Linguistica Silesiana nr 46/1, 2025
ISSN 0208-4228
DOI: 10.24425/linsi.2025.155072

EDYTA ŹRAŁKA
University of Silesia in Katowice, Poland
Institute of Linguistics
ORCID: 0000-0002-2139-0955
edyta.zralka@us.edu.pl

QUALITY ASSESSMENT AND POST-EDITING IN RAISING THE EFFICACY OF LEGAL MACHINE TRANSLATION

With the occurrence of Machine Translation (MT) tools, their reduced efficiency was observed; hence, rules for translation evaluation were introduced among professionals. This paper examines the issue of Google Translate's (GT's) performance in the translation of legal texts from English into Polish based on an excerpt from *The Louisiana Civil Code*, which is studied in terms of errors made by GT and quality improvement in the post-editing (PE) process based on a manual evaluation metric chosen. In a comparative study, possible GT errors were classified in order to carry out PE and streamline legal MT while maintaining quality.

Keywords: Machine Translation, GT performance, translation quality assessment, post-editing, translation metrics

1. Introduction

The idea of translation quality assessment (TQA), promoted by House (1997) and Hatim (1998), became particularly important with the spread of Machine Translation (MT), given the need to post-edit machine-translated texts by companies and freelance professional translators in order to accelerate the speed of the translation process and improve the quality of the translation product. The occurrence of free MT tools like Google Translate (GT) encouraged not only people knowledgeable in the trade but also ordinary users to start applying MT tools for professional or amateur tasks connected with translation. The possibly reduced reliability of such tools was quickly noticed. At the same time, a quest for methods to improve communication and GT outputs' quality even in specialised areas was embarked on.

The paper aims to examine GT performance in the English-Polish translation of a legal text based on Roman provenance (Młodawska 2012). The text is an

excerpt from *The Louisiana Civil Code* concerning partnership, used as the research material studied in order to find GT errors and propose ideas for quality refinement in the process of post-editing (PE).

The reason for recognising GT as a tool for handling specialised translations is its accessibility and improvement in translation correctness together with efficiency after the introduction of Neural Machine Translation in 2016, as Kenny states (2018: 434). Even if it has not been particularly recognised as a trustful tool so far, compared with, e.g. DeepL, it arouses researchers' curiosity how effective GT translations might be in specialised texts. However, a certain range of errors specific to MT justifies caution. In the case of legal texts, a discouraging factor could also be so-called risk management, referring to such issues as data breaches or liability for the effects of wrong translations (Nitzke et al. 2019). The research questions investigated in this paper are whether GT acceptably renders legal texts, and how to improve the translation quality in the PE process.

In the paper, errors made by GT in the initial search for equivalents were classified and quality assessment criteria were proposed based on a comparative study of the Source Text (ST) in English and the Target Text (TT) in Polish. The TT was evaluated based on particular criteria selected according to legal language characteristics described by Mellinkoff (2004) and Jopek-Bosiacka (2006).

The research reported on in the paper provides a quantitative and qualitative GT assessment of the sample translation. It also provides a set of ideas for a manual metric that could be used to evaluate GT translations and then perform the PE process.

2. Concepts of TQA in the evaluation of legal translations by GT

When MT quality assessment is undertaken, one has to be aware that, regardless of whether it is carried out by computer programmes or human beings, the logical principle of correctness must be applied, so the standard rules of assessment concerning translation quality must be adhered to.

2.1. Rules of TQA in legal translations

Views on the quality assessment of TT over the years have been based on equivalence (House 1997, 2015). Seeking equivalence applies to all kinds of translations, regardless of whether they are carried out by human translators or MT systems. House (2015: 10) states that the 'equivalence of response' in the Source Language (SL) and the Target Language (TL) is a criterion that helps formulate less subjective statements in the TQA. This idea can be adopted in the

quality assessment of legal translations. An evaluator must consider Hatim's suggestion (1998), derived from House, which states that what matters in the assessment is to combine the equivalence with the function of translation. This function is considered a mixture of language functions and text functions (based on the views of Reiss 1971, Halliday 1978, and Bühler 1990). House believes that "different language functions can coexist within the so-called individual text's function" (House 1997: 31). When translating legal texts, it is essential to be aware of their specificity (e.g. lexis, grammar, phraseological structures, syntax, punctuation) and their cultural dependence. These are elements that imply the application of the general idea of a source-text oriented translation. However, the TT still has to be readable to the target receiver in its functional aspect and, first and foremost, there cannot be misinterpretations leading to the choice of wrong equivalents. This strategy is considered essential in legal translation by Matulewska (2016: 65), who insists on "adjusting the target text to the communicative needs and requirements of the community of recipients" but considers "source-text legal reality" to be one of the dimensions in her parametrisation methodology (2016: 67, 75-76 and 2013). She encapsulates the whole problem claiming that:

> Practising translators rarely recognize the need to adjust the translation of a given term to the communicative needs of translation recipients. They look for universal equivalents, good for any communicative situation. However, a target text which is communicatively effective in one situation, may not be equally effective in another. (2016: 77)

Taking into consideration different communicative situations in TTs, she formulates Rule 1 of her methodology stating the dependency of TT equivalents on both the ST and TT communicational needs.

Hatim observes a similar correlation. He realizes (1998: 95) that a full reproduction of specific ST functions is not possible in translation because, as House states (1997: 67), "the ST is tied to a specific non-repeatable historic event in the source culture [...] or because of the unique status that the source text has in the source culture". Hatim (1998: 95) concludes that if "situationality" existing between the two texts cannot be fully reproduced, "a second-level function" is to be fulfilled by the translation. The function, understood as such, has to combine ways of perception by ST and TT readers; in other words, it "must hold not only for the contemporary target language readers but also for their counterparts in the source culture" (House 1997: 191).

Ramos (2015) and the authors mentioned by him refer to the criteria of translation evaluation incorporated into legal translations. The most systematic of them, at the same time sharing the features of other classifications, seems to be the one by Brunette. The author differentiates the following criteria of TQA (2000: 175–177):

– Logic (depending on coherence and cohesion), directed to the target audience, not the ST,
– Purpose (effect and intention),
– Context, defined as "non-linguistic circumstances surrounding the production of the discourse to be assessed" (2000: 178),
– Language norm (rules and conventions of the language).

Through addressing the purpose and context, she refers to the criteria introduced by House. Language rules and logic remain standard linguistic components of assessing any text, also translated, and here the author cares about the quality of TT by stressing its logical structure.

The classifications by Mossop (2007: 125–139), Colina (2008: 103-106), and Angelelli (2009: 40-41) mostly refer to typical comparative elements of source and target texts to be taken into consideration when evaluating. Mossop additionally includes layout and organisation among the evaluative criteria, which might be crucial in some legal texts. Angelelli sees "translation skill" as a criterion of evaluation in translation, which seems to be an element encapsulating all other criteria.

2.2. Evaluation of MT quality in the translation of legal texts

In order to produce both linguistically and functionally good quality machine-translated legal texts, translators and evaluators should follow the rules of translation quality that apply to legal texts.

Regarding the evaluation of MT, Forcada (2010), Maučec and Donaj (2020), and many other scholars distinguish manual MT evaluation from automatic MT evaluation. The former is referred to as "human" evaluation, the latter as "an algorithm that can be coded into a program and run by a computer that calculates the evaluation score, which tells the user how good a translation is" (Maučec and Donaj 2020: 149). Using these two types of evaluation, the quality of MT output is measured as a final product, or the MT usability for subsequent corrections to be made by human translators, namely PE, is assessed.

Computer programs used for quality assessment are based on language rules and existing human translations. The aim in using them is to "try to measure how close each raw machine-translated sentence is to one or more reference human translations" (Forcada 2010: 221). In the case of a human assessment, the professionals use their linguistic and cultural knowledge in one or both languages. The human evaluation is normally accompanied by PE.

As stated by Forcada (2010: 221):

> MT quality in general terms has proven to be very difficult, and indeed, the adequacy of raw output may vary from one purpose to another purpose. For instance, a raw machine

> translated text may be almost perfectly understandable to a native speaker of the TL but may still need heavy post-editing to make it fit for publishing. And, conversely, MT errors that make a substantial part of the raw text unintelligible for that native speaker may be very easy to spot and correct by a skilled post-editor.

According to Maučec and Donaj (2020: 150), it is the human evaluation that remains the most common option for evaluating MT quality. The reason might be Kenny's (2018: 437) belief that:

> MT also remains a technology that cannot explain or take responsibility for its decisions in the way a human translator might be expected to, and it is still prone to errors that might only be spotted by an informed bilingual human. For these reasons, bilingual humans will continue to be important arbiters in professional workflows that use MT.

In Human Assessment Methods, the following criteria are applied (Han et al. 2021: 2):

– Traditional: 1. intelligibility and fidelity, 2. fluency, adequacy, comprehension, 3. further development to the criteria of manual evaluation,
– Advanced: 1. task-oriented, 2. extended criteria, 3. utilizing, post-editing, 4. segment ranking, 5. crowd, source intelligence, 6. revisiting, traditional criteria.

This classification is a set of guidelines to be considered when creating manual metrics. Still, automatic evaluation metrics are lower-cost alternatives to human evaluation (Maučec and Donaj 2020: 151), and this is the reason why they are so common in MT assessment.

Automatic Quality Assessment Methods are based on the following criteria (Han et al. 2021: 2):

– Traditional: N-gram surface matching (1. word order, 2. precision and recall, 3. edit distance) and deeper linguistic features (1. syntax, 2. semantics),
– Advanced: Deep Learning Models.

As Han et al. (2021: 4) state, the assessment based on the above-mentioned criteria is not always applicable in a practical context. The PE for quality improvement is not always necessary. According to Forcada (2010: 217), MT can be applied for the aim either of 'assimilation' or of 'dissemination'. The former occurs when "one does not understand the SL and wants to have an approximate idea of the content of the text, its gist" (Forcada 2010: 217). PE is in this case not vital at all. In the case of the dissemination, "texts are machine-translated as an intermediate step in the production of a document in the TL that will be published (disseminated); raw MT results have to be post-edited" (Forcada 2010: 217).

If the raw texts have to be post-edited by professionals, as has been stated, they are evaluated simultaneously as a rule. This means that there must be some general criteria known to evaluators. Among the criteria put forward by Han et al. (2021), two deserve particular attention as they are generally considered to be the most important ones:

– FLUENCY, also referred to as intelligibility (Forcada 2010: 221), requiring an expert fluent in the TL, where the degree of adherence to the TT and TL norms for grammaticality and clarity is assessed, with no ST relevance,
– ACCURACY, also referred to as fidelity or adequacy (Forcada 2010: 221), requiring a bilingual evaluator and access to both ST and TT, where, based on ST norms and meaning, the evaluation reveals how well the TT represents the ST content.

According to Maučec and Donaj (2020: 151), adequacy and fluency are usually judged on a five-point scale. Accuracy, characterised by Han et al. (2021: 2) as the criterion that "includes the requirement that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original", can capture:

– All meaning – five points,
– Most meaning – four points,
– Much meaning – three points,
– Little meaning – two points,
– None of the meaning – one point (Maučec and Donaj 2020: 151).

Fluency evaluation determines whether a sentence is well-formed and fluent in context (Han et al. 2021: 2). In the ARPA (the Advanced Research Projects Agency) project from the 1990s, fluency is evaluated on the basis of whether the translation is good English without reference to the correct translation. Again, typically, evaluation metrics differentiate five levels of quality within fluency:

– Flawless language – five points,
– Good language – four points,
– Non-native language – three points,
– Disfluent language – two points,
– Incomprehensible language – one point (Maučec and Donaj 2020: 151).

Similar methods among the so-called Subjective Evaluation Criteria are discussed by Tomás et al. (2003). They are based on human intervention in the process of evaluation. Among the most widely used methods, the authors describe the metrics called Subjective Sentence Error Rate (SSER), in which each sentence is scored in ten categories according to its translation quality. These categories are:

0 – nonsensical,
1 – some aspects of the content are conveyed,
[…]
5 – comprehensible, but with important syntactic errors,
[...]
9 – OK. Only slight style errors,
10 – perfect (Tomás et al. 2003: 28).

Based on the evaluation, a human evaluator can do PE, which is usually undertaken when quality matters. PE occurs when a professional translator applies an MT tool for his/her purposes to produce a better-quality translation and when companies employ evaluators to raise the quality level of machine-translated texts.

3. Research material and research methodology

In the process of evaluation and PE, which can be assumed to raise the quality of any machine-translated text, error analysis is a crucial factor. In order to carry out the analysis systematically, instruments to aid the process, called metrics, adjusted to the particular needs of the evaluated language, are created to retrieve errors.

3.1. Research methodology

The methodology applied in this research is based on the criteria introduced in the 1990s by the Advanced Research Projects Agency (ARPA) (Han et al. 2021: 2), despite the recent promotion of the model of evaluation called "document-level", which incorporates discourse features in MT evaluation (Maruf et al. 2019: 2). The ARPA model represents a systematic view, more verifiable in the analysis, and which is still used, for example, in the very common BLEU and METEOR automatic metrics. According to the ARPA model's rules, referred to as a sentence-level evaluation, units based on syntactic constituents containing sufficient information are assessed in terms of adequacy on a scale of 1 to 5. While in the ARPA methodology results are calculated based on the average of the judgments over all of the decisions in the translation set, in the present analysis, a percentage of the errors observed in the meaningful units of the GT translation has been computed. Each error has been noted, and then, the number of errors has been compared with the total word count in the text analysed. Regarding fluency, as in the ARPA project, judgments have been made based on each particular sentence and the evaluation exposes how well-formed and fluent in context the sentences are. In practice, the fluency assessment results are correlated with the adequacy criterion.

In order to make the analysis more referred to every textual element of the GT translation than just to counting errors in juxtaposed sentences, a more detailed manual evaluation has also been performed (by a translator having over 20 years of experience in the profession). It is based on the model described, allocating five points for accuracy and five points for fluency to the lines of the GT translation as organised by *LF Aligner* programme. There are 250 lines (excluding the introductory eight limes with instructions). Each line of the ST is paired with the GT translation and they have been individually assessed for accuracy and fluency.

Neural Machine Translation (NMT), which operates in GT, is often evaluated based on automatic metrics. An obstacle to use such a metric in this study is the lack of access to human-translated texts for reference. In order to compare the results of a manual evaluation with an automatic one, the quality evaluation by the BLEU metric has been performed, for which the GT post-edited translation has been used as reference. The BLEU algorithm provides a precision score, calculated by counting the number of words in the machine-translated text that match the reference translation and dividing that number by the total number of words in the machine-translated text. Higher BLEU scores indicate better translation quality, while lower scores indicate lower quality (high translation scores are considered to be in the range of 40-60%, not 100%, which is impossible to be obtained in practice).

This BLEU metric evaluation has been carried out in order to apply the existing methods rather than to obtain significant results. Due to the lack of professional translation of the ST to be used in the BLEU metric evaluation, the score obtained cannot be treated as illustrative. Rivera-Triguero (2022: 597) observes that the existence of good quality human translation is a decisive element in the automatic assessment reasonability. The author believes that:

> […] it must be taken into account that the majority of automated metrics require reference translations created by humans and, in many cases, the quality of these translations is assumed, but not verified, which could introduce an element of subjectivity […].

3.2. Characteristics of the research material considered in the evaluation

As a legal text, *The Louisiana Civil Code*, influenced by multiple revisions now, possesses a variety of features called by Mellinkoff "the chief characteristics of the language of the law" (2004: 11), among which he and Charrow et al. (2015: 175) mention:

– frequent use of common words with uncommon meanings (e.g. *action* for *lawsuit*),
– deliberate use of words and expressions with flexible meanings,

– attempts at extreme precision of expression by using synonyms or near-synonyms combined in binomials (the so-called doublets or triplets),
– frequent use of archaic Old and Middle English words (*aforesaid*, *whereas*, *said* and *such* as adjectives, etc.),
– frequent use of Latin words and phrases (*in propria persona*, *amicus curiae*, etc.),
– use of French words not included in the general vocabulary (e.g. *lien*),
– use of terms of art – or what is called jargon (*month-to-month tenancy, negotiable instrument*, etc.),
– use of argot – in group communication or "professional language" (*pierce the corporate veil, damages, due car*).

Veretina (2012: 104-105) mentions other lexical features of legal English, such as:

a) Archaisms:

– g. adverbs: *hereinafter*, verbs: *darraign*, nouns: *surrejoinder*, and adjectives: *aforesaid*,
– prepositional phrases, e.g. *pursuant to*, *prior to*, and *subsequent to*,
– expressions beginning with *here* and *there* (*therein*, *hereunder*, *thereof*, and *thereto)*.

b) Technical terms (or terms of art);

– pure legal terms, e.g. *tort, patent, share, royalty, bailment,* and *abatement*,
– common words with uncommon meanings, i.e. polysemous lexemes which have specific meanings in legal English, e.g. *attachment*, *action, consideration, execute*, *party*.

Jopek-Bosiacka (2006: 23-24) adds to the features mentioned a grammar rule concerning the differentiation between the descriptive and prescriptive sense of *shall*, the first referring to legal norms when the present tense is used in Polish and the second denoting the normative sense of laws with the future tense structures used in Polish.

Veretina (2012: 103) also claims that specific stylistics of legal English rests in syntactic and textual characteristics. However, due to their insignificance in this research (substantial lack of the errors discussed), they are not included in the description.

4. Text analysis

As mentioned earlier, the analysis concerns an excerpt from *The Louisiana Civil Code*, Title XI – Partnership (3,349 words). Besides possessing some branch-specific features, the Source Language (SL) reveals some other language

characteristics that are substantially different from the Target Language (TL) properties. Maučec and Donaj (2020) clarify that MT is less successful if the translation is from a morphologically less complex to a morphologically more complex and highly inflected language, which is the case in the pair English-Polish. They enumerate the following problems arising from the morphological complexity of the TL:

– a large number of inflected word forms leads to data sparsity, which results in unreliable estimates in statistical MT and wrong declensions,
– the target word order may differ in the target sentence,
– an inaccurate translation of pronouns may occur,
– there is an opportunity to drop the subject in a sentence,
– there are differences in the expression of negation.

Such errors must be expected and should constitute the essence of the analysis. They belong to the grammatical category, but lexical errors are also frequently encountered.

4.1. Results

According to the results obtained, the GT's performance is more than satisfying. There were 125 errors altogether: 104 within the lexical category, 18 grammar mistakes, two classified as syntactic category errors and one error in spelling. For 3,349 words constituting the whole sample, 3.73% of the text was erroneously translated. Considering the accuracy criterion and its five-point scale, attributing one point on the scale to every 20% of text erroneously translated, we have to grant the translation five points and classify the accuracy on Level 5, indicating all meaning and flawless language. The same applied to fluency. The level here must also be the highest, and the TT read smoothly enough to function as communicative .It means that GT can produce a specialised TT at a quality level that is, no doubt, a good starting point for PE.

When more specific results are concerned (Figure 1), the largest category of errors was in lexis (3.1% of text), then the grammatical category (0.53% of text), the syntactic category (0.05% of text), and, finally, the spelling category (0.02% of text). Therefore, this was a negligible share of wrongly selected equivalents in the entire GT translation.

A more detailed analysis based on the alignment of *FL Aligner* programme shows that the general scores obtained when all points have been allocated to the lines, summed up and divided by 250, is 4.58 for accuracy and 4.63 for fluency. The results reach over 90% of the highest scores of five points in the case of accuracy and fluency (91.6% and 92.6%, respectively) and are not distant from the percentage of the correct translation calculated while enumerating errors (96.27% of correct translation). Despite the subjectivity of assessment, the

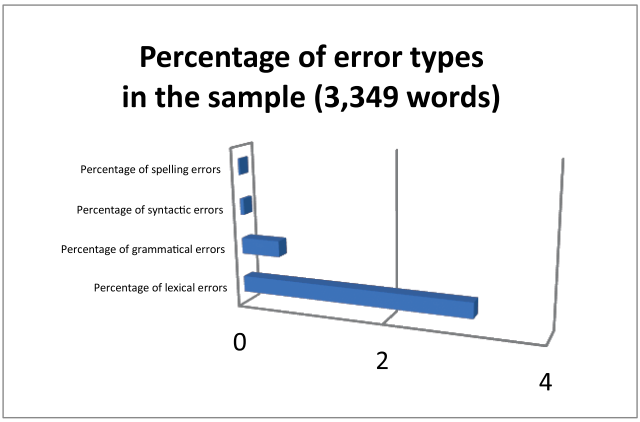

Figure 1. Percentage of different error categories in the sample

difference is natural when the analysis considers individual lines of the whole text and the fluency category is separately evaluated, not in the context of accuracy. Out of 250 lines, 69 have scored the highest five points in both categories. There are also seven lines for which a value lower than three has been given in both or one category. The examples are quoted in Table 1.

The total score obtained for quality by the BLEU metric is 10.37 (Fig. 3). Taking into account the use of GT post-edited translation in the BLEU assessment as a reference, not a real human translation, the result should not be

Table 1.GT translation lines with the lowest scores in the manual evaluation

Line no. according to *LF Aligner*	**ST line**	**GT translation**	**Accuracy**	**Fidelity**
63	[…] person. A partner may share his interest in the partnership with a third person without the consent of his partners, but he cannot make him a member of the partnership.	Wspólnik może dzielić się swoim zainteresowaniem spółką z osobą trzecią bez zgody wspólników, ale nie może go uczynić wspólnikiem spółki.	2	2
71-72	CHAPTER 3 - RELATIONS OF THE PARTNERSHIP AND THE / PARTNERS WITH THIRD PERSONS	ROZDZIAŁ 3 – RELACJE PARTNERSTWA I PARTNERÓW Z OSOBAMI TRZECIMI	2	4

Table 1. cont.

Line no. according to *LF Aligner*	ST line	GT translation	Accuracy	Fidelity
99-103	A. A partner ceases to be a member of a partnership upon: his / death or interdiction; his being granted an order for relief under / Chapter 7 or confirmation of a plan of liquidation or the / appointment of a trustee of his estate under Chapter 11 of the / Bankruptcy Code; his interest in the partnership being seized and not released as provided in Article 2819; his expulsion from the partnership; or his withdrawal from the partnership.	A. Wspólnik przestaje być wspólnikiem spółki osobowej z chwilą: śmierci lub zakazu; otrzymanie/a postanowienia o upłynnieniu jego majątku na podstawie rozdziału 7 lub potwierdzenia planu likwidacji lub powołania syndyka masy spadkowej na podstawie rozdziału 11 Kodeksu upadłościowego; jego interesu w zajęciu i niezwolnieniu spółki osobowej zgodnie z art. 2819; jego wydalenie/a ze spółki; lub jego wystąpienie/a ze spółki.	2	2
145	If the object becomes impossible, the partnership may be continued for a different object.	Jeżeli przedmiot stanie się niemożliwy, partnerstwo może być kontynuowane na inny przedmiot.	2	2
186	The liquidation of a partnership is not final until all its assets have been collected and applied to its obligations and its remaining assets, if any, have been appropriately distributed to the partners.	Likwidacja spółki nie jest ostateczna, dopóki cały jej majątek nie zostanie zebrany i zaspokojony jej zobowiązaniami, a ewentualny pozostały majątek nie zostanie odpowiednio rozdzielony między wspólników.	2	2
187	CHAPTER 7 –PARTNERSHIP IN COMMENDAM	ROZDZIAŁ 7 – PARTNERSTWO W KOMENDAM	2	2
192	Partnership in commendam; definition.	Partnerstwo w komendzie; definicja.	2	2

Figure 2. BLEU metric results

treated as illustrative, as already stated. Still, there is another factor to discredit the sense of BLEU metric application in this study observed by Jansen (2020), who states that:

> […] the neural method gives better results when assessed by humans than the automatic BLEU metric would indicate. This already known fact is explained by the specific construction of the BLEU metric, which favours "locally correct" translations. Neural translation, however, is focused more on analyzing connections between words that are distant from each other.[1]

4.2. Categories of errors

When the particular categories of errors are considered, it was clear that the lexical category contained the highest percentage of errors. It constituted in round numbers 83% of all the errors evidenced in the translation. The next category was the grammatical one. Grammar needed to be corrected in 14% in round figures. The remaining two categories – syntactic and spelling – were almost irrelevant, producing 2% and 1% of all errors in round numbers in this GT rendering. Such a ratio means that the post-editor must be primarily ready for lexical and grammar corrections and, simultaneously, fluent in terminology and specific grammar features of legal texts.

[1] Translation – Edyta Źrałka.

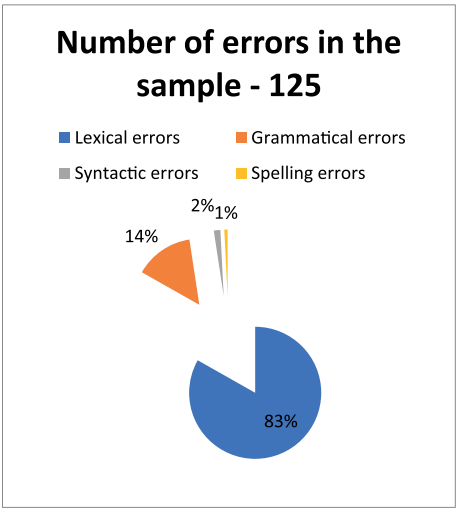

Figure 3. Ratio of the error categories in the sample

4.2.1. Lexical errors

The prevalence of lexical errors in the MT translation is connected with the fact that some terms and errors that they generate tend to be repeated. There were 37 cases when terms were wrongly used or partly correct (there were wrong grammatical elements in phrases constituting terms). The examples found comprise:

Wrong or partly correct terms (37):

(1) Art. 2812: share his interest in the partnership – *dzielić się* ***swoim zainteresowaniem*** *spółką* [instead of ***swoimi udziałami*** *w spółce*][2]
(2) Art. 2814: mandatary – ***zleceniodawca*** (meaning "mandator") [instead of ***zleceniobiorca***, which is, however, used twice correctly]
(3) CESSATION OF MEMBERSHIP – ***Ustanowienie*** *członkostwa* (meaning "establishing") [instead of ***ustanie***]
(4) Art. 2818 and 2826: interdiction – ***zakaz*** (meaninig "prohibition") [instead of ***ubezwłasnowolnienie***]
(5) Art. 2830: termination – ***wypowiedzenia*** [instead of ***rozwiązania spółki***]
(6) Art. 2837, 2838, 2839, 2840, 2844 B. (7) (f), Art. 2844 B. (7) (g) and 2844 B. (9): partnership/partner in commendam – ***partnerstwo/partner w komendzie*** [instead of ***spółka komandytowa*** – the equivalent normally given to "limited partnership", or ***komandytariusz*** – the equivalent mostly chosen for "a general partner", in this GT translation best rendered as *wspólnik komandytu*]

[2] The parts of the TT that are incorrectly rendered and the correct translations are written in bold.

Other renderings*:*

– *spółka partnerska/wspólnik (partner) w komanda* – Art. 2838, 2840, 2841, 2842, 2844 *partner w komandzie* – Art. 2839
– *wspólnik z polecenia/wspólnik w poleceniu* – 2842, 2843**.**

Another type of lexical errors the GT produced was retaining ST elements in phrases. This occurred when a phrase was not recognised as a whole equivalent in the TL. There were two such cases found in the sample.

Retention of ST elements in TT terms (2):

(7) Art. 2836: partnership in commendam - ***partnerstwa in commendam*** [instead of ***partnerstwo w spółce komandytowej***]
(8) Art. 2840: partner in commendam – ***partner in commendam*** [instead of ***komandytariusz***].

For a professional translator, especially one who serves as a court translator in Poland and wants to employ MT to accelerate the speed of the translation process (without violating the principle of confidentiality), a lexical error can also be the lack of original names of institutions and legal acts that are required or habitually recommended to be added to the translated versions of legal documents according to *Kodeks zawodowy tłumacza przysięgłego* (The Code of the Polish Court Translator, par. 47) and experts (Świgońska 2021).[3]

Lack of ST original quotations (3):

(9) Art. 2806: *z tytułem XI III księgi Kodeksu cywilnego* [the legal act quoted was not accompanied by the original name of the act; "Title XI of Book III of the Civil Code"]
(10) Art. 2826: *na podstawie rozdziału 7 Kodeksu Upadłościowego* [no "Chapter 7 of the Bankruptcy Code"]
(11) Art. 2818 A: *na podstawie rozdziału 11 Kodeksu upadłościowego* [no "Chapter 11 of the Bankruptcy Code"].

An evaluator must be particularly sensitive to MT outputs regarding homonyms existing in the TL. There were two such mistakes in the sample analysed.

Homonyms as equivalents (2):

(12) Art. 2807: to permit a partner to withdraw – *zezwolenia wspólnikowi* ***na wystąpienie*** [instead of ***na wystąpienie ze spółki***; *wystąpienie* is also "performance" so the use of *wystąpienie* without the prepositional phrase *ze spółki* accompanying it can be treated as a homonym meaning "performance" and be misleading]
(13) Art. 2844 B (7) (a): dissolution, termination – *rozwiązanie,* ***rozwiązanie*** [instead of *rozwiązanie,* ***zakończenie działalności***; *rozwiązanie* can have

[3] *Kodeks Zawodowy Tłumacza Przysięgłego. https://tepis.org.pl/kodeks-tlumacza-przysieglego/*

two meanings – referring to the termination of, e.g. a contract and cessation of something, of which the second meaning must be chosen here].

It is common in GT translation that all words in English doublets or triplets obtain an equivalent in Polish, whereas Polish legal language is not as precise, and one item is usually selected instead of redundancies if there is no difference in meaning.

Redundant doublets (2):

(14) Art. 2837: the powers, rights, and obligations – ***uprawnienia, prawa*** *i obowiązki* [instead of ***prawa*** *i obowiązki*]

(15) Art. 2844: rights and powers – ***praw i uprawnień*** [instead of ***uprawnień***].

Word-for-word translation leading to stylistic errors in (1):

(16) rights granted by law – ***praw*** *przyznanych przez prawo* [instead of ***uprawnień*** *nadanych zgodnie z prawem*; here *prawa* ("rights") and *prawo* ("law") coincide].

A fundamental problem in GT translations, as observed in the analysis, is the need for more consistency in selecting TL terms. Consistency in TTs, recommended as one of the features of a good translation, is not achieved, especially with regard to the terminology. Even if there are no errors in the formal sense, in such cases, a TT receiver is confused, and the translation should be revised.

Terminological inconsistency (57):

(17) Art. 2803: the distribution of assets – *podziału* ***majątku*** [occurring six times in the whole Chapter)

Art. 2804: *podziału* ***aktywów*** [used four times in the whole Chapter, seems a better equivalent]

(18) Art. 2805: partnership – ***spółka*** [used 140 times in the whole Chapter, being an adequately chosen equivalent], and on the other hand *partnerstwo* [occurring 18 times in the whole Chapter]

(19) Art. 2805, 2842: partners/partner (in commendam) – ***wspólnicy/partner*** *(w komenda* etc.) *partnerów* [31 times] or *wspólników* [79 times in the whole Chapter, at the same time a better equivalent, however, the preferable one is *komandytariusz*]

(20) Art. 2806: in writing – ***pisemna*** [1 case in the whole Chapter] or ***sporządzona na piśmie*** [4 times in the whole Chapter, being a more natural equivalent in legal Polish]

(21) Art. 2806: retroactivity – ***retroaktywność*** [one case in the whole Chapter] and ***moc wsteczna*** [two occurrences in the whole Chapter, at the same time a habitual equivalent in legal terms, even if it requires more complicated sentence structures].

4.2.2. Grammar errors

The next numerously represented category of errors was the one connected with grammar. There were 18 errors in this category found in the sample. The most serious mistakes were wrongly selected cases, which in the pair English-Polish is a common problem, as English is a non-inflected language contrary to Polish, which is highly inflected.

Wrongly selected cases, treated by Maučec and Donaj (2020: 146) as an obstacle in MT concerning highly inflected languages, could be observed ten times.

Wrong case (10): [NOM instead of GEN – 4 cases; NOM instead of INSTR – 5 examples; ACC instead of INSTR – 1 occurrence]:

(22) Art. 2806: between the date of acquisition and the date that the partnership was created – *między dniem nabycia oraz* ***datę*** [instead of ***datą***: ACC not INSTR] utworzenia spółki

(23) Art. 2818: his being granted an order for relief – ***otrzymanie*** *postanowienia* [instead of ***otrzymania***: NOM not GEN – the reason is probably using quotation marks "…"and taking what is after as a new part of a sentence]

(24) Art. 2818: his interest in the partnership being seized and not released – ***jego udział*** *w zajęciu i niezwolnieniu spółki osobowej* [instead of ***jego udziału***: NOM not GEN]

(25) Art. 2818: his expulsion from the partnership – ***jego wydalenie*** *ze spółki* [instead of ***jego wydalenia***: NOM not GEN]

(26) Art. 2818: or his withdrawal from the partnership – ***jego wystąpienie*** *ze spółki* [instead of ***jego wystąpienia***: NOM not GEN]

(27) Art. 2826: a partnership is terminated by: […] a judgment of termination – *rozwiązanie spółki następuje* ***za: […]*** *orzeczenie o rozwiązaniu umowy* [instead of ***orzeczeniem***: NOM, not INSTR, again a wrong selection of case is explained here by the use of punctuation – colon and square brackets]

(28) Art. 2826: the granting of an order for relief to the partnership – ***wydanie*** *nakazu uwolnienia spółki* [instead of ***wydaniem***: NOM not INSTR]

(29) Art. 2826: the reduction of its membership to one person – ***ograniczenie*** *jego członkostwa do jednej osoby* [instead of ***ograniczeniem***: NOM not INSTR]

(30) Art. 2826: the expiration of its term – ***wygaśnięcie*** *jego terminu* [instead of ***wygaśnięciem***: NOM not INSTR]

(31) Art. 2826: the attainment of, or the impossibility of attainment of the object of the partnership – ***osiągnięcie lub niemożność osiągnięcia*** *celu partnerstwa* [instead of ***osiągnięciem lub niemożnością osiągnięcia***: NOM not INSTR].

A general conclusion coming from the analysis of the choice of a proper case in Polish as the TL is that the GT system tended to treat each initial part of a string of words as the subject of a new sentence being a Nominative case, which, in English as the SL, is a supreme rule.

The next category of grammar errors was the wrong selection of gender. It is a case referring to pronouns that can only be agreed upon in the context of a whole sentence, which is why the GT can have trouble picking a proper equivalent. Two errors of this type were found.

Wrong gender (2):

(32) Art. 2812: make him a member – ***go*** *uczynić wspólnikiem* [instead of ***jej*** *uczynić wspólnikiem*, referred to "third person" – *osoba trzecia* which is feminine in Polish, not masculine]

(33) Art. 2839: its use – ***jego*** *użycie* [instead of ***jej*** *użycie,* referred to "name" – *nazwa,* which is feminine in Polish, not masculine].

Another type of contextual error is the one resulting from a wrong interpretation of the word function in a sentence, which causes choosing a wrong part of speech. There were three errors of this type in the sample.

Wrong part of speech (3):

(34) Art. 2826: Termination of a partnership; causes – *Rozwiązanie spółki;* ***powoduje*** [instead of *rozwiązanie spółki;* ***powody*** – the example referred to the verb in the third person singular, not a plural noun]

(35) Art. 2832: paid in preference to – *w pierwszej kolejności* ***niż*** [instead of *w pierwszej kolejności* ***przed*** – here a conjunction *niż* was used, not a preposition *przed*]

(36) Art. 2839: use – ***posługiwać się*** [instead of ***użycie*** – in this example adverb is selected by GT, not a noun].

Interestingly, GT tended to properly interpret the two different meanings of modal *shall*. Seven of eight uses in the sample were translated correctly, and in only one case, a prescriptive sense was used instead of the descriptive required.

Difference between the prescriptive and descriptive meaning of a modal verb *shall* (1):

(37) Art. 2815: a partner shall not participate in losses – *wspólnik* ***nie będzie uczestniczył*** *w stratach* [future tense instead of the present required *wspólnik* ***nie partycypuje/uczestniczy*** *w stratach*].

4.2.3. Syntactic and orthographic errors

Syntactic errors might be produced by GT in cases when, e.g. in the TL a different grammar structure is preferred than in the SL. It concerns, for example, the case when a noun phrase is used instead of an infinitive of purpose. The example below results in the lack of prepositions and the need to rephrase the structure.

Lack of preposition (2):

(38) Art. 2808: obligation of a partner to contribute – *obowiązek wspólnika* ***wnoszenia wkładu*** [instead of *obowiązek* ***wnoszenia wkładu przez*** *wspólnika*]

(39) Art. 2842: do not force the partner in commendam to restore – *nie zmuszą wspólnika z polecenia* ***zwrotu*** [instead of *nie zmuszą wspólnika z polecenia* ***do zwrotu***].

Other syntactic omissions might be direct objects. There was one such case in the sample:

Incomplete syntax (1):

(40) Art. 2839 B: to prevent the use – *zapobiec użyciu* [lack of object *nazwiska* with a transitive verb].

GT also needs aid selecting a proper option when a complex grammar structure is considered. It tends to treat STs linearly. No doubt, then, that such complicated structures as a Saxon Genitive with a longer noun phrase are wrongly parted and rendered as a kind of word-for-word translation. One example is quoted here:

Parting noun phrases (1):

(41) Art. 2844: the partner in commendam's conduct – *wspólnika w postępowaniu komanda* [instead of Saxon Genitive referring to a whole phrase „the partner in commendam", producing an equivalent ***postępowania wspólnika*** *w spółce komandytowej/komandytariusza,* GT only referred it to the last element of the phrase "commendam"].

To discuss to the spelling category, one type of error has to be mentioned. It is the lack of consistency in using capital letters in names of institutions, legal acts, etc.:

Spelling inconsistency (1 error):

(42) Art. 2826: *Kodeksu Upadłościowego* [for „The Bankruptcy Code"]
Art. 2818 A: *Kodeksu upadłościowego* [for „The Bankruptcy Code"].

To summarise, the categories of errors found in this sample support previous analyses performed by the author of this paper (Źrałka, 2019). It has, however, been observed that some kinds of error typical of GT translations have not been numerous in this analysis, e.g.:

– redundant doublets - two errors: (14), (15),
– wrong interpretation of the meaning of *shall* – one case: (37),
– lack of concord between nouns and pronouns – two errors: (32), (33),
– wrong syntax – only four errors: (38), (39), (40), (41),
– inconsistency in spelling – one case: (42).

Compared with other analyses, the correctness of GT renderings remains at around 1.1 % of mistakes per 1.000 words. PE does not seem to be a very challenging task then.

5. Conclusions

In the English-Polish language pair, GT proves a tool that is similar in the level of correctness to other MT systems awarded trust by researchers (e.g. DeepL, according to Domínguez Mora and Ibáñez Moreno 2022). It can be trusted to the extent to which an MT can guarantee accuracy and fluency. How much it can improve the tempo and effectiveness of specialised texts' translation retainig the human-like quality can be a task for future research.

The GT renderings are a good starting point for PE of the texts translated via the system. Principles of more conscious use of GT propositions to serve professional purposes should be:

– checking the correctness and consistency in using terminological equivalents (once leading terms are established, a translator should replace all inconsistent ones with the correct terms),
– checking grammar (declensions, gender, parts of speech used, meanings of *shall*),
– retaining ST elements (original names of, e.g. institutions, legal acts),
– removing ST elements (doublets/triplets, acronyms),
– and checking spelling consistency.

These principles could also serve as a basis for establishing metrics facilitating the evaluation. Such metrics could be based on the criteria including the check of:

– term selection and consistency,
– grammar subtleties: declensions, concord of pronouns, structure of phrases, tenses concerning a modal *shall*,
– redundancies,
– spelling in different types of names.

To conclude, the GT quality assessment based on properly adjusted criteria (metrics) can both raise the level of correctness and accelerate the tempo of translation, especially when it is post-edited by a human translator in the process of applying the tool. It can be stated referring to sound results within TT adequacy and fluency obtained in the research.

Finally, the statement worth stressing is the belief held by Maučec and Donaj (2020: 760), who claim that "Integration of human and machine translation (MT) is a promising workflow for the future. Machine translation will not replace human translation, but it can serve as a tool to increase productivity in the translation process". This belief is confirmed by Zouhar et al. (2021: 10204), who, having reprised research on MT quality in 2016 after its first go with rule-based MT, claim that "better MT systems indeed lead to fewer changes in the sentences in this

industry setting. The relation between system quality and post-editing time is however not straightforward." This fact only confirms the observation made in this research concerning the performance of GT, and the necessity of human presence in the translation process. Pym (2021: 39) shares this belief when he discusses a gradual switch in translators' duties from translation to PE. All in all, it seems unavoidable that the translators' tasks in the future will transmute from being responsible for rendering the ST content into the TL towards controlling the translation quality by incorporating an organised PE procedure.

References:

Bühler, K. 1990. *Theory of Language*. *The Representational Function of Language*. Amsterdam: John Benjamins Publishing.

Charrow, V.R., J.A. Crandall and R.P. Charrow 2015. Characteristics and functions of legal language. In J. Lehrberger and R. Kittredge (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, 175-190. Berlin: De Gruyter.

Domínguez, M.E., and A. Ibáñez Moreno 2022. Comparative analysis of Google Translate and DeepL in Spanish-English literary translation: The case of collocations in Don Quixote. Paper presented at *Positive Impacts of Language Technology: TISLID 22*, UNED, Madrid, Spain, 27 May 2022.

Forcada, M.L. 2010. Machine translation today. In Y. Gambier and L. van Doorslaer (eds.), *Handbook of Translation Studies 1*, 215-223. Amsterdam: John Benjamins Publishing Company.

Han, L., G. JF Jones and A.F. Smeaton 2021. Translation quality assessment: a brief survey on manual and automatic methods. In Y. Bizzoni, E. Teich, C. España-Bonet, J. van Genabith (eds.), *Proceedings for the First Workshop on Modeling Translation: Translatology in the Digital Age*, 15-33. Online: Association for Computional Linguistics. https://aclanthology.org/2021.motra-1.3/

Halliday, M., and A. Kirkwood 1978. *Language as Social Semiotic*. London: Edward Arnold.

Hatim, B. 1998. Translation quality assessment: setting and maintaining a trend. *The Translator* 4 (1): 91-100.

House, J. 1997. *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag.

House, J. 2015. *Translation Quality Assessment*. London: Routledge.

Jansen, K. 2020. Jak ocenić jakość tłumaczenia automatycznego? *Blog POLENG MT*. Accessed May 14, 2024 .https://blog.polengmt.com/blog/jak-ocenic-jakosc-tlumaczenia-automatycznego

Jopek-Bosiacka, A. 2006. *Przekład prawny i sądowy*. Warszawa: Wydawnictwo Naukowe PWN.

Kenny, D. 2018. Machine translation. In P. Rawling and P. Wilson (eds.), *The Routledge Handbook of Translation and Philosophy*, 428-445. London: Routledge.

Maruf, S., F. Saleh and G. Haffari 2019. A survey on document-level machine translation: methods and evaluation. *ACM Computing Surveys* 54 (2): 1-36.

Matulewska, A. 2013. *Legilinguistic Translatology. A Parametric Approach to Legal Translation.* (Vol. 171). Bern: Peter Lang.

Matulewska, A. 2016. Walking on thin ice of translation of terminology in legal settings. *International Journal of Legal Discourse* 1(1): 65-85.

Maučec, M.S., and G. Donaj 2020. Machine translation and the evaluation of its quality. In A. Sadollah and T.S. Sinha (eds.), *Recent Trends in Computational Intelligence*, 143-162. London: IntechOpen. https://www.intechopen.com/chapters/68953pdf.

Mellinkoff, D. 2004. *The Language of the Law*. Eugene, Oregon: Resource Publications.

Młodawska, A. 2012. *Advanced Legal English for Polish Purposes*. Warszawa: Wolters Kluwer Polska.

Nitzke, J., S. Hansen-Schirra and C. Canfora 2019. Risk management and post-editing competence. *The Journal of Specialised Translation* 31: 239-259.

Pym, A., and E. Torres-Simón 2021. Is automation changing the translation profession? *International Journal of the Sociology of Language* 2021 (270): 39-57.

Reiss, K. 1971. *Moglichkeiten und Grenzen der Ubersetzungskritik.* Munchen: Hueber.

Rivera-Trigueros, I. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation* 56(2): 593-619.

Sadollah, A., and T.S. Sinha (eds.). 2020. *Recent Trends in Computational Intelligence*. London: IntechOpen.

Świgońska, R. 2021. Jak tłumaczyć tytuły ustaw na j. polski i na j. obcy? *Blog o tłumaczeniu prawniczym i sądowym.* Accessed June 10, 2023. https://www.tlumaczeniaprawnicze.com.pl/2021/03/25/jak-tlumaczyc-tytuly-polskich-ustaw-na-j-obcy/

Tomás, J., J.À. Mas and F. Casacuberta 2003. A quantitative method for machine translation evaluation. In K. Pastra (ed.), *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are Evaluation Methods, Metrics and Resources Reusable?*, 27-34. Columbus, Ohio: Association for Computational Linguistics. https://aclanthology.org/W03-28pdf.

Veretina, I. 2012. Characteristics and features of legal English vocabulary. *Studia Universitatis Moldaviae (Seria Ştiinţe Umanistice)* 54(4): 103-107.

Zouhar, V., A. Tamchyna, M. Popel and O. Bojar 2021. Neural machine translation quality and post-editing performance. *arXiv preprint arXiv:2109.05016*: 10204-10214.

Źrałka, E. 2019. Google translate evaluation in the context of specialised culture-bound texts. *SKASE Journal of Translation and Interpretation* 12(2): 17-36.