Quality Matters: Measuring the Effect of Human-Annotated Translation Quality on English-Slovak Machine Translation

Matúš Kleštinec
Constantine the Philosopher University in Nitra
Tr. A. Hlinku 1, 949 01 Nitra, Slovakia
matus.klestinec@ukf.sk

Daša Munková
Constantine the Philosopher University in Nitra
Tr. A.Hlinku 1, 949 01 Nitra, Slovakia
dmunkova@ukf.sk

Abstract

This study investigates the influence of human-annotated translation quality on the performance of machine translation (MT) models for a low-resource language pair—English to Slovak. We collected and categorized 287 student translations from a national competition, annotated by expert translators into three quality levels. Using the mT5-large model, we trained six neural MT models: three on the full dataset without validation splitting, and three using training/validation splits. The models were evaluated using a suite of automatic metrics (BLEU, METEOR, chrF, COMET, BLEURT, and TER), with TER serving as the validity criterion. Statistical analyses revealed that data quality had no significant effect when training without validation, but did have a significant impact under fine-tuning conditions ($p < 0.05$). Our results suggest that fine-tuning with combination with validation splitting increases the model's sensitivity to the quality of training data. While the overall effect size is modest, the findings underscore the importance of high-quality, annotated corpora and modern training strategies for improving MT in low-resource languages.

1 Introduction

Machine translation (MT) refers to the use of algorithms and machine learning models to translate texts from one natural language into another (Keary, 2023). Modern MT systems increasingly rely on artificial neural networks, which can autonomously learn to perform translation with high accuracy - often achieving levels of accuracy comparable to those of human translators (Young, 2024). Building a high-quality MT model typically requires access to large volumes of training data. Although neural approaches have reached state-of-the-art performance in MT, they suffer from the high cost of acquiring large-scale parallel corpora (Wang et al., 2021).

A neural MT model θ translates a source sentence x into a target sentence y. Using a parallel training corpus C, the model θ is trained by minimizing the negative log-likelihood loss. The encoder-decoder structure (based on recurrent neural networks, convolutional neural networks or transformer) is commonly employed in neural MT, where the encoder transforms the source sentence into a sequence of hidden representations and the decoder generates target words based on these representations and the previously generated target words (Wang et al., 2021). For high-resource language pairs such as English-French, data availability is less problematic, as substantial parallel corpora have been compiled over time. However, the requirement for large amounts of parallel data is often unrealistic for many of the 7000+ languages spoken worldwide, which presents a major challenge for low-resource languages (Ranathunga et al., 2023). The low-resource problem may stem either from a language itself is low-resourced (underrepresented) or from specific domains lack sufficient data (Hedderich et al., 2021).

In the case of the Slovak language, the limited availability of text data categorizes it as a low-resource language (Do et al., 2014). Such languages are underrepresented in digital spaces compared to high-resource languages, making it difficult for speakers to use the advaced technologies in their daily lives - including effective neural MT systems (Tonja et al., 2023).

The research objective:

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 72–81, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-009

The aim of this study is to investigate how both the quality of parallel texts (fair, good, and excellent translations) and the distribution of the dataset (corpus) influence MT system performance, specifically the quality of neural MT output as measured by automatic evaluation metrics.

The structure of this study is as follows: Section 1 introduces the research problem, motivation, and contributions. Section 2 reviews related work on data quality in MT and prior studies on evaluation metrics. Section 3 describes the dataset, tokenization process, model setup, and evaluation metrics. Section 4 presents the experimental results, including statistical analyses of models trained on both the full dataset and split dataset. Section 6 concludes the study and outlines directions for future work.

2 Related work

A recent case study demonstrated that carefully targeted data collection can significantly improve MT performance in a low-resource language pair (Hasan et al. 2020). Data is arguably the most critical factor in modeling (developing) translation systems (Haddow et al., 2022). When applying data-driven MT to a specific language pair, the initial step involves assessing available data resources and identifying effective strategies for collecting additional data. In the context of low-resource MT, Haddow et al. (2022) classify research approaches into four main categories: searching existing data sources, web-crawling for parallel data, data creation, and test data development. In our research, we focus on creating a new parallel dataset comprising student translations from English into Slovak.

Several researchers have explored the use of multiple references in MT. Wu et al. (2024) measured semantic similarity among reference translations and categorized them into different training subsets based on their degree of variation. They fine-tuned two pre-trained large language models - LLAMA-2-7B and mT5-large - using datasets containing multiple references. Their results showed that using source texts with semantic similarity scores between 0.45 and 1.0 led to better performance than unfiltered datasets. Similarly, Zouhar et al. (2021) investigated how the quality and quantity of reference translations affect the reliability of automatic MT evaluation metrics. They found that low-quality or overly diverse references may distort metric scores, whereas carefully selected multiple references enhance evaluation robustness. Our study builds on these findings by combining both perspectives: we employ multiple reference translations per source sentence while accounting for diversity in human-annotated translation quality. Unlike prior studies that primarily focused on semantic variation, we examine how quantity and quality of human-annotated translations influences MT model training and quality of MT outputs.

3 Methodology

3.1 Data collection and pre-processing

The texts used in this study were obtained from the Young Translator public competition, which is open to high school students interested in translation. A total of 287 student translations were included in this study, most of which were translations of literary texts. Two professional translators - both university lecturers in translation and interpreting - evaluated the translations and classified them into three quality categories: 1 – fair translation, 2 – good translation, and 3 – excellent translation. Since the collected translations were available only in printed form, several pre-processing steps were required before training.

The following pre-processing steps were applied:

- Optical character recognition (OCR)
- Text editing for alignment
- Alignment of English and Slovak texts
- Additional text editing prior tokenization and training
- Tokenization

Optical character recognition

Because the original documents were available only as scanned PDFs, it was necessary to convert them into machine-readable text. This was achieved using the Tesseract OCR library (Tesseract OCR, 2025). Although the student translations (essays) were typewritten, many contained handwritten annotations—often in black or colored ink—as part of the evaluation process. In cases where colored pens were used, color filtering was applied to improve OCR accuracy. After recognition, the output was stored in txt format for further processing.

Text editing for alignment

The OCR output required extensive cleaning. Common issues included misrecognized characters, extra punctuation marks (e.g., quotation marks), incorrect spacing (e.g., multiple spaces), and line breaks not corresponding to sentence boundaries. All empty lines were removed to prevent alignment errors. Additionally, the texts were anonymized to remove any personally identifiable information.

Alignment of English and Slovak sentences

After cleaning, the English source texts and Slovak translations were aligned. Each English sentence corresponded to multiple Slovak translations (a 1-to-*n* alignment), reflecting the multiple student versions. To facilitate semantic alignment, we employed LaBSE (Language-agnostic BERT Sentence Embedding), a model trained on more than 100 languages, including English and Slovak (Feng et al. 2020). A similarity threshold of 0.6 was applied.

The aligned data were merged into two larger txt files - one for English and one for Slovak - structured for training purposes. Each model requires one text file in English and one corresponding text file in Slovak. Duplicate sentence pairs were removed, and the dataset was randomly shuffled. Since the number of sentence pairs varied across the three quality categories, the sets for scores 2 and 3 were downsampled to match the smallest set (score 1), ensuring balanced training data and avoiding bias in model evaluation (Table 1). For English and Slovak, the number of words was (54,827 EN | 43,354 SK) for model_1, (54,056 EN | 42,849 SK) for model_2, and (55,514 EN | 42,717 SK) for model_3. Number of tokens was (100,841 EN | 97,284 SK) for model model_1, (101,698 EN | 96,707 SK) for model_2, and (101,802 | 96,406) for model_3.

	Model_1	**Model_2**	**Model_3**
Slovak sentences	3130	3130	3130
English sentences	3130	3130	3130

Table 1: Number of sentences for each model

Tokenization

For tokenization and training of MT models, we utilized the pre-trained *mT5-large* model. This model is based on the transformer architecture and was trained on a multilingual dataset containing sentences from 101 languages, including Slovak (Xue et al., 2020). The *mT5-large* model was used to tokenize both the English and Slovak texts in preparation for training. We selected this model because it covers the English-Slovak language pair and offers a balance between model capacity and training efficiency.

Although the *mT5* model includes Slovak in its pre-training data, fine-tuning on domain-specific datasets is still necessary to achieve optimal performance. In this study, we trained separate models for each quality category, resulting in a total of six models:

1) Three models trained on the full dataset for each category (fair, good, and excellent translations).

2) Three models trained on a split dataset for each category (fair, good, and excellent translations).

All training was conducted on Google Colab using an NVIDIA A100 GPU.

3.2 Models trained on the full dataset

In the first experiment, we trained three models using the entire dataset. Each model corresponded to one of the three quality categories - fair, good, and excellent translations.

The training parameters for these models are summarized in Table 2:

Hyperparameters	**Values**
Per_device_train_batch_size	4
Num_train_epochs	3
Learning_rate	1e-4
fp16	False

Table 2: Hyperparameters for training

After training, three MT models were obtained. Their performance was evaluated using a reference file containing all unique English–Slovak sentences, which had been excluded from the training data to ensure a fair and unbiased evaluation.

3.3 Models with data split

The key difference between the initial three models and the subsequent three lies in the data split

strategy. For these latter models, the dataset was randomly divided into 90% for training and 10% for validation. The 10% validation set was used to fine-tune the models during training. The hyperparameters employed for training all three fine-tuned models are listed in Table 3.

Hyperparameters	Value
Per_device_train_batch_size	4
Num_train_epochs	5
Learning_rate	1e-4
fp16	False
eval_strategy	steps
eval_steps	500

Table 3: Hyperparameters for training

3.4 Evaluation metrics

The trained models were evaluated using a range of automatic metrics: BLEU, METEOR, COMET chrF, TER, and BLEURT.

BLEU (BiLingual Evaluation Understudy) is a precision-based metric that evaluates MT output by comparing n-grams in the hypothesis (MT output) with those in one or more reference translations. It does not consider word order beyond matching n-grams and tends to reward exact matches. A higher BLEU score indicates closer overlap with the reference and therefore better translation quality (Papineni et al., 2002).

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is a metric that aligns words and phrases between the hypothesis and reference translations using synonyms, stemming, and paraphrasing. It calculates scores based on unigram precision, recall, and F-score, which are combined via a weighted harmonic mean. Score ranges from 0 (poor translation) to 1 (perfect translation) (Banerjee et al. 2005).

COMET is a neural framework that considers both source and reference translations. Trained on human judgment data, it predicts sentence-level quality score. This study employed several versions, including wmt20-comet-da, wmt21-comet-da, wmt21-comet-qe-da and wmt22-comet-da. Metric wmt22-comet-da integrates quality estimation techniques using OK/BAD tags from human-annotated datasets and combines multiple models via hyperparameter optimization to produce a single quality score (Rei et al. 2020, Rei et al. 2022). Scores typically range from 0 (poor quality) to 1 (high quality).

BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) is a regression-based evaluation metric built on BERT. Fine-tuned on human ratings of translation quality, it captures subtle semantic differences between translations. BLEURT scores generally range from 0 to 1, though values may occasionally exceed this range due to the nature of the regression output (Sellam et al., 2020).

chrF is a character n-gram F-score metric that evaluates translation quality at the character level rather than the word level. This approach is particularly effective for morphologically rich languages or those with flexible word order. It computes F-scores over character n-grams (e.g., 6-grams), combining precision and recall into a single score, with higher values indicating better translation quality (Popović, 2015).

TER (Translation Edit Rate) measures the number of edits (insertions, deletions, substitutions, and shifts) required to transform the hypothesis into the reference translation. Lower TER score indicates higher translation quality, as fewer edits are needed (Snover et al., 2006).

4 Results

To facilitate interpretation and comparison of MT model performance, the evaluation metrics were grouped according to their scale and underlying evaluation strategy. Three metric groups were defined:

- Group 1 (within-group factor: Metric1*): Includes BLEU, METEOR, chrF, wmt22-comet-da and wmt21-comet-qe-da. These metrics primarily assess surface-level or structural similarity between hypothesis and reference translations.
- Group 2 (within-group factor: Metric2*): Includes BLEURT, wmt20-comet-da, and wmt21-comet-da. These metrics capture deeper semantic similarity, often leveraging pre-trained language representations and human rating data.
- Group 3: TER, treated as a separate metric due to its nature as an error-based measure, serving as a validity criterion for the accuracy metrics.

We hypothesize that statistically significant differences will exist among the examined metrics, between within-group metric (Metric1*/Metric2*)

and the translation quality levels of the training data (between-group factor: quality levels 1–3).

TER as the only metric explicitly measuring edit distance/error rate, is used as a benchmark validity measure to evaluate the reliability and consistency of the other metrics.

4.1 Models trained on full dataset

To assess the assumption of homogeneity of variances across the independent variable (quality levels 1, 2, and 3) a nonparametric Levene's test was used. The results were non-significant (Table 4), indicating that the assumption of equal variances between independent groups was not violated.

	MS Effect	**MS Error**	**F**	**p**
bleu	0.019	0.018	1.038	0.355
meteor	0.004	0.013	0.323	0.724
chrf	0.002	0.011	0.183	0.833
wmt22-comet-da	0.003	0.006	0.447	0.640

Table 4: Levene' test for homogeneity of variances

However, the assumption of sphericity - which concerns the equality of variances of the differences between all combinations of dependent metrics in Metric1 (BLEU, METEOR, chrF, and wmt22-comet-da) - was violated (Table 5).

	W	**Chi-Sqr.**	**df**	**p**
Metric1	0.496	396.360	5	0.0000

Table 5: Mauchley sphericity test

To preserve statistical power and ensure the validity of the analysis, adjusted univariate tests for repeated measures were applied. These tests evaluated the effects of translation quality and its interaction with evaluation metric (metric1 × quality) on translation performance (Table 6).

	Epsilon	**Adj. df1**	**Adj. df2**	**Adj. p**
Metric1	0.705	2.116	1199.867	0.0000
Metric1 x quality	0.705	4.232	1199.867	0.4542

Table 6: Adjusted (G-G) univariate tests for repeated measure

Statistically significant differences ($p < 0.05$) were observed only among the evaluation metrics themselves (Table 6). The effect of between-group factor (quality level) did not have a statistically significant on the evaluation outcomes ($p > 0.05$), indicating that the quality categories (1, 2, and 3) did not significantly influence scores across the metrics.

The results of the multilevel comparison (Table 7) further clarify the relative behavior of individual metrics. Specifically, a statistically significant difference was found between the BLEU metric and remaining metrics, whereas no statistically significant difference was observed between meteor and chrf metrics. Even when considering the quality levels (Table 7), no statistically significant differences were found between the individual quality categories (1, 2, and 3) with respect to the metrics included in Metric1 (BLEU, METEOR, chrF, and wmt22-comet-da).

Quality	**Metric1**	**mean**	**1**	**2**	**3**
3	bleu	0.255		****	
1	bleu	0.282		****	
2	bleu	0.299		****	
3	meteor	0.593	****		
3	chrf	0.598	****		
1	meteor	0.606	****		
1	chrf	0.616	****		
2	meteor	0.628	****		
2	chrf	0.635	****		
1	wmt22-comet-da	0.820			****
3	wmt22-comet-da	0.825			****
2	wmt22-comet-da	0.837			****
Note: **** - p > 0.05, homogeneous group					

Table 7: Multi-stage comparison

We applied the same analytical procedure to Group Metric2, taking into account deviations from the assumption of normality.

Statistically significant differences were observed only among the evaluation metrics within-group Metric2 ($p = 0.000$). The effect of the between-group factor (translation quality level) on evaluation scores was not statistically significant ($p = 0.552$), indicating that the assigned quality

categories did not influence the metric scores in this group. Statistically significant differences were found between all three metrics in this group ($p < 0.05$). When incorporating the translation quality factor, no significant interaction effects based on translation quality were observed (Table 8).

quality	metric2	Mean	1	2	3
3	wmt21-comet-da	0.050	****		
1	wmt21-comet-da	0.056	****		
2	wmt21-comet-da	0.071	****		
3	bleurt	0.096	****	****	
1	bleurt	0.134	****	****	
2	bleurt	0.179		****	
3	wmt20-comet-da	0.625			****
1	wmt20-comet-da	0.638			****
2	wmt20-comet-da	0.696			****
Note: **** - p > 0.05, homogeneous group					

Table 8: Multi-stage comparisons

metric2	MS Effect	MS Error	F	p
bleurt	0.019	0.055	0.346	0.7077
wmt20-comet-da	0.053	0.110	0.487	0.6147
wmt21-comet-da	0.000	0.008	0.001	0.9994

Table 9: Levene' Test for Homogeneity of Variances

	W	Chi-Sqr.	df	p
metric2	0.728	179.615	2	0.0000

Table 10: Mauchley Sphericity Test

	Epsilon	Adj. df1	Adj. df2	Adj. p
metric2	0.786	1.572	891.567	0.0000
metric2 x quality	0.786	3.145	891.567	0.5521

Table 11: Adjusted (G-G) Univariate Tests for Repeated Measure

Similar to the first group of metrics, deviations from normality were identified for the second group of metrics. Based on the results of the nonparametric Levene's test (Table 9), we conclude that the assumption of equality of variances between independent samples (quality: 1, 2, and 3) is not violated. In the case of dependent samples (metric2: bleurt, wmt20-comet-da, wmt21-comet-da), the sphericity condition of the covariance matrix was violated (Table 10). In order not to reduce the power of the statistical tests, we use adjusted univariate tests for repeated measures (Table 11) to assess the quality of the translation as a function of the interaction of the within-group and between-group factors (metric2 x quality).

Statistically significant differences were observed only among the metrics themselves ($p < 0.05$), while the between-group factor, translation quality, did not have a significant effect on evaluation outcomes ($p > 0.05$) (Table 8). A multilevel comparison (Table 8) indicates that the wmt21-comet-da metric is statistically the most rigorous metric ($p < 0.05$), whereas wmt20-comet-da is statistically the least rigorous ($p < 0.05$). Statistically significant differences were found between all three metrics ($p < 0.05$).

The reliability analysis of the MT assessment procedure indicates that the selected set of evaluation metrics - BLEU, METEOR, chrF, BLEURT, wmt22-comet-da, wmt20-comet-da, and wmt21-comet-da - demonstrates acceptable internal consistency (*Cronbach's* $\alpha > 0.6$), suggesting that the metrics collectively form a coherent measurement construct (*Average inter-item corr.* > 0.5).

The MT evaluation procedure explains nearly 70% of the variability in MT error rate (Table 12). Based on the validity analysis (Table 12), the procedure demonstrates acceptable criterion validity. The TER metric, which directly represents MT error rate, was employed as the validity criterion (Munk et al., 2018), confirming that the combined use of BLEU, METEOR, chrF, BLEURT, wmt22-comet-

da, wmt20-comet-da, and wmt21-comet-da provides a valid estimation of translation accuracy.

	Summary for scale
Multiple R	0.830
Multiple R2	0.689
F(7,562)	178.169
p	0.0000

Table 12: Validity analysis

4.2 Models with data split

As with the first three models, evaluation of the split-data models was performed using the same reference file. Due to deviations from normality and differences in the range of the evaluated scores, the metrics were again divided into three groups, following the same grouping strategy as in the first experiment.

We hypothesize that statistically significant differences will exist between within-group metric (Metric1*/Metric2*) and the translation quality levels of the training data (between-group factor: quality levels 1–3).

As in the previous analysis, the TER metric (ter_ref.) was employed as the validity criterion, since it directly reflects the MT error rate. As in the previous analysis, we observed a violation of the sphericity assumption for the covariance matrix, as indicated by the Mauchly's Test of Sphericity ($p < 0.05$), which pertains to the use of repeated (dependent) measures (metric1*: bleu_ref., meteor_ref., chrf_ref., wmt22-comet-da, wmt21-comet-qe-da).

We applied adjusted univariate tests for repeated measures to evaluate translation quality as a function of the interaction between within-group (metric1)* and between-group (quality level) factors. The results indicated statistically significant differences among the evaluated metrics ($p = 0.000$), as well as a significant effect of translation quality on the evaluation outcomes ($p = 0.004$).

Multilevel comparisons (Table 13) further confirmed statistically significant differences among all metrics ($p < 0.05$). Additionally, a significant effect of translation quality was observed across nearly all metrics, except wmt21-comet-qe-da, for which the effect was not statistically significant ($p > 0.05$).

quality	**metric1***	**Mean**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	wmt21-comet-qe-da	0.106	****						
2	wmt21-comet-qe-da	0.107	****						
1	wmt21-comet-qe-da	0.107	****						
3	bleu_ref.	0.218		****					
1	bleu_ref.	0.271			****				
2	bleu_ref.	0.283			****				
3	meteor_ref.	0.546				****			
3	chrf_ref.	0.555				****			
1	meteor_ref.	0.596					****		
2	meteor_ref.	0.611					****		
1	chrf_ref.	0.619					****		
2	chrf_ref.	0.627					****		
3	wmt22-comet-da	0.801						****	
1	wmt22-comet-da	0.821						****	****
2	wmt22-comet-da	0.842							****
Note: **** - p > 0.05, homogeneous group									

Table 13: Multi-stage comparisons

quality	**metric2***	**Mean**	**1**	**2**	**3**	**4**
8	bleurt_ref.	0.019	****			
8	wmt21-comet-da	0.024	****			
6	wmt21-comet-da	0.062	****			
7	wmt21-comet-da	0.072	****			
7	bleurt_ref.	0.172		****		

6	bleurt_ref.	0.174		****		
8	wmt20-comet-da	0.54				****
6	wmt20-comet-da	0.65			****	
7	wmt20-comet-da	0.708			****	
Note: **** - p > 0.05, homogeneous group						

Table 14: Multi-stage comparisons

For the second group of metrics (metric2*: bleurt_ref., wmt20-comet-da, wmt21-comet-qe-da), the sphericity assumption was also violated (Mauchley Sphericity Test: $p < 0.05$). In order not to reduce the power of the statistical tests, we use modified univariate tests for repeated measures to assess the quality of the translation as a function of the within-group and between-group interaction (metric2* x quality) (Table 15).

	Epsilon	**Adj. df1**	**Adj. df2**	**Adj. p**
metric2*	0.787	1.575	892.795	0.0000
metric2* x quality	0.787	3.149	892.795	0.0130

Table 15: Adjusted (G-G) univariate tests for repeated measure

Statistically significant differences (Table 15) were again demonstrated between the metrics themselves ($p < 0.05$), and the effect of translation quality was likewise significant ($p < 0.05$).

When including translation quality as a factor in the multilevel comparison (Table 14), a statistically significant influence of quality level was confirmed for almost all metrics, with the exception of wmt21-comet-da ($p > 0.05$).

	Summary for scale
Multiple R	0.890
Multiple R2	0.792
F(7,562)	267.144
p	0.0000

Table 16: Validity analysis

The MT evaluation procedure explains nearly 80% of the variability in the MT error rate (Table 16). Based on the results of the validity analysis (Table 16), we conclude that the procedure demonstrates acceptable criterion validity. The TER metric was employed as the validity criterion (Munk et al., 2018), confirming that the combined use of BLEU, METEOR, chrF, BLEURT, wmt22-comet-da, wmt20-comet-da, and wmt21-comet-da provides a valid estimation of translation accuracy.

5 Conclusion

The study demonstrates that the quality of annotated training data influences the performance of neural MT systems for the English–Slovak language pair. However, the extent of this effect depends strongly on the training strategy. When models were trained on the full dataset without validation splitting, translation quality level showed no significant impact on performance ($p > 0.05$). In contrast, when the dataset was split into training and validation subsets, translation quality level significantly affected the evaluation metrics ($p < 0.05$). This suggests that fine-tuning with held-out validation data increases the model's sensitivity to training data quality.

Despite minor deviations and variations across individual metrics, the overall evaluation procedure explains a significant proportion of the variance in translation error rates. These findings indicate that for low-resource languages such as Slovak, enhancing the quality of human-annotated parallel corpora can lead to measurable gains in MT performance - particularly when modern training strategies like fine-tuning on held-out validation sets are employed. Nonetheless, the effect size remains relatively small, and further improvements may require not only higher-quality data, but also larger and more diverse training corpora.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-23-0554.

References

Anthony Young, 2024. Exploring *Machine Translation: Output Quality, Learner Reflection, Teacher Detection*. In *Thailand TESOL Conference Proceedings 2024.*

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh and Grigori Sidorov. 2023. *Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data.* In *Applied Sciences.* 2023; 13(2):1201. https://doi.org/10.3390/app13021201

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl and Alexandra Birch; *Survey of Low-Resource Machine Translation*. In *Computational Linguistics 2022; 48 (3):* 673–732. https://doi.org/10.1162/coli_a_00446

Britannica (2025): *OCR*. https://www.britannica.com/technology/OCR

Cong-Thanh Do, Lori Lamel and Jean-Luc Gauvain 2014. *Speech-to-text development for Slovak, a low-resourced language*. In 4th Workshop on Spoken Language Technologies for Under-resourced Languages, pages 176-182.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan and Wei Wang 2020. *Language-agnostic BERT Sentence Embedding.* In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics. arXiv:2007.01852. Version 2

Hasan, Tahmid, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. *Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation.* In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pages 2612–2623. https://doi.org/10.18653/v1/2020 .emnlp-main.207

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *A Method for Automatic Evaluation of Machine Translation*. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL),* pages 311-318. Philadelphia, Pennsylvania. https://doi.org/10.3115/1073083.1073135

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua and Colin Raffel. *2020. mT5: A massively multilingual pre-trained text-to-text transformer*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, pages 483-498. arXiv:2010.11934

Maja Popović. 2015. *chrF: character n-gram F-score for automatic MT evaluation*. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. *A Study of Translation Edit Rate with Targeted Human Annotation.* In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*. 2545–2568.

Michal Munk, Daša Munková and Lubomír Benko. 2018. *Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics.* In *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology 34(5)* https://doi.org/10.3233/JIFS-169505

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur and André F. T. Martins. 2022. *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*. In *Proceedings of the Seventh Conference on Machine Translation (WMT).* Association for Computational Linguistics, pages 578-585. Abu Dhabi, United Arab Emirates.

Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. 2020. *COMET: A Neural Framework for MT Evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 2685-2702.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin and Tie-Yan Liu. 2021. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Survey Track*. Pages 4636-4643.

Satanjeev Banerjee and Alon Lavie. 2023. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In *Proceedings of the ACL Workshop on Instrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, pages 65-72.

Si Wu, John Wieting and David A. Smith. 2024. *Multiple References with Meaningful Variatons Improve Literary Machine Translation*. arXiv:2412.18707

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar and Mehreen Alam, Rishemjit Kaur. 2021. *Neural machine translation for low-resource languages*: In *A survey." ACM Computing Surveys 55(11),* pages 1-37. https://doi.org/10.48550/arXiv.2106.15115

Tesseract OCR (2025): Tesseract-ocr/tesseract. Github. https://github.com/tesseract-ocr/tesseract

Tim Keary. 2023. *Machine Translation (MT)*. https://www.techopedia.com/definition/machine-translation-mt

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. *BLEURT: Learning Robust Metrics for Text Generation.* In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Vilém Zouhar, Ondřej Bojar. 2024. *Quality and Qunatity of Machine Translation References for Automatic Metrics.* In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 1-11. arXiv:2401.01283. Version 4