DOI: https://doi.org/10.31392/NPU-nc.series9.2022.24.04
UDC: 81'25

Milda Povilaitienė
MA in Translation,
Faculty of Social Sciences, Arts and Humanities,
Kaunas University of Technology,
Kaunas, Lithuania,
https://orcid.org/0000-0003-2077-7493
e-mail: milda@baltijosvertimai.eu

Ramunė Kasperė
Dr. (Linguistics), Professor,
Vice Dean for Research,
Faculty of Social Sciences, Arts and Humanities,
Kaunas University of Technology,
Kaunas, Lithuania,
https://orcid.org/0000-0003-0782-3758
e-mail: ramune.kaspere@ktu.lt

MACHINE TRANSLATION FOR POST-EDITING PRACTICES

Bibliographic Description:

Povilaitienė, M., Kasperė, R. (2022). Machine Translation for Post-Editing Practices. *Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development*, 24, 47–62. https://doi.org/10.31392/NPU-nc.series9.2022.24.04

Abstract

It has been proven in many quality-focused studies that machine translation output in some language pairs is still far from publishable (Koponen, 2016). Even so, post-editing has become a daily practice among translators and translation service providers, especially with language pairs where machine translation demonstrates good human parity. The fast development of machine translation and its quality improvement

have led to a growing demand of post-editors. This study attempts to evaluate the quality of the most popular machine translation tools for the Lithuanian language in order to find the correlation between the results of automatic quality estimation (i.e., the BLUE score), human / manual evaluation of machine translation output quality following the multidimensional quality metrics (MQM) and the most common machine translation engines used by freelancers and language service providers.

The conclusions are based on the findings of a survey and the automatic vs human / manual machine translation quality analysis. The findings demonstrate and support previous research that automatic machine translation quality estimation may not be taken for granted. Human / manual machine translation quality evaluation is still a better indicator whether a machine translation tool fits the purpose of translation. The study brings to the fore some insightful findings that may be beneficial for translator and post-editor trainers from the pedagogical perspective as well as for translation industry from the practical perspective.

Keywords: *machine translation, post-editing, BLEU score, multidimensional quality metrics (MQM).*

1. Introduction.

Researchers agree that the quality of machine translation may be far from publishable in some language pairs (Koponen, 2016), yet it is improving in a galloping manner. There are speculations that in the near future translators may be able to translate many more pages a day instead of a dozen. There will be less post-editing effort after machine translation, fewer errors in the raw machine translated text, inconsistencies, distortions of meaning or failed translations overall.

Using machine translation as an aid, a source of information, a dictionary and a reference has become a daily routine in the translator's work. It would be irrational not to take advantage of this technology. Thus, machine translation post-editing is becoming a common practice among translators and translation service providers worldwide. However, it must also be acknowledged that in some language pairs, especially in low-resource languages where machine translation does not reach good human parity, the machine-processed text must be treated with the utmost caution.

Widespread and transparent use of and trust in machine translation post-editing practice is making its first steps in Lithuania, hence there is the need to analyze post-editing practices. It is essential as good quality of machine translation output creates a good experience in post-editing and thus increases the number of machine translation post-editing users and productivity of translators.

2. Literature Review.

Following the European Language Industry Survey (EUATC, 2018), the year 2018 was that of machine translation, since more than half of the respondents declared using the technology in one or another way. Despite negative attitude towards machine translation of some translation industry participants, reported by Levanaitė (2021), the urge for machine translation use has been increasing rapidly, and thus, the need of well-trained post-editors, relatively a new player on the board of translation industry, and the need to evaluate the quality of machine translation output are emerging.

A number of studies have been conducted to evaluate the impact of machine translation and to find out the degree to which machine translation and post-editing shows potential in being implemented in the market of language services. The TAUS survey of 2010, overviewing language service providers operating in Europe, America and Asia, revealed that almost half (49.3%) of the 75 respondents offered machine translation post-editing services (Joscelyne & Brace, 2010, in Christensen & Schjoldager, 2016). Note that at the time neural machine translation did not yet exist. The report on machine translation market as of 2014 showed that 38.63% of language service providers offered machine translation services (Van deer Meer & Ruopp, 2014, in Christensen & Schjoldager, 2016). The Translation Technology Insights Research, carried out in 2016, demonstrated that 40% of the total of 2784 respondents used machine translation. According to the results of

the ELIS survey (2022) where answers of 1342 respondents have been obtained, standard human translation is prevalent, but post-editing has been ranked second (21%); it has also been found to have the highest growth potential (64%).

According to the International Standard for Translation Services – post editing of machine translation output – Requirements (ISO 18587, 2017), machine translation refers to automatic translation of the text by using a computer application. The standard states that usage of machine translation allows language service providers to improve translation productivity, times resources and to competitiveness with clients who are increasingly demanding to use machine translation in translation processes. The Standard ISO 18587 (2017) also emphasizes that there is no machine translation engine offering output that could be qualified equal to the one produced by a human translator. Therefore, the final quality of the machine translation output still depends on qualification of post-editors. Although the Standard (ISO 18587, 2017) defines two types of post editing, i.e. full (process that seeks to obtain an output comparable to that obtained by human translation) and light (process that aims to obtain a merely comprehensible text), it must be mentioned that the confines between the two types have been decreasing with the increasing quality of machine translation when the output requires almost no post-editing effort for the light version.

There is some research covering Europe's existing practices of machine translation post-editing (e.g., EUATC, 2022, among others); however, there are not many details on smaller member countries with less populous languages. It should not come as a surprise that languages like Lithuanian have been receiving less attention from machine translation developers. The smaller number of language users, the less corpus is available. Thus, the quality of machine translation for the Lithuanian is not that great compared with machine translation quality for the most popular language pairs. According to Maučec and Donaj (2019), machine translation does not work well with morphologically rich languages, "especially when translation is done from a morphologically less complex to a morphologically more complex language". Being a morphologically rich language, Lithuanian, like many other languages spoken by fewer people, is categorized as a low-resource language.

There are several machine translation systems operating with Lithuanian, like *DeepL, Google Translate, Tilde, Globalese, Microsoft Translator, eTranslation*, among many. Most of them can be integrated in CAT tools, thus providing post-editors with a comfortable environment for post-editing and generating translation memories. According to the data provided in the ELIS 2022 survey, the most popular machine translation systems in Europe are *DeepL, Google Translate, Globalese* and *Microsoft Translator* (EUATC, 2022).

As machine translation and post-editing are playing an increasingly important role in today's routine of translators and post-editors, they are to make a decision choosing the best machine translation system for the purpose. The question what quality of machine translation output is good enough is of paramount importance. The process of quality assessment is complicated due to cognitive, social, cultural, and technical linguistic peculiarities. There is still no clear consensus on the agreement on various machine translation evaluation – both human and machine – criteria (fluency, adequacy, meaning, severity, usefulness, etc.) and on what kind of an assessment approach is the best (Castilho et al., 2018). It has been accepted that "a single gold standard measure of quality" is non-existent (Way, 2018).

Yet before even undertaking post-editing, it is important to identify weaknesses of machine translation output as it might help to improve the output quality, especially for low-resource and morphologically rich languages, such as Lithuanian. With increasingly improving quality of machine translation systems where differences between them are hard

to spot instantaneously, it becomes more and more difficult to select the best solution for translation (Rossi & Carré, 2022).

The output of machine translation can be evaluated by human, automatically or combining both manual and automatic efforts. The aim of automatic evaluation is to compare a machine translated text to its similarity to one or more reference translations created by humans, that, because of a large number of corpora, are assumed to be correct (Castilho et al., 2018). Despite the fact that there is no one single correct translation, machine translation output quality is compared to the reference translations. During the automatic evaluation, an algorithm calculates the evaluation score. The calculated score is an indication to the user of the level of a translation quality. The methods of automatic translation evaluation count automatically detected inaccuracies on word and/or sentence level, but not on the general text level (Maučec & Donaj, 2019).

Among the most popular metrics used for automatic evaluation of machine translation output quality is Bilingual Evaluation Understudy (BLEU) (Castilho et al., 2018), which has been used widely for the past decades. The metric has been found to be a valid evaluation technique based on an assumption that it may well predict the appropriateness of machine translation systems (Reiter, 2018). Some research studies have reported varying correlation findings with human evaluation, both good and bad (Coughlin, 2003; Dey et al., 2022; Reiter, 2018; Mathur et al., 2020).

BLEU scores are in the range from 0 to 1 or from 0 to 100. The score above 0.30 or 30 shows that the translation is understandable while the score above 0.50 or 50 shows that translation is good and fluent (Seljan, 2012). The BLEU metric is statistically based and is suitable for every language. Yet, it does not consider lexical relations of words (Chauhan et al., 2021), which is an important issue for highly inflective languages such as Lithuanian. Since BLEU scores have been proven to correlate less well with human evaluations in such languages, researchers and developers have proposed alternative or modified evaluation scores, e.g. AdaBLEU, which takes into account lexical and syntactical properties of morphologically rich languages (Chauhan et al., 2021). Besides, the reliability of the metrics with low-resource languages has not been confirmed either (Kocmi et al., 2021).

However, the most common and more reliable way to evaluate machine translation quality is human evaluation, which has been for long considered a golden standard and its superiority has been proven in numerous case studies. Even though there has been considerable research on the topic, researchers find a lack of a commonly accepted standard procedure (Freitag, 2021). To evaluate the quality of human and machine translations, a framework called multidimensional quality metrics (MQM) (Lommel et al., 2013) was developed. European Union-funded QTLaunchPad project has developed the Multidimensional Quality Metrics (MQM) that helps to organize many types of translation errors into a particular hierarchy. This framework includes more than 100 issue types. Human evaluators must meet certain requirements, so that results of evaluation could be counted as valid. Evaluators have to be trained and familiar with the subject area of the texts (Rivera-Trigueros, 2021). Thus, it is a time-consuming, expensive and to some extent very subjective task.

It is assumed that the best way to establish the quality of the machine translation output is to apply a combination of methods for a short extract of a longer text and to choose between the most realistic and pragmatic method. However, if the results differ significantly, it is better to rely on human quality evaluation. If the results of automatic estimation and manual evaluation of the quality are similar, the rest of the machine translation output could be evaluated automatically.

3. Aim and Objectives.

The study attempts to evaluate the quality of the most popular machine translation tools in Lithuania in order to find the correlation between the results of automatic and human machine translation output quality evaluation and the use of the most common machine translation engines.

The ***objectives*** are as follows:

– to analyze the post-editing practices in Lithuania in order to find out the prevalence of machine translation usage and identify the most popular machine translation engines among the professional users;

– to evaluate machine translation output quality of the most popular machine translation engines identified using automatic and human/manual machine translation output quality evaluation metrics;

– to compare the results of manual and automatic machine translation output quality evaluation in order to find out if they match and if automatic quality evaluation can be trusted.

4. Methodology.

In this study, qualitative research methods are employed along with frequency calculations. The research combines the survey of language service providers and freelance language specialists about machine translation and post-editing practices and analysis of the machine translation output quality. Then the quality of machine translation output was measured using BLEU metrics (automatic quality evaluation) and MQM metrics (human quality evaluation) on a text consisting of 67 segments. The automatic testing with the BLEU tool was performed with an interactive BLEU score evaluator. To obtain a human reference translation, a translator with 14 years of experience translated a selected text from English into Lithuanian. Qualitative data analysis by segments might also yield valuable findings where the significant differences between automatic and human / manual quality evaluation are noted (Zaretskaya et al., 2020). Qualitative analysis of some segments is provided for illustration purposes.

5. Results.

5.1. Results of the Survey.

The survey was open for 2 months in March through April, 2022, and was advertised via various channels including Facebook groups of translators. 142 answers were collected from Lithuania (110 from freelance language specialists and 31 from language service providers). Out of them, almost three-fourths of freelancers (n = 78) indicated using machine translation on a daily basis. Meanwhile, representatives of language service providers (heads, owners and/or project managers) were almost equally divided into users (n = 16) and non-users (n = 15) of machine translation in routine business (see Figure 1), which may imply that language service providers are not fully aware of its drawbacks, benefits and associated risks.

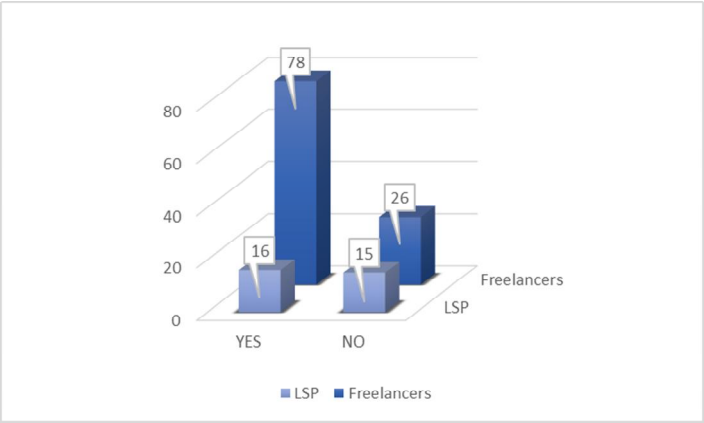

Figure 1. Use of machine translation in everyday practice and business

Both target groups were asked a question regarding the reasons behind using machine translation post-editing (see Figure 2). The main reason chosen by language service providers (n = 9) and freelancers (n = 12) was indicated to be poor quality or raw machine translation output, followed by lack of knowledge about machine translation post-editing (n = 6, in case of LSPs, vs. n = 7, in case of freelancers). Freelancers (n = 9) also noted low rates given for post-editing as the second main reason why they did not use it in routine practice.

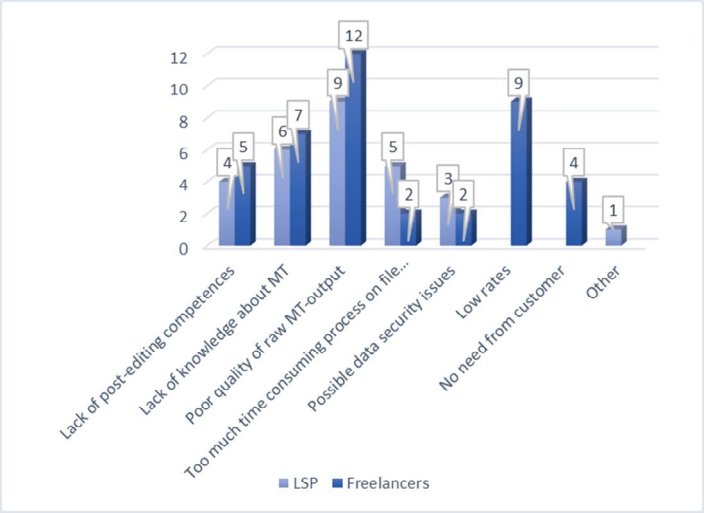

Figure 2. Reasons why machine translation post-editing is not used

Freelancers were also asked to indicate the main reasons why they chose to do machine translation post-editing (see Figure 3). Over 40% (n = 33) indicated to do post-editing upon a request from a customer and more than half (n = 45) indicated to perform on their own initiative, which might be considered a finding causing concern. If freelancers choose to do post-editing without the customer's consent, such practice may lead to damaged reputation on the translator's and or language service providers' part. Besides, an assumption could be made that freelancers engaging in such practice are not fully aware how to perform post-editing following the standards and good practices and/or do not completely understand the risks and threats exposing the data to freely available machine translation tools may bring about.

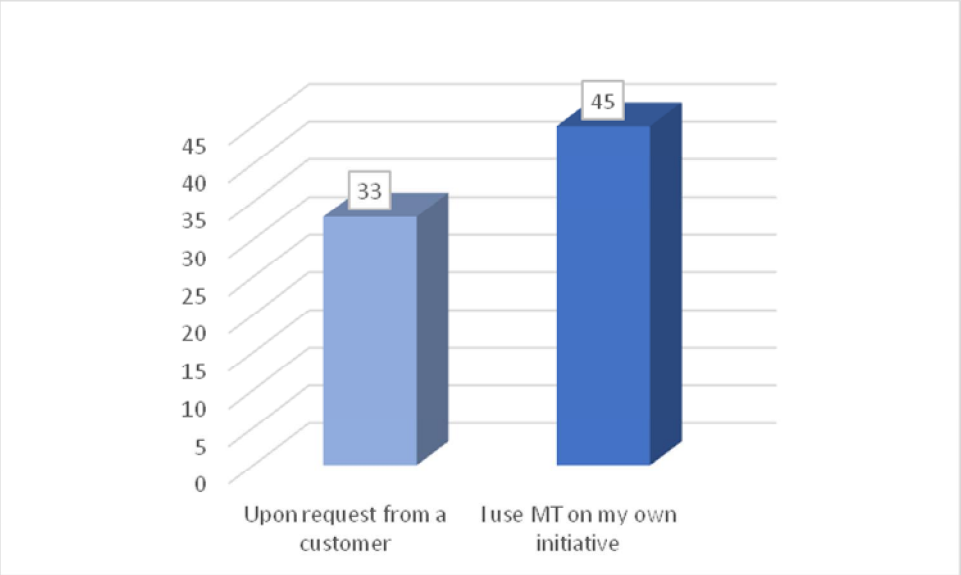

Figure 3. Reasons why machine translation post-editing is used indicated by freelancers

Among the aims of the survey was the question about the most common machine translation tool employed by LSPs and freelancers. *DeepL* and *Google Translate* were mentioned most frequently (n = 36 each). These options were followed by a good number of the respondents indicating CAT-embedded (n = 32) or client-provided (n = 31) machine translation tools. *Tilde* and *eTranslation* were less often selected options (n = 14 and n = 12, respectively). These findings might point to a few assumptions. On the one hand, *DeepL* and *Google Translate* are best known and most popular machine translation tools, the fact confirmed in many other studies with professionals and professional users of machine translation (Kasperė et al., 2021; Vieira et al., 2021). On the other hand, CAT-embedded and client-provided machine translation tools mentioned by approximately one-third of our respondents are safer options, which indicates more transparency and consideration given as to the use of machine translation post-editing in routine practice and business.

5.2. Machine Translation Quality Assessment.

Following the survey results, in the second stage of the study, quality assessment of raw machine translation output was performed. First, machine translation output quality was automatically evaluated with the BLEU tool. An extract 1015 words length was selected from a BBC article in order to use it in machine translation and to find out if the translation provided is accurate and renders meaning properly. The output of *DeepL*, *Google Translate* and *Tilde* machine translation systems was taken for analysis. As it can be seen from Figure 4, according to the interactive BLEU score evaluator, the best quality of the output is that of the machine translation performed with *Google Translate* (the obtained score is 39.12), followed by *Tilde* (37.72) and *DeepL* (27.64).

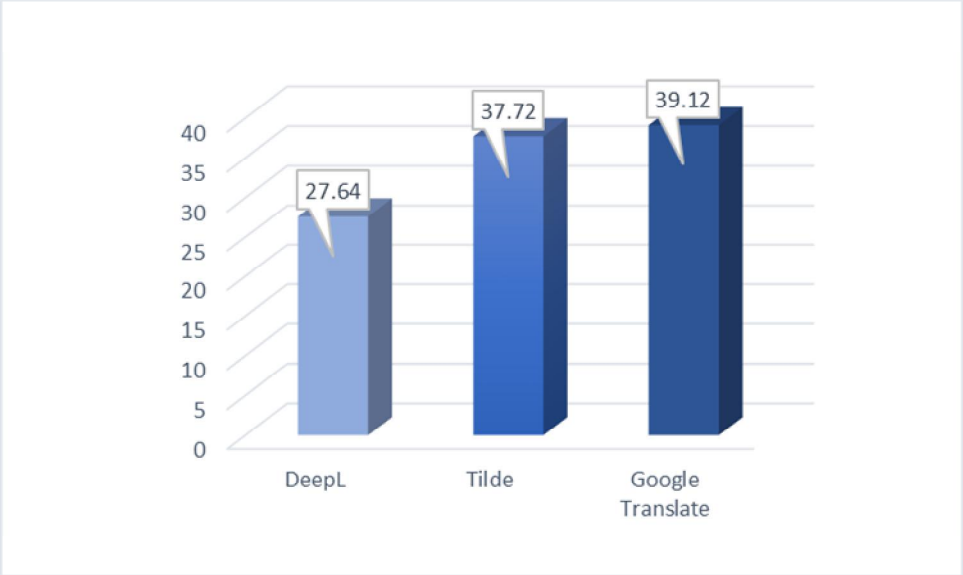

Figure 4. BLEU scores for machine translation output quality in DeepL, Google Translate, and Tilde

As mentioned above, the closer the BLEU score is to 100, the better the quality it reports. However, only translations that receive the BLEU score as high as 50 may be considered of sufficient quality. This means that, in our case, none of the three machine translation systems employed demonstrated sufficient quality. The BLEU score below 30 indicates very low quality, and in our case, the machine translation from *DeepL* falls within this category. This might be taken as a surprising fact since generally it is known and considered that *DeepL* is not only one of the most popular, but also most trusted machine translation system.

On the other hand, automatic evaluation is quite superficial. It does not count the severity of errors. It only counts the distance of machine translation from human reference translation, evaluating and comparing it sentence by sentence. It is well known that there is no such thing as one correct version of the translation. Therefore, even two correct translation versions of the same source text may result in different BLEU scores. For this reason, such automatic evaluation may only be taken as indicative, but not definitive.

Further, the translation provided by the machine translation systems *DeepL*, *Google Translate* and *Tilde* were analysed using MQM metrics. The analysis of machine translated files was carried out in the TAUS error typology evaluation template (retrieved from https://info.taus.net/dqf-mqf-error-typology-template-download). The higher the score, the worse the quality of the machine output.

The DeepL output showed the worst result after testing it with interactive BLEU quality evaluator. However, the human analysis of the machine translation output quality with MQM metrics demonstrated the highest quality score with DeepL – 139. The score was calculated taking into the account the number and the severity level of errors. There was the highest number of errors totally (47 errors), but the severity of the errors was mostly neutral (25 errors). 10 minor errors and 13 major errors were found out. No critical errors were found (see Figure 5).

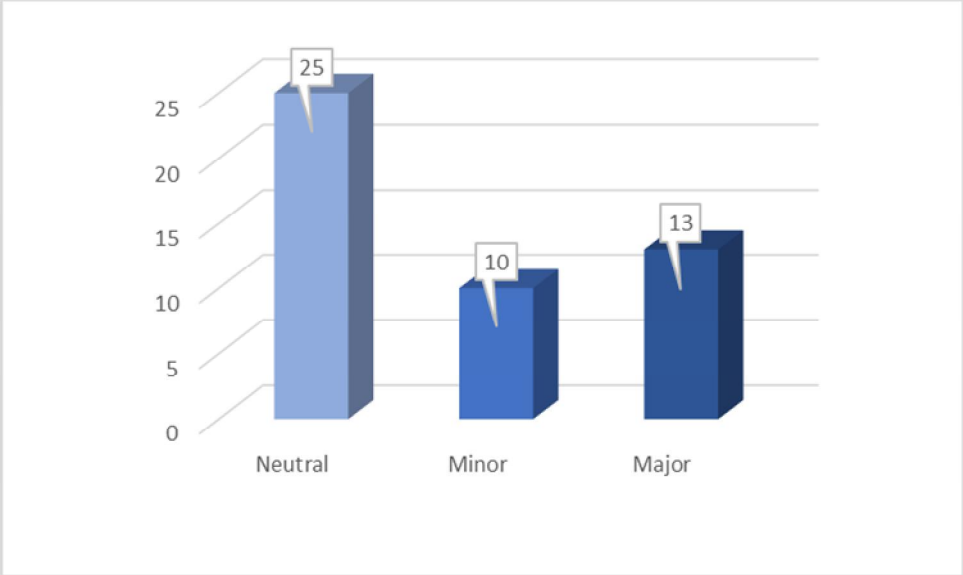

Figure 5. DeepL machine translation output error severity with MQM metrics

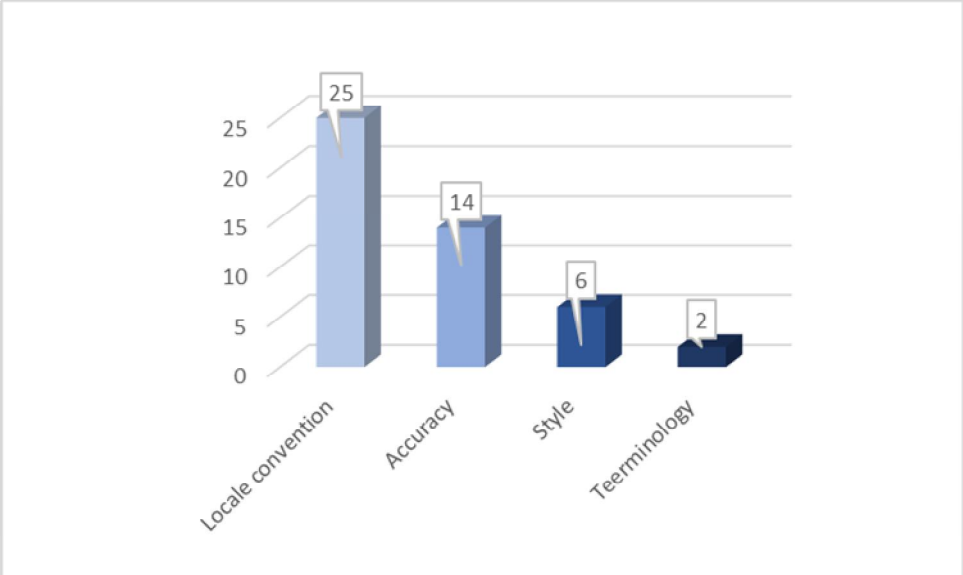

Figure 6. DeepL machine translation output error categories with MQM metrics

Most of the errors found were locale convention errors, because of the non-localised quote marks (25 errors) that fall under category of neutral errors, 14 accuracy errors (13 major and 1 minor accuracy error), 6 style errors falling under minor error category, and 2 terminology errors falling under minor category (see Figure 6). Of all three machine engines tested, only *DeepL* did not localize quote marks. Translations with the most quote marks which were left non-localized earned some of the smallest BLEU scores (4.24), e.g.:

Example 1

Source: *The BBC asked to interview Apple, Google, Meta and Twitter, but Google and Twitter declined to comment, and Apple and Meta didn't respond.*

DeepL output: *BBC paprašė apklausti "Apple", "Google", "Meta" ir "Twitter", tačiau "Google" ir "Twitter" atsisakė komentuoti, o "Apple" ir "Meta" neatsakė.*
Human translation: *„BBC“ paprašė „Apple“, „Google“, „Meta“ ir „Twitter“ interviu, tačiau „Google “ ir „Twitter“ komentuoti atsisakė, o „Apple“ ir „Meta“ apskritai neatsakė.*

Non-localised quote marks in this case do not change the meaning and would not result in a substantial workload for a post-editor. Therefore, this type of errors can be classified as neutral errors and is easy to correct.

Google Translate machine translation engine showed the highest quality result while measuring the machine translation quality level with BLEU metrics; yet, the human quality analysis revealed a score of 262. The total amount of errors in the selected sample was 37. The text included 25 major errors and 12 minor errors (see Figure 7). Most of the major errors were related to the accuracy and mistranslation of the text, and minor errors were mostly related to the fluency.

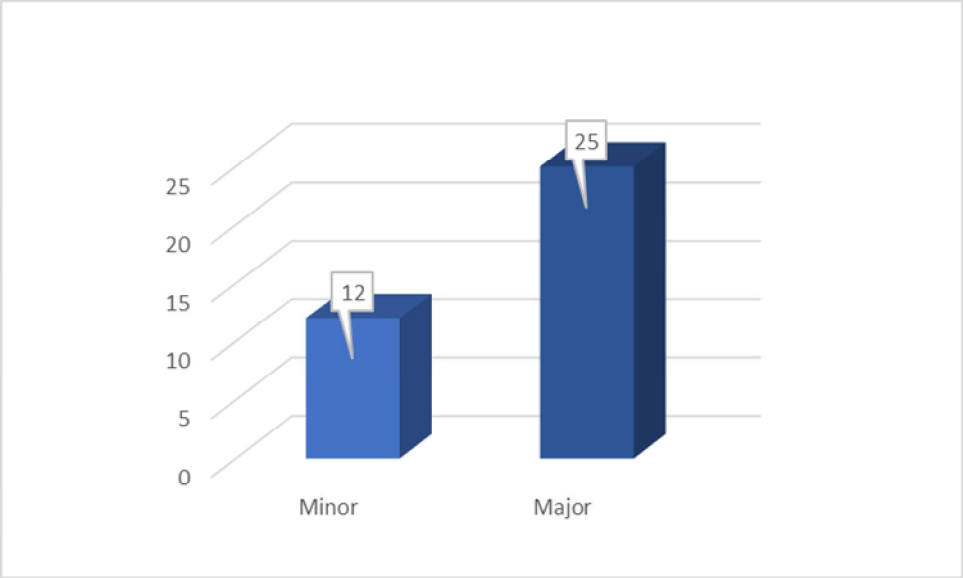

Figure 7. Google Translate machine translation output error severity with MQM metrics

The *Google Translate* processed text contained four types of errors (Figure 8). Most of the errors fall under to the accuracy category. There were 23 accuracy errors, of which 22 errors were major and 1 minor error of accuracy. 7 errors of fluency were found in the translation (2 major and 5 minor errors); 3 terminology errors appeared (1 major and 2 minor errors); and 4 minor style errors were observed.

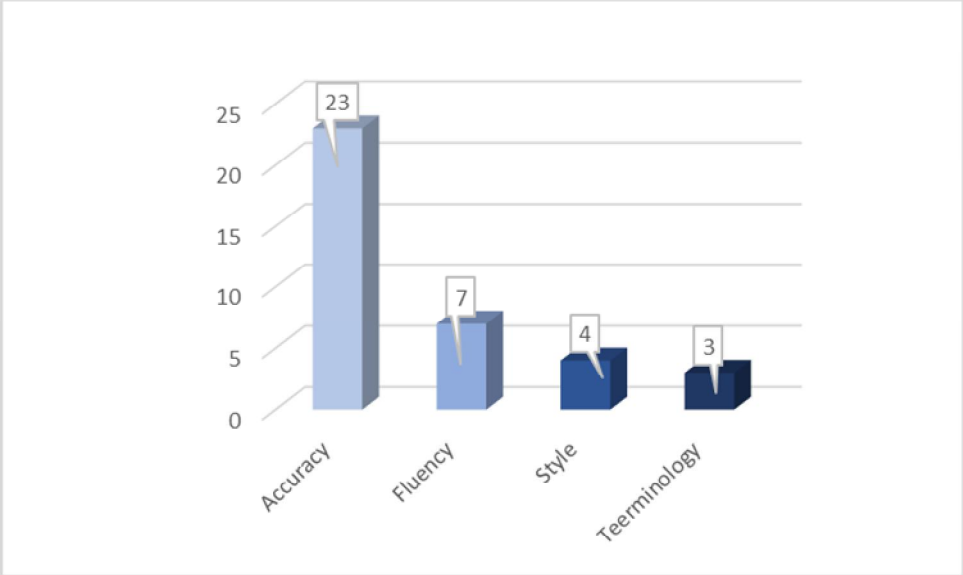

Figure 8. Google Translate machine translation output error categories with MQM metrics

Errors of accuracy were divided into over-translation error (1 error), under-translation errors (3 errors) and mistranslation errors (19 errors). An instance of mistranslation and poor translation can be seen in Example 2 (BLEU score 16.51):

Example 2

Source text: *Elsewhere in the world, the tweet still appears as normal.*
Google Translate output: *Kitur pasaulyje tviteris vis dar atrodo įprastas.*
Human translation: *Kitur pasaulyje įrašas vis dar matomas įprastai.*

The low score is not surprising as the machine translated variant barely renders the meaning of the original, but rather distorts it by extending the subject of the sentence from "įrašas" (en. *tweet*) to „tviteris" (en. *tweeter*).

Tilde has received the middle BLEU score (37.72) and MQM score (188), demonstrating its quality level lover than *DeepL* and higher than *Google Translate*. The total amount of errors in the selected text found through human evaluation following MQM metrics was 27. There were 18 major errors and 9 minor errors found (Figure 9).

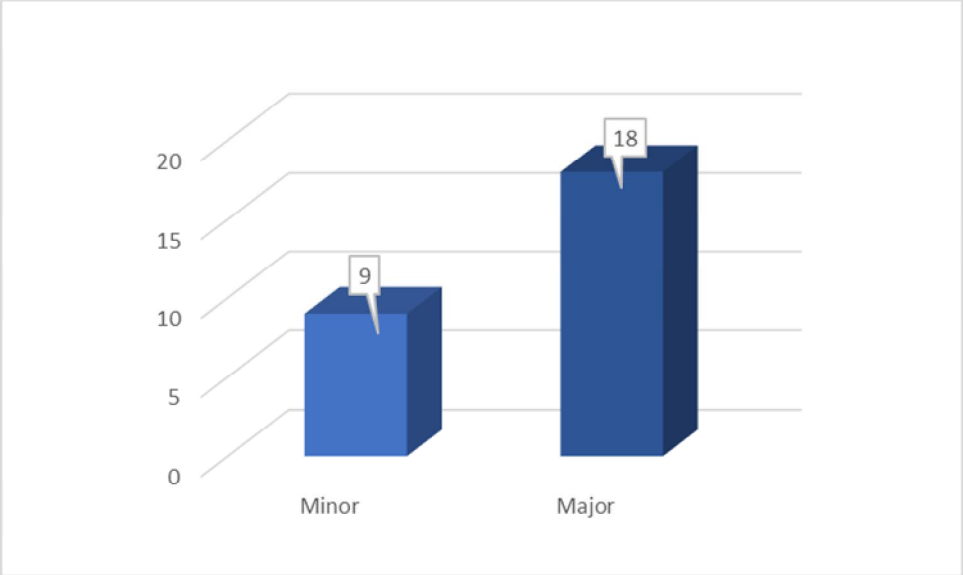

Figure 9. Tilde machine translation output error severity with MQM metrics

The Tilde machine translation output contained 15 accuracy errors (13 major and 2 minor errors), 8 style errors (2 major and 2 minor errors), 2 fluency errors (1 major and 1 minor errors), and 2 minor terminology errors (Figure 10).

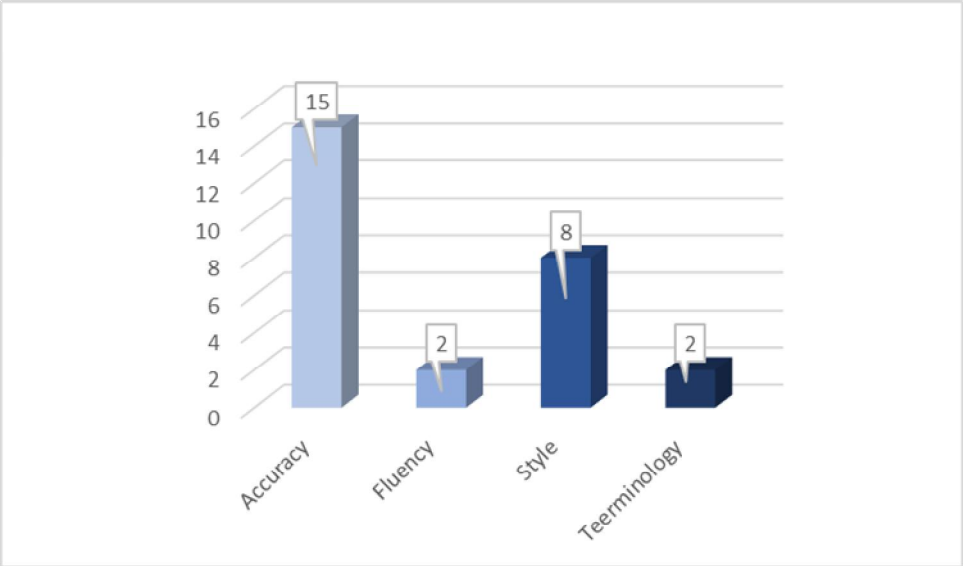

Figure 10. Tilde machine translation output error categories with MQM metrics

Accuracy errors were divided into 2 over-translation errors and 13 mistranslation errors. Example 3 (BLEU score 11.20) demonstrates an instance of mistranslation falling within the MQM accuracy error type:

Example 3

Source text: *We haven't yet resorted to these fines, but we will.*
Tilde output: *Mes dar nesiėmėme šių baudų, bet tai padarysime.*
Human translation: *Kol kas šių baudų nepritaikėme, bet pritaikysime.*

In Example 3, the English verb "resorted" is machine translated as "nesiėmėme", which is not completely unsuitable in this case, but rather vague and unclear, especially because the clarity of the meaning of the second part of the sentence "bet tai padarysime" (Eng. *but we will do*) depends on the translation of the first. The human reference translation variants "nepritaikėme" and "pritaikysime" are well chosen and express the intended meaning.

The human evaluation of the outputs of three machine translation engines demonstrated that *DeepL* offered the best quality. *Google Translate* showed the worse result, while *Tilde* was the second best. It is opposite to the results achieved using automatic machine translation output quality evaluation with the BLEU metrics.

6. Discussion.

Taking into account the survey findings on machine translation post-editing practices, we may draw attention to seemingly growing numbers of translators doing post-editing. Although the results are hard to compare due to different scopes and target groups of conducted research, the findings of this study are in line with those of previous research targeting multiple countries worldwide. A survey by Blagodarna (2018) has addressed 26 countries, including Lithuania, and found that the frequency of performing post-editing was varied between "never" to "very often" but a tendency was noticed for the answer "sometimes" to prevail. In our case, freelancers were more prone to do post-editing than language service providers (i.e., companies) on a day to day basis.

A study by Cadwell et al. (2018), which included analysis of the data obtained through focus group interviews with 90 translators of 24 source languages, reported that the majority of DGT translators in Brussels and Luxembourg used machine translation daily. However, a lot fewer translators of a language service provider based in the United Kingdom employed it routinely and almost half said they did that only a few times per year. Inconsistent terminology issues and poor quality for certain language pairs were found to be among the reasons for reluctance using machine translation, which is supported by our findings that low machine translation quality hinders translators from relying on it. Poor quality of machine translation to Lithuanian was also noted as a reason for not using machine translation by Lithuanian translators and translation companies (Levanaitė, 2021).

The common use of *DeepL* and *Google Translate*, as the most popular and/or best machine translation systems, has been proven in other studies conducted on the perceptions of various end-users (Kasperė et al., 2021; Vieira et al., 2021). The findings of our study are in line with previous research. *Tilde* as the third most common machine translation engine, revealed in our study, is a popular engine in Latvia, as well as in Lithuania, since the company is based in Latvia and has gained international visibility due to its constant innovation and focus on small and morphologically rich languages, stemming from the Baltic (Latvian and Lithuanian) language family perspective. For example, the first neural machine translation system for small languages was launched by Tilde right after it was launched for big languages by Google (Pinnis & Bergmanis, 2020).

Our findings are also consistent with those obtained in a plethora of studies that test whether automatic machine translation quality estimation and human / manual quality evaluation correlate. Although BLEU is considered the most reliable metric and in some studies its correlation with human evaluation scores seems to have been proven, our results corroborate the evidence of previous research where poor correlations between BLEU and human evaluation have been found. Therefore, despite the fact that BLEU metrics is kept closest to human evaluation of machine translation output quality, it should not be blindly trusted.

7. Conclusion.

This study presents a survey of language service providers and freelance language professionals to find out the prevalence of machine translation post-editing practices, and the most popular machine translation engines. The conclusions that may be drawn are the following.

Machine translation is relied upon by the majority of freelancers in Lithuania and, to a somewhat lesser degree, by language service providers, with the latter seemingly undecided whether the post-editing brings more benefits or harms. Three most popular machine translation engines used by professionals in Lithuania are neural machine translation engines *DeepL*, *Google Translate* and *Tilde*. Although some freelancers claim to use CAT-embedded or client-provided machine translation tools, which shows greater awareness as to the benefits and threats of machine translation within the professional community, further research is necessary to disclose the practices and perceptions towards the technology.

Altogether, an automatic and human/manual machine translation quality evaluation of a news text with three most commonly used machine translation engines shows contradicting results. Automatic estimation of machine translation output quality using the interactive BLEU score evaluator on a text processed with the three engines from English to Lithuanian shows that the best output quality is that of *Google Translate*, followed by *Tilde* and *DeepL*. On the other hand, the human / manual quality evaluation carried out using MQM metrics shows quite the opposite result, i.e., the highest quality is that of *DeepL*, followed by *Tilde* and *Google Translate*, the latter being almost twice as worse as that of *DeepL*. A conclusion might be drawn that the results of automatic estimation and human / manual evaluation of machine translation quality do not correlate. Therefore, reliance on automatic machine translation quality estimation should not be excessive or complete when making a decision whether machine translation post-editing is a cost-effective and time-consuming option for a given text type on the language service provider's part and sufficient quality on the customer's part. So far, the better option to evaluate the quality of machine translation output seems to be human evaluation even though it is expensive and time-consuming.

The limitation of this study is the scope, type and domain of the text chosen for the analysis. Future studies may be focused on replication of the analysis with different types and amounts of the source text and with different source languages in the pair with Lithuanian.

R e f e r e n c e s

Blagodarna, O. (2018). Insights into post-editors' profiles and post-editing practices. *Revista Tradumàtica. Tecnologies de la Traducció, 16*, 35–51. doi: 10.5565/rev/tradumatica.198

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens J., S. Castilho, F. Gaspari, S. Doherty (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, 9–38.

Chauhan, S., Daniel, P., Mishra, A., & Kumar, A. (2021). AdaBLEU: A Modified BLEU Score for Morphologically Rich Languages. *IETE Journal of Research*. doi: 10.1080/03772063.2021.1962745

Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. In Proceedings of Machine Translation Summit IX: Papers, New Orleans, USA. Retrieved October 15, 2022, from ACL Anthology website https://aclanthology.org/2003.mtsummit-papers.9/

Christensen, T. P., & Schjoldager, A. (2016). Computer-aided translation tools – the uptake and use by Danish translation service providers. *JosTrans – The Journal of Specialised Translation*, 25. Retrieved October 15, 2022, from the JosTrans website https://www.jostrans.org/issue25/art_christensen.php

Dey, S., Vinayakarao, V., Gupta, M., & Dechu, S. (2022). Evaluating commit message generation: to BLEU or not to BLEU? In Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER '22). Association for Computing Machinery, New York, NY, USA, 31–35. doi: 10.1145/3510455.3512790

EUATC, GALA, FIT, Elia, EMT & LIND-Web. 2018 Language Industry Survey. Retrieved April 3, 2022, from the website https://aticom.de/wp-content/uploads/2018/08/2018_Language_Industry_Survey_Report.pdf

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics, 9*, 1460–1474. doi: doi.org/10.1162/tacl_a_00437

ISO 18587:2017 (2017). Translation Services – post editing of machine translation output – Requirements.

Kasperė, R., Horbačauskienė, J., Motiejūnienė, J., Liubinienė, V., Patašienė, I., Patašius, M. (2021). Towards sustainable use of machine translation: usability and perceived quality from the end-user perspective. *Sustainability*, 13(23), art. no. 13430, 1–17. doi: 10.3390/ su132313430.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., & Menezes, A. (2021). To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. arXiv. doi: 10.48550/arXiv.2107.10821

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation, 25*, 131–148. *JosTrans – The Journal of Specialised Translation,*. Retrieved October 15, 2022, from the JosTrans website https://www.jostrans.org/issue25/art_koponen.php

Levanaitė, K. (2021). Lietuvos vertimo rinkos dalyvių požiūris į mašininį vertimą ir postredagavimą. *Vertimo studijos*, 14, 22–39.

Lommel, A.R., Burchardt, A., & Uszkoreit, H. (2013). Multidimensional quality metrics: A flexible system for assessing translation quality. In: *Proceedings of ASLIB: Translating and the Computer*, 35. Retrieved April 3, 2022, from the ACL Anthology https://aclanthology.org/2013.tc-1.6.pdf

Maučec, M. S., & Donaj, G. (2019). Machine Translation and the Evaluation of Its Quality. In A. Sadollah, & T. S. Sinha (Eds.), Recent Trends in Computational Intelligence. IntechOpen. doi: 10.5772/intechopen.89063

Mathur, N., Baldwin, T., & Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4984–4997, Online. Association for Computational Linguistics. Retrieved November 3, 2022, from the ACL Anthology website https://aclanthology.org/2020.inlg-1.45

Pinnis, M., & Bergmanis, T. (2020). Tilde's neural machine translation technology. In: Latvian Academy of Sciences Yearbook 2020, 83–87. Latvia: Latvian Academy of Sciences. ISBN 978-9934-549-94-6

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: a systematic review. *Language Resources & Evaluation*, 56, 593–619. doi: 10.1007/s10579-021-09537-5

Rossi, C., & Carré, A. (2022). How to choose a suitable neural machine translation solution: Evaluation of MT quality. In Dorothy Kenny (ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence, 51–79. Berlin: Language Science Press. doi: 10.5281/zenodo.6759978.

Seljan, S., Vicic, T., & Brkic, M. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2143–2148, Istanbul, Turkey. European Language Resources Association. doi: 10.13140/RG.2.1.4374.3204

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11), 1515−1532. doi: 10.1080/1369118X.2020.1776370

Way, A. (2018). Quality Expectations of Machine Translation. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. Machine Translation: Technologies and Applications, Vol. 1. Springer, Cham. doi: 10.1007/978-3-319-91241-7_8

Zaretskaya, A., Conceição, J., & Bane, F. (2020). Estimation vs Metrics: is QE Useful for MT Model Selection? In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 339–346, Lisboa, Portugal. European Association for Machine Translation. Retrieved November 3, 2022, from the ACL Anthology website https://aclanthology.org/2020.eamt-1.36

Бібліографічний опис:

Повілайтієне, М., Каспере, Р. (2022). Машинний переклад для постредагування текстів. *Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9. Сучасні тенденції розвитку мов*, 24, 47–62. https://doi.org/10.31392/NPU-nc.series9.2022.24.04.

Анотація

У багатьох дослідженнях із якості перекладу було доведено, що результати машинного перекладу в деяких мовних парах все ще далекі від досконалості (Копонен, 2016). Подальше редагування машинного перекладу, тобто постредагування, стало звичайною практикою серед перекладачів та постачальників перекладацьких послуг, особливо з тими мовними парами, де машинний переклад демонструє відповідність із точки зору семантики не достатньою мірою. Швидкий розвиток нейронного машинного перекладу та підвищення його якості призвели до зростання попиту на редакторів текстів, попередньо перекладених машиною.

У цьому дослідженні зроблено спробу оцінити якість найбільш популярних інструментів машинного перекладу литовською мовою з метою пошуку кореляції як між результатами оцінки якості машинного перекладу, виконаного автоматичним способом (із алгоритмом BLEU score), ручним способом (людиною) із застосуванням багатовимірних показників якості (MQM), так і найбільш поширеними програмами машинного перекладу, які використовують фрілансери та інші постачальники перекладацьких послуг.

Висновки ґрунтуються на результатах опитування та аналізу якості перекладу, виконаного автоматичним способом та ручним способом (людиною). Отримані дані демонструють і підтверджують попередні дослідження, що оцінку якості машинного перекладу, виконаного автоматичним способом, не можна сприймати як належне. Оцінка якості машинного перекладу, відредагованого людиною, як і раніше, є найкращим індикатором того, чи відповідає такий інструмент машинного перекладу меті перекладу.

Дослідження висуває на передній план деякі важливі узагальнення, які можуть бути корисні з дидактичного погляду тим, хто навчає перекладачів і редакторів машинного перекладу, а також цікаві для перекладачів-практиків загалом.

Ключові слова: *машинний переклад, постредагування, алгоритм оцінки якості машинного перекладу (BLEU score), багатовимірні показники якості (MQM).*