/2024.unlp-1.6/.

17. Papineni, Roukos, Ward and Zhu (2002) Bleu: a Method for Automatic Evaluation of Machine Translation, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318 URL: https://aclanthology.org/P02-1040/.

18. Pennington, Socher and Manning (2014) GloVe: Global Vectors for Word Representation, Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543 URL: https://aclanthology.org/D14-1162/.

19. Ramos, Fernandes, Farinhas and Martins (2024) Aligning Neural Machine Translation Models: Human Feedback in Training and Inference, European Association for Machine Translation (EAMT), Sheffield, UK, pp. 258–274 URL: https://aclanthology.org/2024.eamt-1.22/.

20. Rei, Stewart, Farinha and Lavie (2020) COMET: A Neural Framework for MT Evaluation, Association for Computational Linguistics, Online, pp. 2685–2702 URL: https://aclanthology.org/2020.emnlp-main.213/.

21. Siegel (2024) hdaSprachtechnologie/ukrajinet URL: https://github.com/hdaSprachtechnologie/ukrajinet.

22. Snover, Dorr, Schwartz, Micciulla and Makhoul (2006) A Study of Translation Edit Rate with Targeted Human Annotation, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, pp. 223–231 URL: https://aclanthology.org/2006.amta-papers.25/.

23. Thompson and Post (2020) Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing, Association for Computational Linguistics, Online, pp. 90–121 URL: https://aclanthology.org/2020.emnlp-main.8/.

24. Yan, Wang, Zhao, Huang, Chen and Wang (2023) BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training, Association for Computational Linguistics, Toronto, Canada, pp. 5428–5443 URL: https://aclanthology.org/2023.acl-long.297/.

25. Kocmi, Federmann, Grundkiewicz, Junczys-Dowmunt, Matsushita and Menezes (2021) To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation, arXiv URL: http://arxiv.org/abs/2107.10821.

26. Korobov (2015) Morphological Analyzer and Generator for Russian and Ukrainian Languages, arXiv URL: http://arxiv.org/abs/1503.07283.

27. Mikolov, Chen, Corrado and Dean (2013) Efficient Estimation of Word Representations in Vector Space, arXiv URL: http://arxiv.org/abs/1301.3781.

28. Perrella, Proietti, Cabot, Barba and Navigli (2024) Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics, arXiv URL: http://arxiv.org/abs/2410.05183.

29. Rei, Guerreiro, Pombal, Stigt, Treviso, Coheur, Souza and Martins (2023) Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task, arXiv URL: http://arxiv.org/abs/2309.11925.

30. Rei, Treviso, Guerreiro, Zerva, Farinha, Maroti, Souza, Glushkova, Alves, Lavie, Coheur and Martins (2022) CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task, arXiv URL: http://arxiv.org/abs/2209.06243.

31. Sellam, Das and Parikh (2020) BLEURT: Learning Robust Metrics for Text Generation, arXiv URL: http://arxiv.org/abs/2004.04696.

32. Servan, Berard, Elloumi, Blanchon and Besacier (2016) Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources?, arXiv URL: http://arxiv.org/abs/1610.01291.

33. Sun and Wang (2024) Textual Similarity as a Key Metric in Machine Translation Quality Estimation, arXiv URL: http://arxiv.org/abs/2406.07440.

34. Zhang, Kishore, Wu, Weinberger and Artzi (2020) BERTScore: Evaluating Text Generation with BERT, arXiv URL: http://arxiv.org/abs/1904.09675.