SKASE Journal of Translation and Interpretation, 2023; 16(2): 5–22
doi: 10.33542/JTI2023-2-2

Medical terminology issues: a feasibility study of machine translation in a low-resource language

Ramunė Kasperė, Jurgita Mikelionienė, Dalia Venckienė

Abstract
Medical knowledge may be targeted at a multitude of audiences, including researchers, health professionals, patients and the general public. For many reasons, like in other spheres, machine translation has gained its way into medical and clinical settings. Medical translation is complicated not only because of terminological issues, but also because of a variety of genres and text types ranging from clinical practice to education, research and dissemination. Adequacy and fluency of the translated text are all equally crucial. The aim of this study is to analyze machine translation output of medical texts with a focus on terminology issues in English to Lithuanian language pair. The results of this study on human evaluation of machine translated output reveal that the performance of machine translation systems is insufficient, and the raw output requires careful review. While the accuracy of terminology translation and adequacy of longer segments may be regarded fairly acceptable for non-professional uses, the fluency of the raw translation mostly produces poor readability. Our research corroborates the findings of other studies that machine translation may be employed as a complementary tool only and cannot be relied on as the only source, but its significance for medical professionals and the broad public cannot be ignored.

Keywords: *machine translation, terminology, accuracy, adequacy, fluency*

1. Introduction

Management and standardization of medical terms, as well as digitization of terminological resources, have received special attention among researchers. Translation of medical terms has also been investigated as the demand for translation of medical texts is huge: patients planning to consult doctors or undergo treatment in foreign countries, or those willing to read documents written by foreign doctors in a native language need translations. The texts requiring translation are often written in a hurry, may contain errors, multiple abbreviations, and medical jargon. Translation service providers providing medical translation emphasize that medical texts are specific and complex; nevertheless, high quality translation is expected. Therefore, descriptions of translation service providers' activities include quality requirements and touch upon various aspects of medical translation.

The communicative aspect is vital to production and translation of medical texts. Communication may take place between professionals (this includes the use of language in conferences, reports, etc.); semi-professional communication is established between doctors and patients; non-professional communication takes place when discussing health problems on an everyday basis. Doctors and patients possess different levels of medical knowledge, may understand concepts and terms encoding the concepts differently, and this may complicate communication between the two parties. Given the professionalism of the participants of the communicative act, terms may be categorized into scientific denotations of concepts used by medical professionals, and general medical terms used by the doctor to convey the message to

the patient. Medical professional jargon, which occurs in informal professional language, may also be used in various communicative situations. Fage-Butler & Nisbeth Jensen (2016: 644) point to the existence of improper use of medical terms and medical jargon and note that problems arise when patients do not understand such lexis.

An individual facing his or her own diagnosis or a diagnosis of a close person may desperately search for all readily available information, including information in foreign languages. In pursuit of the answers, trying to save time and financial resources, lacking bilingual dictionaries and encyclopedias, as well as the knowledge about sophisticated search techniques, such individuals often resort to the use of machine translation engines, by copying the text into machine translation systems. The quality of machine translation in low-resourced languages may be substantially reduced in comparison with bigger languages that receive huge investments into data acquisition and machine learning. It becomes crucial to consider the reliability of machine translation engines and to be able to assess the quality of the provided service.

Abundant use of specialized terminology and, sometimes, the use of different denotations to name the same concept (e.g. en. *carcinoma* and *cancer*, lith. *tuberkuliozė* and *džiova* (lot. *Tuberculosis*) are among key characteristics of popular science texts. Popular science, also referred to as "popularization of science", "pop science", "expository science", "science communication", "public science", "public understanding of science" (Leane 2007; Manfredi 2019: 64), is ained at ordinary people as opposed to the field specialists; therefore, it may be defined as a specialised or technical subject explained and described "in a generally intelligible or appealing form" (Manfredi 2019: 64).

In Jacobi's (1987, cf. Ciapuscio 2003: 209) view, first, science uses "hermetic and distinct" language that needs some "decoding", and second, that "a mediator is necessary to "translate" the scientific language into everyday language" (Manfredi 2019: 64). Calsamiglia & van Dijk (2004: 370) view "popularization" as various genres that transform specialized knowledge into "lay" knowledge, in which scientific discourse is recontextualized. In Tinker Perrault's (2013: xiii) opinion, "science popularization" refers to "science-related communication directed at nonspecialist audiences".

Manfredi points out that "[i]n the framework of linguistics", most existing research explores popular science "within the broader field of scientific discourse" (2019: 66). Popular science can be viewed as *science journalism*, as it shares features of both science and journalism and its writing is the result of "a relationship" between science and the media (Nelkin 1987 in Manfredi 2019: 66). News articles in consumer and specialized magazines are viewed as examples of "science journalism" (Myers 1990: 187; Manfredi 2019: 66). Medical writers are expected to deliver medical news to lay readers as people have a general interest in science focusing on health-related issues; therefore, the press continues to be a source of information and learning for many people (Nelkin, 1987), and the role of journalists as specialists spreading precise and "balanced" public health information is crucial (Chase 2006: 162). For this reason, popular science articles falling within the field of Medical and Health Sciences of the OECD's *Revised Field of Science and Technology Classification* (2007) are taken as the focus of analysis.

Therefore, it is essential to explore if principles of terminology use and management are adhered to in the process of machine translation and whether machine-generated terms and output satisfy the requirements set for the quality of translated medical texts. This study aims at an analysis of popular science texts as units designed for semi-professional communication, focusing on Lithuanian machine-translated equivalents of English medical terms, and provides

inference about the quality of machine-generated output in terms of adequacy and fluency. The findings of this study contribute to other research on machine translation with low-resourced languages, but offer new insights into the issues of terminology.

2. Literature overview

2.1.Machine translation in Lithuanian as a low-resource language

Machine translation in low-resourced languages receives less attention from developers but is an area of interest of researchers. The scientific literature discussing machine translation quality criteria and the outcome in Lithuanian is scarce and mostly focused on the English-Lithuanian language pair translation (Daudaravičius 2006; Rimkutė & Kovalevskaitė 2007a, 2007b, and 2008; Petkevičiūtė & Tamulynas 2011; Cvilikaitė 2008; Daubarienė & Ziezytė 2013; Stankevičiūtė et al. 2017; Kasperavičienė et al. 2020). Cvilikaitė (2008) performed an analysis of lexical errors identified in the output of statistics-based and rule-based machine translation systems. The research revealed that the errors produced by the machine translation system were closely related to cases of polysemy, homonymy, the use of multi-word lexical units, and lack of equivalents in the target language (Cvilikaitė 2008: 35).

Petkevičiūtė & Tamulynas (2011) performed a thorough analysis of taxonomies proposed by other researchers for classifying machine translation errors and developed a list of errors characteristic of the machine translation in English to Lithuanian language pair. In this study, the errors in Lithuanian are categorized into two groups: linguistic (morphological and lexical) and systemic (errors in dictionaries and program codes, for which there is no linguistic, and sometimes no reasoned logical, explanation). The research also revealed that morphological errors (related to expression of gender, basic verb forms, grammatical number, and parts of speech) and lexical errors (i.e. relating to untranslated words or polysemy) occurred most frequently in the Lithuanian output of statistics-based and rule-based machine translation systems (Petkevičiūtė & Tamulynas 2011: 39–41).

Significant improvements in the quality of machine translation occurred around 2015 with the introduction of neural machine translation technology. The Latvian company Tilde was the first to employ the neural machine translation technology for the purpose of translating into low-resourced, morphologically rich languages, following Google's introduction of this technology for the purpose of translating between well-resourced languages (Pinnis & Bergmanis 2020). A number of research works specifically dedicated to the machine translation quality of under-resourced languages, such as Spanish-Galician (do Campo Bayón & Sánchez-Gijón 2019), Azerbaijani, Belarusian, Galician and Slovak (Xia et al. 2019), etc. have been published. Stankevičiūtė et al. (2017), Kasperavičienė et al. (2020), Miltakienė (2021), Kvyklytė & Mikelionienė (2022) have published the results of quality evaluation in the English to Lithuanian machine translation output generated by neural machine translation systems and have arrived at the conclusion that achieving satisfactory quality in low-resourced languages is a challenging task, even for neural machine translation systems.

As far as research on various aspects of machine translation quality in medical texts is concerned, studies have already been carried out, e.g. (Skianis et al. 2020; Zappatore & Ruggieri 2024; Mehandru et al. 2022). However, there is a lack of specialized research focusing on evaluation of machine-translated scientific and popular science text in terms of adequacy,

intelligibility, readability, and other criteria, and studies of such aspects of medical translation in the English to Lithuanian language direction have not yet been conducted.

Translators and society tend to increasingly rely on machine translation. Nevertheless, several studies show that professional translators employed by translation service providers and various other institutions have mixed views towards the integration of machine translation in daily tasks of translation. Some favour the technology, while others are reluctant to extensively use machine translation systems for a few obvious reasons (Cadwell et al. 2018; Levanaitė 2021; Povilaitienė & Kasperė 2022). The unsatisfactory quality of machine-translated output, especially noticeable in specific combinations of morphologically rich or smaller languages, is repeatedly indicated as a major reason. Translators sometimes note inconsistent use of terminology in machine-translated material, which must be carefully revised and thoroughly checked (Cadwell et al. 2018: 312).

2.2.Machine translation quality assessment

Multidimensional quality metrics (MQM) is a framework for translation quality evaluation that is often used to assess human translation and machine translation (MQM Council n.d.). MQM may be used to evaluate the quality of translation products in all languages, and has been recently used in studies focusing on the quality of machine translation. Following Rivera-Trigueros, MQM is reported to be employed in approximately one-third of the analyzed studies (2022: 609). MQM's hierarchy of error types lists more than 100 types. This universal metrics is designed to support the analysis of errors in various content types, for instance, human translation, machine translation, post-edited content, etc. The MQM's applicability "goes far beyond quality evaluation of translation products", and MQM is implemented "in the language industry, in institutional translation, in translation tools, and in research projects" (MQM Council n.d.). Multidimensional Quality Metrics was originally proposed by Lommel et al. in 2014. The original 2014 version of the metrics was comprised of 10 major categories that further, hierarchically, branched into more subcategories (Lommel et al. 2014: 458). The approach is based on international standards ISO DIS 5060:2022, *Translation services – Evaluation of translation output – General guidance* and ASTM WK46396: *Standard Practice for Analytic Translation Quality Evaluation* (MQM Council, n.d.). In this evaluation system, the main error typology is comprised of the following fundamental error types:

- *Terminology errors*, for example, the term *pool* in a text on a game resembling billiards is rendered as *baseinas* (= pool of water);
- *Accurac*y errors, including those of addition, mistranslation, omission, untranslated text;
- *Errors of linguistic conventions* (*fluency* in the previous version of MQM), including grammar, register, consistency, spelling, typos, unintelligible text;
- *Design and markup errors*, related to visual content, for example, text formatting;
- *Errors of locale conventions*, referring to compliance of the translated content to formal local requirements, for example, proper date format;
- *Style errors*, for example, when a text meant for children is translated in a way that is too complicated for them to comprehend;
- *Audience appropriateness errors* (*verity* in the previous version of MQM) related to the translation output inappropriate for the target locale or target audience.

Terminology errors are further classified into three types (MQM Error Typology, n.d.):

- *inconsistent with terminology resource(s)*, i.e. "use of a term that differs from term usage required by a specified termbase or other resource" (hereinafter also MQM1), e.g. a terminology resource provides the term *USB memory stick*, but instead of it, *USB flash drive* is used.
- *inconsistent use of terminology*, i.e. "use of multiple terms for the same concept in cases where consistency is desirable" (hereinafter also MQM2), e.g. the terms *brake release lever* and *brake disengagement lever* are used interchangeably in the same translation text, even if they both refer to the same concept.
- *wrong term*, i.e. "use of term that is not the term a domain expert would use or because it gives rise to a conceptual mismatch" (hereinafter also MQM3), e.g. the English *river* is translated into French as *rivière*, when *fleuve* is the right translation.

Besides, error severity levels may be different, i.e. errors may have different impact on the quality of the target text, also affecting the reader's perception of the text. Following this approach, severity levels fall into four categories, supported by MQM (Lommel 2018). Due to critical errors, the text becomes misleading and unusable (does not fulfil its function). Even one such critical error may inhibit comprehension of the text. Major errors have a negative impact on the meaning of the text and hinder comprehension. Although the impact of such errors is not critical, they nevertheless are noticeable and may cause the reader's irritability and dissatisfaction with the text. Minor errors do not hinder comprehension. Many a time the reader will mentally correct such errors and continue reading or will even not notice them. Instances that are classified as null in the MQM taxonomy are not errors in fact, but they may be an outcome of changes made post submission of translation (Lommel 2018).

In modern technologized environments, the border between human and machine translation becomes blurred (Castilho et al. 2018: 3; Doherty 2016: 953). The concept of translation quality is "difficult to operationalise and measure", and in the theoretical discussion of translation quality, the dichotomy between accuracy or adequacy (the source-oriented concepts) and fluency (the target-oriented concept) is highlighted (Castilho et al. 2018: 1). Translation quality assessment (TQA) processes vary and have limitations, and various measures exist for TQA to be performed by humans, for both industry and research purposes. Fluency and/or adequacy as measures in machine translation have been used in studies by Fernández-Torné & Matamala (2021), Doherty et al. (2013), Castilho et al. (2018), Popović (2020), to mention but a few.

Of note, Doherty et al.'s (2013: 11) study focusing on survey of translation and localization buyers and vendors revealed that "the most popular choice in evaluating MT quality is human evaluation (69%)". When TQA is carried out by humans, according to Castilho et al. (2018), evaluators most commonly look at adequacy and fluency as primary measures, while readability, comprehensibility, usability, and acceptability of target texts (MT output especially) can also be assessed (i.e. secondary measures may be taken).

Castilho et al. (2018: 18) define adequacy as "the extent to which the translation transfers the meaning of the source-language unit into the target" and note that authors of certain studies use the terms accuracy or fidelity to refer to adequacy. Fluency, which is also referred to using the term intelligibility by Arnold et al. (1994), is defined as "the extent to which the translation follows the rules and norms of the target-language", regardless of the input text (Castilho et al. 2018: 18). Capitalizing on the findings of Arnold et al. (1994) and Reeder (2004), Castilho et al. (2018) note that fluency may be reduced by untranslated words, mistranslations, grammatical errors, and also incorrect pronouns, prepositions, and

punctuation. Adequacy and fluency are customarily ranked using Likert scales. To evaluate adequacy (accuracy), for example, the following scales may be used: 1 – completely inaccurate, 2 – mostly inaccurate, 3 – somewhat accurate, 4 – mostly accurate, 5 – completely accurate. Evaluation is carried out at a sentence or segment level; when MT output is evaluated, extended context is not considered. Proficiency in the target language is required to evaluate fluency, while bilingual proficiency is required to evaluate adequacy (Castilho et al. 2018).

For research purposes, renderings of the same source text from various machine translation systems are evaluated for comparison. Castilho et al. (2018) state that both amateur and professional evaluators can be involved in quality assessment: though results provided by professionals may be considered more trustworthy, amateur raters may be also helpful and are being increasingly involved in machine translation research projects implementing crowdsourcing techniques as they are more accessible than professional evaluators. In case of group-based translation quality assessment, several raters assess human or machine translation using a set of criteria and then average their scores, to reduce personal biases (Castilho et al. 2018: 15), which was the method employed in this current small-scale research. Castilho et al. (2018: 15) draw their readers' attention to Doherty's (2017) finding that information on human TQA tasks is rather scarce and that available data leads to the assumption that professional or trained evaluators are "the exception" in machine translation evaluation, which points up, to some extent, the uniqueness of the current research. TQA is generally carried out with the aim of minimizing risk, "whether this is a risk to communication, to reputation, or a risk of injury or death", and acceptability of a translation implies "a permitted level of acceptable risk" (Castilho et al. 2018: 23). This study was designed to raise awareness of the general public, translators, and (post)editors about the current quality of machine translation of medical texts from English into Lithuanian, in terms of adequacy and fluency, and being designed so, it also yields some insights into whether reliance on unedited machine translated output of medical content may, to some extent, be risky.

2.3.Peculiarities of terminology translation

Structurally, terms can be classified as simple and compound or single-word and multi-word terms (two-word, three-word, etc.) terms (Kvašytė 2005: 71). Studies on the problems of translating terms point out that multi-word terms, being the most frequent type, are the main way how "concepts are linguistically expressed in specialized domains" (Cabezas-García & Faber Benitez 2017: 193). Since they are complex and not systematically represented in terminological dictionaries and resources, terms are more difficult to render in another language, especially because of a variety of options to select from in the translation process (León-Araúz et al. 2020: 2358).

Finding the correct terminological equivalent in the target language is considered to be one of the most complex tasks in medical translation (Rask 2008: 17). Among the reasons of insufficient or inappropriate rendering of terms in the translation process, Moghadam & Far (2015) mention the emergence of terms (Moghadam & Far 2015). Fóris (2022: 55) notices that terminology should be carefully considered prior to translation: the source text should include proper and precise terms as the quality of the target text depends on the quality of the source text submitted for translation. To follow Cabré (1996), the search of specific terms should be performed in approved dictionaries or other reliable sources in both source and target languages as this enables the translator to avoid undesirable variability of terminology in the translated text. In other words, it is necessary to consider text-normative equivalence (Munday 2004).

Terminology also poses issues for machine translation systems. Multi-words terms as any multiword units "significantly contribute to the robustness of MT systems as they reduce the inevitable ambiguity inherent in word to word matching" (Váradi 2006: 73). Appropriate rendering of terms is critically important in machine translation (Haque et al. 2020: 149). However, research on machine translation of terms in various languages and language directions has provided evidence that neural machine translation systems have been in fact underperforming in rendering terms, compared to phrase-based statistical machine translation (Haque et al. 2019).

Medical terms are vitally important in healthcare as patients need to understand treatments and diagnoses communicated to them by medical staff (Tarasiewicz 2023). Therefore, the communicative aspect needs to be assured. Among the theoretical approaches of contemporary translation-oriented terminology, *the Communicative Theory of Terminology* suggests that terms can be simultaneously researched as linguistic units, cognitive units, and units with the social function; according to this theory, the primary purpose of translation and use of terminology is to convey necessary information and meet communicative needs (Cabré 1999: 45–48; Cabré 2009). Cabré (1999), a founder of this theory, emphasizes the importance of the context in which terminology is used. Translators of medical texts are expected to render synonyms, numerous abbreviations and acronyms, and eponymous terms. Sometimes it may be difficult to recognize terms. Context plays an important role in term recognition, and any lexical unit can become a term in a specific context (Cabré 2003: 190). It is because of its emphasis on the importance of context that this theory of terminology serves as the theoretical backround in a translation-oriented terminology analysis.

3. Methods and materials

For the purposes of this study, a medical term is a unit of meaning consisting of one or more elements, denoting a specific medical concept and having a definition. The most common semantic categories of words in the language of medical professionals were also used to identify medical terms, as proposed by medical terminology researchers. According to Černý (2008: 41), these include names of diseases and disorders, medicines, medical equipment, procedures and treatment methods.

Randomly selected popular science texts from the New Scientist web portal on different medical topics (consequences of coronavirus, gout, booster vaccine, neonatal umbilical cord clipping, mold disease, Alzheimer's disease) were taken for analysis. The criteria for selection of articles were relatively recent date of publication (the analyzed articles were published in the period July 16, 2022 – December 16, 2022); topics that do not overlap; similar approximate length (23–31 segments each). The texts were machine translated from English to Lithuanian by *Google Translate* and *Tilde Translator* systems, both general purpose neural machine translation systems. To avoid subjectivity in identifying the terms, the English units selected from the analysed texts as terms were checked manually in terminology resources (the IATE, Eurotermbank, Merriam-Webster Medical Dictionary, SNOMED CT Dictionary of Medical Terms, and Farlex Medical Dictionary) by three terminologists and language specialists in order to verify their status. The analysis then was twofold. First, the translated terms were evaluated according to the MQM taxonomy, classified into three subcategories within the category of Terminology errors, and the severity of the errors found was determined. In the selected corpus, the analysis was conducted with a focus on both the unique terms and their

repeated instances. The selected unique terms (n = 149) were analyzed in terms of two MQM subcategories, i.e. the use of terminology inconsistent with resources (MQM1) and wrong term (MQM3). Unrecognized and therefore untranslated terms were also attributed to the MQM3 category, with the exclusion of untranslated Latin terms, the use of which is a common feature in medical content. All terms, i.e. all instances of unique terms, (n = 308) were analyzed within the selected corpus in terms of the subcategory of the inconsistent use of terminology (MQM2).

Then the translated segments (sentences) containing a term were manually evaluated for adequacy (lexical/terminological and grammatical inaccuracies) and fluency (syntax and style errors) by three experts: two translators and a terminologist. In total, there were 154 segments, which were ranked on the Likert scale from 1 to 5 for adequacy where 1 – completely inadequate; 2 – mostly inadequate; 3 – somewhat adequate; 4 – mostly adequate; 5 – completely adequate; and for fluency where completely non-fluent; 2 – mostly non-fluent; 3 – somewhat fluent; 4 – mostly fluent; 5 – completely fluent. The three raters assessed the adequacy and fluency of the segments independently.

4. Results

4.1.Quality evaluation of machine translated medical terms

The corpus of the terms that were evaluated following the MQM contained 48% of single word terms (n = 72), 34% of two-word terms (n = 51) and 18% of multiword terms (n = 26).

The performed detailed analysis of machine-translated terms indicates that a major part of the terms (81%) were translated appropriately (77% from Tilde Translator and 85% from Google Translate) (see Figure 1).

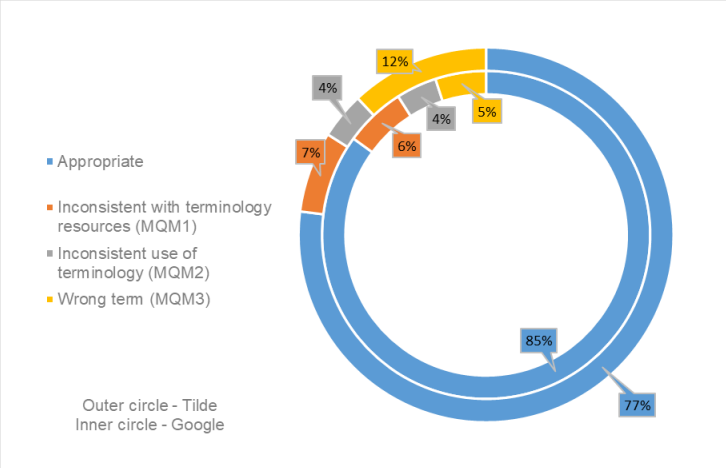

Figure 1 Quality ratings of machine-translated terms

The majority of single-word terms – international (derived from classical languages) or English – were rendered into Lithuanian appropriately, by selecting optimal single-word international or Lithuanian equivalents. Some of the examples of properly rendered terms are as follows: concepts of human body structure: *placenta – placenta, lungs – plaučiai, blood – kraujas, kidneys – inkstai, antibody – antikūnas, proteins – baltymai*; procedures and treatment

methods: *chemotherapy – chemoterapija, resuscitation – gaivinimas, freezing – užšaldymas*; symptoms, illnesses, conditions: *gestation – nėštumas*, *breathlessness – dusulys*, *death – mirtis*; denominations of persons: *doctor – gydytojas, infant – kūdikis*, etc.

The majority of two-word terms were also rendered into Lithuanian appropriately, by selecting proper two-word terms comprised of a major term component (noun) with a coordinated attribute (e.g. *immune response – imuninis atsakas*, *genetic material – genetinė medžiaga, stem cells – kamieninės ląstelės*) or a major term component (noun) with an uncoordinated attribute (*blood volume – kraujo tūris, cellular immunotherapy – ląstelių imunoterapija*).

The quality evaluation of multiword terms was subject to close scrutiny of both single-word terms and two-word terms, depending on which component the basis of a multiword term is formed. Thorough analysis of 26 machine-translated multiword terms showed that machine-translation systems erred more often transferring such multiword items. Nevertheless, it should be noted that the engines handled translation of eponymous terms well, including the term *chimeric antigen receptors – chimeriniai antigeno receptoriai* with an eponym as its component.

According to the MQM, the category of **the use of terminology inconsistent with resources (MQM1)** includes cases when the use of a term differs from term usage required by a specified termbase or other resource. If an English term rendered into Lithuanian resulted in a different term than recommended in terminological resources, it was considered as inconsistent with resources irrespective of the fact whether it was fully rendered inappropriately or whether only part of it was inappropriate. About 7% of terminology errors were assigned to category MQM1: 7.1% from *Tilde Translator* and 6.2% from *Google Translate*. Erroneous equivalents still carried some meaning and belonged to the medical or health domain. Specific errors are typical, and a machine translation system may have stuck to a more frequent usage, though this usage was not a standard equivalent of the corresponding term in the source text, e.g.: *Antifungal drugs* were rendered by the two systems as *priešgrybeliniai vaistai* instead of *vaistai nuo grybelio*, which is recommended by the State Lithuanian Language Commission. Multiword terms were not always treated as a whole; therefore, components of specific multiword items were rendered as separate words (instead of rendering those English multiword items into medical terms proper), e.g. *sperm cells – spermos ląstelės* instead of *spermatozoidai* [= *spermatozoons*]. As an additional example, the Latin, anatomical term *Loccus coeruleus* was left untranslated and, thus, also assigned to the MQM1 category, adhering to the principle that as much of the information as possible should be translated into the target language in popular science texts.

According to the MQM typology, the category of **inconsistent use of terminology (MQM2)** comprises errors arising when, in translation, "multiple terms are used for the same concept in cases where consistency is desirable" (MQM Council n.d.). About 4% of terminology errors were assigned to category MQM2: 4.9% from *Tilde Translator* and 3.9% from *Google Translate*. Such errors were relatively infrequent, and this may be ascribed to a short length of the content selected for analysis.

For instance, *non-vigorous infants* was rendered in two ways: as *neenergingi* [= lacking vitality] and *nestiprūs* [= limp] (instead of the proper equivalent *neaktyvūs kūdikiai* [= inactive infants], which would be the term conforming to organizational terminology standards, used by neonatal care specialists.

In addition, translators and (post)editors of English popular science texts should note variability of the terms used in the source text to denote a specific concept and should ensure

consistent use of respective equivalents in the target texts. However, in our analyzed corpus, it was not always the case. For example, three words – *infant* (a higher register word, may acquire the meaning of a legal term *minor* in specific contexts), *baby* (a less formal word, may be used as slang in specific contexts), and *newborn* were used to refer to a recently born child in the text on neonatal care. Another example of term variability in the source text refers to two terms, *brain injury* and *brain damage*, which were used to discuss the effects of ischemic encephalopathy (a brain injury caused by a lack of oxygen, moderate-to-severe). In the source text, *brain injury* marked *ischaemic encephalopathy*, while *brain damage* was used in the concluding remark to discuss the effects of the condition in a broader sense. Both machine translation systems did not produce errors in translation: Lithuanian versions of the two aforementioned English terms were correct and multiple equivalents (*smegenų pažeidimas* – appropriate Lithuanian use for the source concept) were applied consistently.

About 8% of terms were assigned to the category of **wrong terms (MQM3)**: 11.4% from *Tilde Translator* and 5.2% from *Google Translate*. These errors are the most conspicuous, occur when the term a domain expert would not use is used, and, according to the MQM Typology, give rise to conceptual mismatches. The following examples illustrate the usage of terms in the machine-translated output that are not related to the medical or health fields: *early umbilical cord clamping – ankstyvasis laido* [= *cord*, technical field] *suspaudimas* instead of *ankstyvasis virkštelės perspaudimas*; *... flu shots – ... gripo nuotraukos* [= *photos*, general domain]; *booster – stiprintuvas* [= *booster*, technical field] instead of *stiprinamoji vakcina.*

Sometimes the system(s) failed to recognize abbreviations used in the source text. For instance, the abbreviation *CAR* (used in isolation) or in the term *CAR-T cells* was rendered into *automobiliai* [= *automobiles*] instead of, for instance, original allowed *CAR*. In specific cases, the system(s) failed to recognize and render terminology, for instance, the term *early umbilical cord clamping* was rendered into *ankstyvas virkštelės CLAMP* (treated as an abbreviation).

It was noted that errors assigned to category MQM3 (and sometimes those assigned to category MQM1) can be regarded as severe if not critical. For example, the term *strain of the virus* was rendered by *Google Translate* into substandard borrowing *viruso štamas,* which is not recommended by the State Lithuanian Language Commission, and instead *padermė* is to be used.

Minor and insignificant errors do not prevent the reader from understanding the term or the general meaning of the text; nevertheless, such errors must also be eliminated. Examples of such minor errors in the target texts include improper use of cases, Latin terms not provided in italic, omitted letters, etc.

Based on the above, the question arises to what extent this is influenced by the fact that terminology is translated in sentences from popular science texts (rather than scientific ones). It is evident from specific examples that a specialized medical or health-related context does not always help machine translation tools to select adequate equivalents. For example, *delivery [of infants]* was rendered by *Tilde Translator* into *pristatymas* [= *delivery of a parcel*]. It may be hypothesized that the system was misled by abstract nature of the term *delivery* and produced a critical error, instead of selecting medical equivalents of the term, *gimimas* or *gimdymas*.

In addition to being filled with specialized terminology, popular science texts, especially English, include units of figurative language and professional jargon. Lithuanian popular science texts, however, are more neutral (Ringailienė 2014: 141). When rendering such figurative units or units of professional jargon into Lithuanian, machine translation systems are not always capable of selecting precise and stylistically appropriate Lithuanian equivalents,

and the context is not always helpful. Some cases in point are provided further. For example, the word "jabs" was rendered by *Tilde Translator*, possibly misled by the beginning of the word, into *žandikaulis* [= *jaw*], while *Google Translate* rendered the word into *dūris* [colloq. for *injection*], which would be an adequate equivalent for the informal British word (*jab: chiefly British, informal*), though less suitable for a more neutral Lithuanian style. In the text on neonatal care, two terms (a more stylistically neutral *newborn* and a stylistically marked *newborn baby*) were used, in parallel; both cases were rendered into stylistically neutral *naujagimis* [= *newborn*].

4.2.Adequacy and fluency of machine translated medical texts

Since the majority of the terms were translated accurately in many cases in the segments, the ranking analysis conducted by three independently working raters was aimed at checking the context surrounding the terms in terms of its adequacy and fluency. Figure 2 and Figure 3 demonstrate the distribution of the scores from 1 to 5 given by three raters to the segments machine translated from English to Lithuanian in terms of two criteria, i.e. adequacy and fluency. Approximately one-third of the segments (33%) from *Tilde Translator* and approximately two-fifths of the segments from *Google Translate* (43%) were ranked as completely adequate; meanwhile, approximately one-third of the segments from both machine translation systems were ranked as mostly adequate (32% from *Tilde Translator* and 35% from *Google Translate*). What concerns inadequate segments, there were 5% of those from *Tilde Translator* and only 1% from *Google Translate*. There were 9% of mostly inadequate segments from *Tilde Translator* and 4% from *Google Translate*. Having in mind that 77% of terms were translated adequately by *Tilde Translator* and 85% were adequately translated by *Google Translate*, the findings of the entire segment may imply that rendering context adequately is more problematic than rendering the term itself.

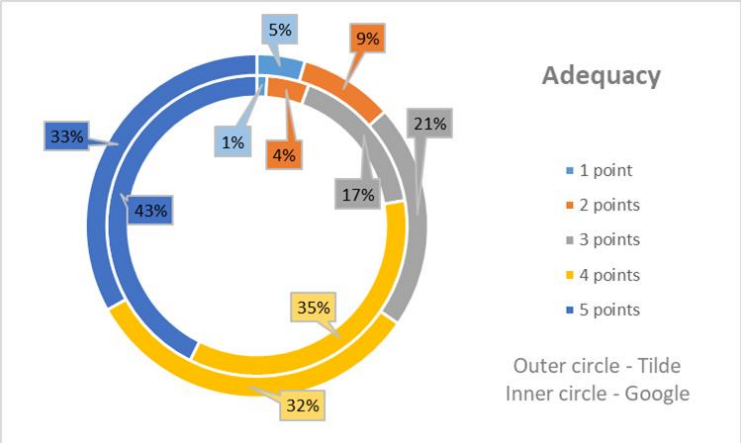

Figure 2 Distribution of scores (in percentages) for adequacy in Google and Tilde machine translation systems

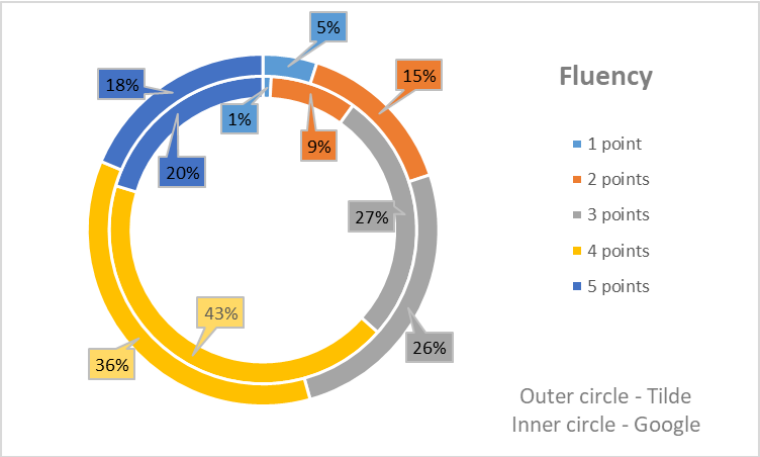

Figure 3 Distribution of scores (in percentages) for fluency in Google and Tilde machine translation systems

From the data in Figure 3, it is apparent that the scores given by the three raters for the fluency criterion to all translated segments containing terms are somewhat divergent from the scores given for adequacy. There were fewer completely fluent segments from both *Tilde Translator* (18%) and *Google Translate* (20%) compared to adequacy scores (33% from *Tilde Translator* and 43% from *Google Translate*, see Figure 2). In terms of mostly fluent segments, there were 36% from *Tilde Translator* and 43% from *Google Translate*. The raters rated 5% of the segments from *Tilde Translator* as completely non-fluent, 15% as mostly non-fluent, and 21% as somewhat fluent; in the same line, 1% of the segments from *Google Translate* were ranked as completely non-fluent, 9% as mostly non-fluent, and 26% as somewhat fluent.

The average adequacy score of the three raters of all segments processed with *Google Translate* was 4.13, while the average score of all segments processed with *Tilde Translator* was 3.80. The average fluency score of the three raters was 3.78 for all segments processed with *Google Translate* and 3.55 for all segments processed with *Tilde Translator*.

One of the insights that can be inferred from these data is that *Google Translate* performs better than *Tilde Translator*, which might have been expected from the beginning. However, taken together, the results indicate that segments processed using any of the two machine translation systems are considered less fluent in comparison with adequacy.

5. Discussion

Cabré's Communicative Theory of Terminology (1999; 2003; 2009) emphasises the importance of context. Our study confirms that context plays a crucial role when evaluating the rendering of terms in scientific popular texts, which aim at communicating science to non-professionals. When analysing how terms are rendered, it is appropriate to do so in the context of a given communicative situation, which is equally important in the analysis of machine translation.

The relevance of machine translation quality assessment is emphasized in various other recent studies, mainly for the reasons of machine translation quality improvement, and evaluation by way of human metrics is considered a useful methodology (Chatzikoumi 2020). The chosen method to collect data for assessment of machine translation quality in this study is in line with similar research on other languages. For example, Rivera-Trigueros (2022: 609)

reports a systematic literature analysis of works on machine translation quality assessment and a finding that as many as 86.7% of studies employed questionnaires using Likert type scales as the method for machine translation quality assessment in isolation or a combination with other methods, along with 52% of the studies employing analyses of errors committed by machine translation.

Lithuanian is considered to be low-resourced language which may demonstrate a lower quality of machine translation. However, our study results may point to two inferences. On the one hand, the findings contradict the conclusions reached in a study by Wu et al. (2011: 1298) who state that translation quality is poor in languages acquiring small training of corpora. A decade since Wu et al.'s study has led to a breakthrough in machine translation with the neural networks so that the quality improved to such an extent that from the perspective of terminology translation it might be regarded as sufficient. On the other hand, even if terminology rendering in the target language offers high accuracy (which is a common finding of this and other previous studies), machine translation cannot replace human translators. As confirmed in Haddow et al.'s (2021) study, among many others, the fluency criterion in longer portions of text is much lower than adequacy of text or accuracy of terms used.

Studies on the quality of machine translation of terminology in the English-Lithuanian language pair are not abundant. One such study on the machine translation of artificial intelligence terms has reported that the quality offered by *Google Translate* is only to some extent sufficient (Kvyklytė & Mikelionienė 2022). The typology for terminology errors in machine translation which was used in the study had been suggested by Haque et al. (2020: 163–164). The conducted study demonstrated that two-thirds of artificial intelligence terms were translated inappropriately. The most common errors were those of omitted, unrecognized or, in general, untranslated terms, which was also found in machine translation of medical terms presented in this paper. The emergence of new terms is noted to be one of the reasons for the lack of translation quality in rendering terms (Moghadam & Far 2015). Thus, the overall ratio of well-rendered to poorly rendered terms is more favorable in the translation of medical texts than in translation of artificial intelligence terms, perhaps due to the lower number of terms denoting innovations, or due to internationally better established and structured medical terminology.

In a study of three machine translation systems focusing on human evaluation, Fernández-Torné & Matamala (2021) compare translation of industrial documentation from Spanish to German. The research included assessment of fluency, adequacy, and ranking at the segment level. At the segment level fluency (on a 1 to 4 scale) and adequacy (measured on another 4-point scale) were assessed, and ranking (placing in order different translated versions of the same original segment from best to worst quality) was implemented. Adequacy of 81% of the segments processed by neural machine translation was rated in the high range (3 and 4 for the most segments), while in terms of fluency at the segment level, 70% of the segments from neural machine translation appeared in the high range (Fernández-Torné & Matamala, 2021: 104–107). Although the study design and the scale are somewhat different from our research, presented in this paper, some similarities might be established regarding adequacy and fluency results. In both studies, the fluency scores were lower than the adequacy scores, even though one study analysed machine translation from a high-resource language (Spanish) to another high-resource (German) language, and another study focused on machine translation from a high-resource language (English) to a low-resource language (Lithuanian). Therefore, the human post-editing effort can not be eliminated and is a crucial aspect in producing machine translation-enhanced high quality translation.

6. Conclusion

The results of this study on human evaluation of machine-translated popular science texts containing medical terminology reveal that the performance of machine translation systems is somewhat sufficient but the output requires careful review. In general, machine translation works well recognizing and rendering terms, i.e. the accuracy criterion is satisfactory. Only a small proportion of terms in our analysis were rendered inappropriately, either as a term that is inconsistent with terminological resources or as a wrong term. However, Lithuanian, considered a low-resourced language, cannot boast sufficient approved terminological resources, especially of medical and healthcare terminology that would be freely available to translators and the general public. Therefore, adding more medical and healthcare terms (even the most basic and general ones) in various term banks would be desirable. Besides, creating specialised corpora that could be used to train machine translation systems are prerequisite so that higher quality machine translation output could be achieved and less post-editing effort would be required for greater productivity and efficiency. For that, large amounts of domain-specific data are needed.

The study also revealed that, in terms of longer segments, adequacy of medical texts containing medical terms was rated to be rather good. However, it is the fluency at the level of segments (sentences) that machine translated texts fail to comply with, which may result in miscomprehension and/or information uselessness. The quality of the output is acceptable only for quick comprehension of medical texts. The topic of machine translation of terminology might be a fruitful area for further work, as several questions still remain to be answered. A natural progression of this work may be to analyze the readability of machine-translated segments by an ordinary reader and end user of machine translation. The research contributes to raising awareness of the importance of efficient and reliable medical communication between healthcare professionals and broad society (patients, etc.) and stresses the need for responsible and ethical use of modern language technologies.

References

Arnold, Doug; Balkan, Lorna; Meijer, Siety; Humphreys, Lee R. & Sadler, Loisa. 1994. *Machine translation: An introductory guide*. Manchester: Blackwell.

Cabezas-Garcia, Melania & Faber Benitez, Pamela. 2017. Exploring the semantics of multi-word terms by means of paraphrases. In: *Temas actuales de terminología y estudios sobre el léxico*, 193–211. Availabale at: https://dialnet.unirioja.es/servlet/autor?codigo=187932

Cabré Castellví, M. Teresa. 1996. Terminology today. In Harold Somers (ed.). *Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager*, 15–34. Available at: https://doi.org/10.1075/btl.18.04cab

Cabré Castellví, M. Teresa. 1999. T*erminology: Theory, methods and applications*. Amsterdam: John Benjamins. Available at: https://doi.org/10.1075/tlrp.1

Cabré Castellví, M. Teresa. 2003. Theories of terminology. Their description, prescription and explanation. *Terminology* 9(2), 163–199.

Cabré Castellví, M. Teresa. 2009. The Communicative Theory of Terminology, A Linguistic Approach of Terms. *Revue française de linguistique appliquée* XIV. 9–15. Available at: https://doi.org/10.3917/rfla.142.0009

Cadwell, Patrick; O'Brien, Sharon & Teixeira, Carlos S. C. 2018. Resistance and accommodation: Factors for the (non-) adoption of machine translation among professional translators, *Perspectives* 26(3), 301–321. DOI: 10.1080/0907676X.2017.1337210

Calsamiglia, Helena & van Dijk, Teun A. 2004. Popularization Discourse and Knowledge about the Genome. *Discourse and Society* 15(4): 369–389. Available at: https://doi.org/10.1177/09579265040437

Campo Bayón, Maria do & Sánchez-Gijón, Pilar. 2019 Evaluating machine translation in a low-resource language combination: Spanish-Galician. *Proceedings of MT Summit XVII, volume 2,* Dublin, Aug. 19–23, 2019. Available at: https://aclanthology.org/W19-6705.pdf

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico & Moorkens, Joss. 2018. Approaches to human and machine translation quality assessment. In: Joss Moorkens, Sheila Castilho, Federico Gaspari & Stephen Doherty (eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, 9–38. Cham: Springer.

Chase, Marilyn. 2006. Infectious Diseases. In *A Field Guide for Science Writers.* 2nd edition, eds. Deborah Blum, Mary Knudson and Robin Marantz Henig. New York: Oxford University Press.

Chatzikoumi, Eirini. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* 26(2). 137–161. DOI:10.1017/S1351324919000469

Ciapuscio, Guiomar E. 2003. Formulation and Reformulation Procedures in Verbal Interactions Between Experts and (Semi-) Laypersons. *Discourse Studies* 5 (2): 207–233. DOI:10.1177/1461445603005002004

Cvilikaitė, Jurgita. 2008. Leksinės mašininio vertimo klaidos: beekvivalenčių žodžių vertimas. *Filologija* [Philology]13. 27–38. Available at: http://talpykla.elaba.lt/elaba-fedora/objects/elaba:6138715/datastreams/MAIN/content

Černý, Miroslav. 2008. Some observations of the use of medical terminology in doctor-patient communication. *SKASE Journal of Translation and Interpretation* 3(1). 39–53. http://www.skase.sk/Volumes/JTI03/pdf_doc/Czerny.pdf

Daubarienė, Audronė & Ziezytė, Greta. 2013. Machine translation: translated texts in terms of standards of textuality. *Kalbų studijos* [Studies about languages] 22, 55–61.

Daudaravičius, Vidas. 2006. Pradžia į begalybę. Mašininis vertimas ir lietuvių kalba. *Darbai ir dienos* [Deeds and days] 45, 7–18.

Doherty, Stephen; Gaspari, Federico; Groves, Declan & van Genabith, Josef. 2013. *Mapping the industry I: Findings on translation technologies and quality assessment. Technical Report.* Available at: https://www.researchgate.net/publication/257227488_Mapping_the_Industry_I_Findings_on_Translation_Technologies_and_Quality_Assessment/link/00b7d524ac2f177643000000/download

Doherty, Stephen. 2016. The Impact of Translation Technologies on the Process and Product of Translation. *International Journal of Communication* 10. 947–969. Available at: https://ijoc.org/index.php/ijoc/article/viewFile/3499/1573

Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In: Dorothy Kenny (ed.). *Human issues in translation technology*, 154–178. London: Routledge.

Fage-Butler, Antoinette M. & Jensen, Matilde Nisbeth. 2016. Medical terminology in online patient–patient communication: evidence of high health literacy? *Health Expectations* 19(3). 643–653. DOI: 10.1111/hex.12395

Fernández-Torné, Anna & Matamala, Anna. 2021. Human evaluation of three machine translation systems: from quality to attitudes by professional translators. *Vigo International Journal of Applied Linguistics 18.* 97–121. DOI: https://doi.org/10.35869/vial.v0i18.3366

Fóris, Ágota. 2022. Terminology Work in the Translation Project Process. The Hungarian Perspective. *Terminologija* [Terminology] 29. 45–65. DOI: doi.org/10.35321/term29-03

Haddow, Barry; Birch, Alexandra & Heafield, Kenneth. 2021. Machine translation in healthcare. in Şebnem Susam-Saraeva & Eva Spišiakova (eds.), *The Routledge Handbook of Translation and Health.* 108–129. London: Routledge.

Haque, Rejwanul; Hasanuzzaman, Mohammed & Way, Andy. 2020. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation* 34. 149–195. Available at: https://doi.org/10.1007/s10590-020-09251-z

Kasperavičienė, Ramunė; Motiejūnienė, Jurgita & Patašienė, Irena. 2020. Quality assessment of machine translation output: Cognitive evaluation approach in an eye tracking experiment. *Texto livre: linguagem e tecnologia* 13(2), 1–16. Belo Horizonte: Universidade Federal de Minas Gerais. eISSN 1983-3652. DOI: 10.17851/1983-3652.13.2.%25p

Kvašytė, Regina. 2005. *Mokomasis terminologijos žodynėlis*. Šiauliai: Šiaulių universiteto leidykla. Available at: https://etalpykla.lituanistika.lt/fedora/objects/LT-LDB-0001:B.03~2005~1367153714479/datastreams/DS.001.0.01.BOOK/content

Kvyklytė, Urtė & Mikelionienė, Jurgita. 2022. Evaluation of machine translated artificial intelligence terms in respect of fluency and adequacy. *TOTh 2022 – Onsite & Online Conference "Terminology & Ontology: Theories and Applications",* 175–194. Chambéry, France, 2–3 June 2022.

Leane, Elizabeth. 2007. *Reading Popular Physics: Disciplinary Skirmishes and Textual Strategies.* Aldershot, Hampshire/Burlington, VT: Ashgate.

León-Araúz, Pilar; Cabezas-García, Melania & Reimerink, Arianne. 2020. Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2358–2367, Marseille, 11–16 May 2020. Available at: https://aclanthology.org/2020.lrec-1.287.pdf

Levanaitė, Karolina. 2021. Lietuvos vertimo rinkos dalyvių požiūris į mašininį vertimą ir postredagavimą. *Vertimo studijos* [Studies in translation] 14. 22–39. DOI: 10.15388/VertStud.2021.2.

Lommel, Arle R.; Uszkoreit, Hans & Burchardt, Aljoscha. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica* 12. 455–463. Available at: https://revistes.uab.cat/tradumatica/article/view/n12-lommel-uzskoreit-burchardt/pdf

Lommel, Arle. 2018. Metrics for translation quality assessment: a case for standardising error typologies. In: Joss Moorkens, Sheila Castilho, Federico Gaspari & Stephen Doherty (eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications*, 109–127. Cham: Springer. Available at: https://doi.org/10.1007/978-3-319-91241-7_6

Manfredi, Marina. 2019. *Popular Science Discourse in Translation: Translating "Hard", "Soft", Medical Sciences and Technology for Consumer and Specialized Magazines from English into Italian.* Universitat Jaume I.

Mehandru, Nikita; Robertson, Samantha & Salehi, Niloufar. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2016–2025. June 2022. Available at: https://dl.acm.org/doi/10.1145/3531146.3533244

Miltakienė, Eglė. 2021. Translation error analysis of non-fiction book titles: a case study on neural machine translation systems. *SMILES 2020: Social Sciences, Arts and Humanities in Contemporary Society*, 13–18. Available at: https://doi.org/10.5755/e01.2783-5820.2021.1

Myers, Greg. 2003. Discourse Studies of Scientific Popularization: Questioning the Boundaries. *Discourse Studies* 5(2), 265–279. Available at: https://doi.org/10.1177/1461445603005002006

Moghadam, Masoumeh Y. & Far, Mansureh D. 2015. Translation of technical terms: A case of law terms. *Journal of Language Teaching and Research* 6(4), 830–835. Available at: https://doi.org/10.17507/jltr.0604.16

Munday, Jeremy. 2004. Advertising: Some Challenges to Translation Theory. *The Translator* 10(2). 199–219. DOI: 10.1080/13556509.2004.10799177

MQM Council. n. d. *MQM (Multidimensional Quality Metrics)*. Available at: https://themqm.org/

MQM Error Typology, n.d. Available et: http://themqm.info/typology/

Nelkin, Dorothy. 1987. *Selling Science: How the Press Covers Science and Technology.* New York: Freeman.

OECD. 2007. *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual.* Available at: https://www.oecd.org/science/inno/38235147.pdf

Petkevičiūtė, Inga & Tamulynas, Bronius. 2011. Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas. *Kalbų studijos [Studies about languages]* 18, 38–45. Available at: https://etalpykla.lituanistika.lt/fedora/objects/LT-LDB-0001:J.04~2011~1367174892606/datastreams/DS.002.0.01.ARTIC/content

Pinnis, Mārcis & Bergmanis, Toms. 2020. Tilde's neural machine translation technology. In Ojārs Spārītis, Ilze Trapenciere & Namejs Zeltiņš (eds.), *Latvian Academy of Sciences*, p. 85. Available at: https://www.researchgate.net/profile/Alexey-Lihachev/publication/343905905_Latvian_Academy_of_Science

Popović, Maja. 2020. Informative manual evaluation of machine translation output. *Proceedings of the 28th International Conference on Computational Linguistics. 5059–5069.* Available at: https://doi.org/10.18653/v1/2020.coling-main.444

Povilaitienė, Milda & Kasperė, Ramunė. 2022. Machine translation for post-editing practices. *Current Trends in Language Development* 24. 5–15. Available at: https://doi.org/10.31392/NPU-nc.series9.2022.24.01

Rask, Nina. 2008. *Analysis of a Medical Translation. Terminology and cultural aspects.* Available at: https://www.diva-portal.org/smash/get/diva2:206300/FULLTEXT01.pdf

Reeder, Florence. 2004. Investigation of intelligibility judgments. In: Robert E. Frederking & Kathryn B. Taylor (eds.). *Machine Translation: From Real Users to Research*. AMTA 2004. Lecture Notes in Computer Science, vol 3265. Berlin: Springer. Available at: https://doi.org/10.1007/978-3-540-30194-3_25

Rimkutė, Erika & Kovalevskaitė, Jolanta. 2007a. Mašininis vertimas – greitoji pagalba globalėjančiam pasauliui. *Gimtoji kalba* [Native language] 9. 3–10.

Rimkutė, Erika & Kovalevskaitė, Jolanta. 2007b. Pastabos apie žmogaus ir mašininį vertimą. *Gimtoji kalba* [Native language] 10. 11–20.

Rimkutė, Erika & Kovalevskaitė, Jolanta. 2008. Linguistic evaluation of the first English-Lithuanian machine translation system. In T*hird Baltic conference on human language technologies: Proceedings.* 257–264. Vilnius: Lietuvių kalbos institutas.

Ringailienė, Teresė. 2014. *Popular Scientific Discourse in English and Lithuanian: A Multimodal Perspective*: Kaunas: Vytautas Magnus University. (Doctoral dissertation). Available at: https://gs.elaba.lt/object/elaba:2121672/

Rivera-Trigueros, Irene. 2022. Machine translation systems and quality assessment: a systematic review. *Lang Resources & Evaluation* 56. 593–619. Avalable at: https://doi.org/10.1007/s10579-021-09537-5

Skianis, Konstantinos; Briand, Yann & Desgrippes, Florent. 2020. Evaluation of Machine Translation Methods applied to Medical Terminologies. *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. 59–69. November 20, 2020. Available at: https://aclanthology.org/2020.louhi-1.7.pdf

Stankevičiūtė, Gilvilė; Kasperavičienė, Ramunė & Horbačauskienė, Jolita. 2017. Issues in Machine Translation. *International Journal on Language, Literature and Culture in Education* 4(1). 75–88. Available at: https://doi.org/10.1515/llce-2017-0005

Tinker Perrault, Sarah. 2013. *Communicating Popular Science: From Deficit to Democracy*. Basingstoke/New York: Palgrave MacMillan.

Váradi, Tamás. 2006. Multiword Units in an MT Lexicon. In: *EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*. 73–77. Trento, Italy, 3 April, 2006. Available at: https://aclanthology.org/W06-2410.pdf

Wu, Cuijun; Xia, Fei; Deleger, Louise & Solti, Imre. 2011. Statistical machine translation for biomedical text: are we there yet? *AMIA Annual Symposium Proceedings.* 1290–1299. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243244/

Xia, Mengzhou; Kong, Xiang; Anastasopoulos, Antonios & Neubig, Graham. 2019. Generalized Data Augmentation for Low-Resource Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5786–5796. Available at: https://arxiv.org/pdf/1906.03785.pdf

Zappatore, Marco & Ruggieri, Gilda. 2024. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language* 84. 1–46. Available at: https://www.sciencedirect.com/science/article/pii/S0885230823001018

Ramunė Kasperė, ramune.kaspere@ktu.lt; Jurgita Mikelionienė, jurgita.mikelioniene@ktu.lt; Dalia Venckienė, dalia.venckiene@ktu.lt

Kaunas University of Technology,
A. Mickevičiaus st. 37, Kaunas LT-44244,
Lithuania