Manual Quality Evaluation and Post-editing in Enhancing the Correctness of MateCat's English-Polish Legal Translations

Author A Edyta Źrałka,

The University of Silesia, Katowice, Poland,Department of Humanities, Faculty of Romance Languages

edyta,zralka@us.edu.pl

Abstract. The reduced credibility of Machine Translation (MT) engines evoked the necessity of quality control and post-editing (PE). Some aiding instruments for evaluation were introduced (Quality Assessment Metrics - QAMs) to facilitate PE and make the process of MT valuable, subject to specified rules. The idea was not only to invent the criteria for metrics and evaluation performance but to record the proofread outcomes and make them repetitive. In such a case, the process of evaluation and PEwould be more effective and lead to better translation quality. A valuable contribution to that need was the creation of MateCat (Machine Translation Enhanced Computer Assisted Translation) tool, a combination of MT engine and CAT tool, enabling a user to translate automatically and edit MT results for better outcomes based on translation memories (TMs) and terminology databases. The outcomes depend on the tool's parallel corpora and databases created by any individual user. Such linguistic data serve to improve the quality of subsequent translations, even if the language is specialised.

The research aims to discover how the quality of legal texts translated via MateCat is enhanced based on quality assessment,PE, andthe creation of new databases.It also seeks to determine how the quality of the tool's performance can be improved andproposes theoretical approaches to Translation Quality Assessment (TQA).

Based on the research, it can be observed that introducing terminological corrections in MateCat translations results in consistently rendered terminology. Grammatical problems tend to be sustained due to fewer chances of contexts' replicability.

Keywords: Translation Quality Assessment, evaluation metrics, post-editing, CAT tools, translation memories

1Research fundamentals and relevance

Recently, the rapid growth of demand for translations in all possible languages in the progressive globalisation resulted in the creation of modern translation tools, some offered online for free, like Google Translate or DeepL. When Machine Translation (MT) tools are considered, they are based on automatic renderings of source texts (STs) into target languages (TLs) using data stored in databases selected by unique algorithms to perform a particular translation. Combining this method with computer-assisted translation (CAT) and applying individually created databases and translation

Proceedings of New Trends in Translation and Technology, pages 261–270
3–6 July, 2024, Varna, Bulgaria. ©NeTTT 2024
https://doi.org/10.26615/issn.2815-4711.2024_022

memories (data ever introduced to the system matched with best translations that are reused in subsequent translations) can produce better results. It is what MateCat offers as open source software, according to the characteristics given by Federico et al.[6], as "a new web-based CAT tool providing translators with a professional work environment, integrating translation memories, terminology bases, concordancers, and machine translation."

In the era of extensive demand for legal translations in the face of societies' migrational lifestyle and internationally conducted businesses, the opportunity offered by CAT tools seems a promising factor. A gradual transformation of translators' duties from traditional human translation into the automatic process followed by post-editing (PE) is observed and predicted as the future standard by Pym [17]. That is why the idea of the research is to perform an automatic translation of selected legal texts in English into Polish based on the MateCat tool, post-edit the raw translations obtained according to a manual metric (namely evaluation criteria formerly worked out for the legal technolect), and, simultaneously, record the results in MateCat for further use and better translation outcomes.The outcome planned to be presented at the conference, to a great extent, is the results of a pilot study carried out to test the research methods under the supervision of Prof. Rozane Rebechi from the Department of Modern Languages at the Federal University of Rio Grande do Sul, Porto Alegre, Brazil, within the funds of *Miniatura -7*project sponsored by *Narodowe Centrum Nauki* (The National Science Centre - a Polish government agency supervised by the Ministry of Education and Science, set up in 2011 to support academic research in Poland).

1.1 Research problems

As fluent in many languages as the MateCat database might be, it still needs much professional evaluation and enrichment. Building a well-equipped database takes time as it builds up along with the systematically performed translations. However, it can be stimulated by anticipating possible wrong renderings in relation to typical errors discovered through QAMs applied in translations by MateCat. The MateCat database can be equipped with new data in a trustful and profitable way when such kinds of research are undertaken (in other words, its engine can be trained) so that it can be a time and effort saver used in a way other CAT software tools function (e.g. SDL Trados Studio), but for free and for everyone – professionals and the common public. The research, when systematically carried out based on a variety of texts, would lead to public access translation tools' reliability. It can also serve as a professional reference for linguists in further research in the field.

There are two main research problems that need elaboration. First, translation quality assessment (TQA) is now one of the most prominent issues in Translation Studies. Since Julian House introduced the backgrounds of her theory in the 1990s

promoting functional equivalence and then developed it in her book *Translation Quality Assessment: Past and Present* published in 2015, her views were analysed, discussed, and adopted about translation in general (e.g. Basil Hatim), or in the field of legal language (e.g. Pietro Ramos), in more particular terms.

Second, the problem of TQA gained particular importance when MT engines occurred, and institutions dealing with the translation business started to replace human translators with automatic translations, followed by TQA and sometimes PE. It was due to the swelling demand for both translations and rapidity to perform them, and, what has to be openly stated, decreasing requirements for correctness. Even if standards are specified for translations, e.g. DIN 2345 and ISO 9000 [3], or ISO 17100 and ISO 18587 [2], first, they concern rather a process than a product, and second, some companies believe that what they need is just the information included in an ST and settle for the MT combined with using automatic metrics for quality control (e.g. BLEU, or METEOR) to feel safe with the percentage of translation correctness that can be accepted, without wasting money on PE. Such a model cannot be accepted for legal texts. It is due to the burden of responsibility for wrongly performed translations and the subsequent consequences to companies and individuals. What scholars need to add is their research to raise the correctness of automatic translations systematically so that they can be reliedupon to as vast an extent as possible. They have much to do with proposing contents of metrics of quality control and criteria for PE to make them both more adjusted to the needs of particular types of texts and ultimate.

1.2 State of the research

As for the linguistic theories of TQA, scholars have referred to both categories considered in evaluation (House, Hatim) and criteria based on which evaluation can be performed (Ramos, Brunette, Mossop, Colina, and Angelelli), the second of which being of more practical use for the research. By definition, categories are some sets of elements that represent "similarities" in any aspect and allow for differentiating them from other collections of similar components. Categories considered when searching for evaluation criteria of any linguistic content, including translation, could reasonably be the following: 1. terminology, 2. grammar features, 3. orthography, and 4. consistency.

"Criteria", on the other hand, adjust to some standards and are subject to judgements. It is then all selected elements that undergo evaluation. Categories represent a broader construct from which one can choose a criterion. Then, manual metrics for MT evaluation could be based on the criteria,including the check of 1. term selection and consistency, 2. grammar subtleties (declensions, concord of pronouns, structure of phrases, tenses concerning prescriptive and descriptive

meanings of a modal *shall*, 3. spelling in different types of names, and 4. redundancies [22].

Apart from just analysing the errors, a post-editor of automatic translation renderings has to be aware of something more than linguistic content. To set the stage for PE, he or she has to consider what House [13] refers to in her theory of TQA based on Halliday's Systemic-Functional Linguistics [10]. According to this theory, texts are always set in some situation, and this situation conditions the texts' functions. Specific situational dimensions characterise the notion of situation: 1. Field refers to the text's subject matter (here, it is law); 2. Tenor: incorporates the people involved in the communication and the relationships between them (level of formality – here, it is formal); 3. Mode: refers to the form of language, spoken or written, used in the interaction (here, it is a written language). It is then necessary to start translation evaluation by realising such basic features as register, style, and the general function of a translated text according to the defined needs.

Considering that, House gathers and groups views on the quality of translation into several categories based on equivalence and the functional treatment of STs and target texts (TTs) in translation. Instead of traditionally understood equivalence, the author allows the functionalistic approach to the TQA as one of the categories is derived not only from Halliday but also from *Skopos* theory of Katherine Reiss and Hans Vermeer of the 1970s, later developed by Christiane Nord. She even claims that "The skopos or purpose is the most important factor in translation, the original text being downgraded to a mere 'offer of information' and the translator often seen as a type of 'co-author'." [13].

Hatim [11] observes that, in House's theory, a function of translation is considered a mixture of language functions and text functions, based on the views of Reiss [19], Bühler [4] and Halliday [9]. Hatim [11] states that a full reproduction of specific ST functions is not possible in translation because, according to House [12], "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 [11] concludes that if "situationality" existing between the two texts cannot be fully reproduced, "a second-level function" combining ways of perception by ST and TT readers is required to be reached by a translation. Such a function "must hold not only for the contemporary target language readers but also for their counterparts in the source culture" [12].

Hatim's and House's observations are crucial to the attitude towards legal texts' TQA. A translator and post-editor have to be aware of the functions and the specificity of legal texts in both languages, comprising their linguistic features (terminology, phraseological structures, grammar, syntax, punctuation, etc.) and cultural dependence. These are all elements that imply a source-text-oriented translation, with the TT receiver kept in mind.

Ramos [18] and the authors he mentioned by him refer to the criteria of translation evaluation incorporated into legal translations. The most systematic of them, while sharing the features of other classifications, seems to be the one by Brunette [3]. The author differentiates the following criteria of TQA: 1. Logic (depending on coherence and cohesion), directed to the target audience, not the ST; 2. Purpose (effect and intention); 3. Context, defined as "Non-linguistic circumstances surrounding the production of the discourse to be assessed."; and 5. Language norm (rules and conventions of the language). Through addressing the purpose and context, she refers to the criteria introduced also by House and others. Language rules and logic remain standard linguistic components of assessing any text. The classifications by Mossop [15], Colina [5], and Angelelli [1] mainly refer to typical comparative elements of source and target texts. Mossop additionally includes layout and organisation among the evaluative criteria. Angelelli sees "translation skill" as a criterion of evaluation in translation, which seems an element encapsulating all other criteria.

For this research, the criteria introduced by Źrałka [22], mentioned earlier, will be incorporated into the analysis on the condition that the pilot study performed in March-August 2024 does not reveal additional needs. The principle of assessing the quality of translations and PE will be orientation to the ST, taking into account the needs of the target recipient.

When assessing MT is taken into consideration, Forcada [7], Maučec and Donaj [16], and others differentiate manual ("human") MT evaluation and automatic MT evaluation, "an algorithm that can be coded into a programme and run by a computer that calculates the evaluation score, which tells the user how good a translation is" [16]. The quality of MT output is measured as a final product, or PE is incorporated to enhance the quality for more official text use. Computer programmes used to assess translation quality operate on language rules, and often, an existing human translation is used for comparisons with the aim to "try to measure how close each raw machine-translated sentence is to one or more reference human translations" [7]. In the case of a human assessment, the professionals use their linguistic and cultural knowledge in one or both languages. Han et al. [8] propose the criteria for such evaluation, divided into traditional (e.g., fidelity, fluency, adequacy, comprehension) and advanced ones (including PE).

Maučec and Donaj [16] point out that MT quality is usually evaluated within adequacy and fluency on a five-point scale. The characteristics of particular levels of the classification are the following:

- Adequacy: all meaning – worth five points, most meaning – four points, much meaning – three points, little meaning – two points, none of the meaning – one point;

- Fluency: flawless language – five points, good language – four points, non-native language – three points, disfluent language – two points, incomprehensible language – one point [16].

For the theoretical considerations of this research, a five-point scale in the criterion of Adequacy and Fluency will be applied as well.

1.3Justification to undertake the research and general research goal

The research is innovative due to the following reasons justifying its performance: 1. Linguistic theories are more practically used in the evaluation of MT products, which should result in a broader and more critical attitude to the assessment of MTs and their PE; 2. Evaluation of legal texts' translations based on the ideas of different scholars refers to the translations performed by CAT tools, which is a way to enhance the quality and pace of the translation process of MateCat in practical terms too; 3. The use of translation evaluation metrics with the incorporation of PE will presumably raise the level of correctness of MateCat translations by enriching the database instead of using the metrics just as tools for their own sake to know how good or bad the translations are (the record of the data will serve future academic purposes); 4. Creating a database taking into account common people's needs concerning matters and problems they could encounter is aimed instead of enriching the existing EU legal terminology databases developed for institutional texts and serving mostly professional translators (ordinary people need a tool to trust when they look for translations of popular legal documents – court decrees, contracts, fiscal and civil documents, etc.); 5. PE teaching ideas within legal texts will occur and serve to fill the gap in the modern education of future translators.

Specifying the research goal, itseeks to gather possible data within a legal language constituting sources of errors in MateCat translations and post-edit them while creating an enlarged database. First of all, it needs the knowledge of a professional to deal with all linguistic and terminological aspects needed to enrich the MateCat existing database. There is still a need to evaluate and post-edit the MateCat translations in the field of law in the pair of languages English-Polish. The research aims to raise the quality of MateCat legal translations within the scope of languages chosen to the highest standards possible through post-editing MateCat renderings of a text collection selected and creating a richer professional database referring to the errors encountered. The idea is to use texts, which are legal documents popular among ordinary people, to enable them to understand what is included in the content that concerns their matters. For the pilot study, different types of employment contracts were selected (two open-ended employment contracts, two fixed-term contracts, and two mandate contracts, comprising roughly 15,000 words altogether).The future data will be gathered based on more diversified documents (other contracts, court decrees, civil procedures' documentation, inheritance documentation, donation documentation,

ownership, mortgage procedures) andcan be used to develop not only MateCat database but also legal translation software for instructional and practical purposes (MT or CAT programmes). The possible modification in the system of translators' training that gradually evolves into the abilities of PE rather than mere translation is, by the way, anotherreason for carrying out the research. Based on the research results, some suggestions will be given within post-editing didactics.

The outcome of the research will be a well-designed database of legal terminology, phraseology, syntactic issues and orthography in the language pair English-Polish and then Romance languages, with the perspective of raised quality of translations performed automatically but post-edited by trained translators according to the challenges observed in the research.

1.4Research questions and hypotheses

The fundamental research question posed at the project's grassroots is the following: 1.Will PE lead to systematically enhancing the MateCat performance? It provokes some other questions, such as: 2. How much can the quality of MateCat translations be improved by enriching the tool's database within legal texts?, and, the one underlying the above-mentioned, 3. Is MateCat performance an auspicious basis for carrying out academic research?

The hypothesis to be tested is that creating the enriched database should result in a higher quality of MateCat legal translations and higher efficiency.

1.5. Methodology

Our main concern in the research will be the analysis of a translation product –the first renderings of corpus texts performed by MateCat from English to Polish. The main method will be corpus research and a comparative study of ST's and TTs within the two main criteria of the manual metric used: the Accuracy criterion (ST and TT comparison), and the Fluency criterion (TT quality exclusively). Discourse analysis and critical discourse analysis will be incorporated within the assessment of Fluency. The analysis will incorporate qualitative and quantitative methods.

The research methodology incorporates post-editing the MateCat's translations according to the manual metric comprising the evaluation of:1. term selection and consistency (including the level of formality); 2. grammar structure of phrases focused on: a. case, b. number, c. gender, d. part of speech, e. meaning of "shall", f. additions, g. omissions, h. syntax, i. word-for-word translation),j. form of a verb/ or any grammar structure differing from the ST;3. redundancies in doublets and triplets; 4. translation of proper names (only SL version, or only TL); 5. spelling in names of legal action participants and legal instruments (general use of capital letters in the TL wherever they are used in the SL). 6. Stylistic mistakes in the TL.

LF Aligner programme is used to align and juxtapose the texts analysed. Simultaneously, corrections are made in the MateCat tool to improve the quality. As a source of terminological reference, the AntConc programme is used, the same as the IATE terminology database and possible references in legal dictionaries. The outcomes assessment based on BLEU is going to be performed at the beginning and the end of the study, after introducing the corrections, and then the two TQA results are compared. Methodological conclusions concerning how to evaluate and post-edit are made, and the evaluation metric is being revised systematically.

2 Research plan

The detailed research plan includes realising three parallel goals: 1. to perform the research material evaluation based on two kinds of metrics – automatic and manual, add PE to the evaluation and classify errors to prepare entries to the MateCatdatabase (initial stage of the research performed within two first months); 2. to create the database, incorporate it in the MateCat system and devise a separate version of it for further academic purposes (during the PE and afterwards up to two additional months after the first stage); 3. To answer the research questions and test the hypothesis posed in two months after stages 1. and 2., which will lead to the final quality assessment of the post-edited texts and conclusions. Collateral research concerning the needs and schemes of a course in post-editing automatic translations of specialised legal texts can be done. Altogether, the research is planned to be conducted within six months, started in March 2024 it is bound to offer the final results in August 2024..

All findings are expected to raise both the practical MateCat's legal translation quality and the theoretical awareness of contemporary needs in translation theoryand practice, together with new approaches proposed.

3 Research results' relevance for the quality enhancement of legal texts'automatic translation

The extended version of the research is essential for upgrading the quality of automatic translations of specialised legal texts in languages that are not yet leading regarding translation combinations (the juxtaposition of Polish with English, but, ultimately, Romance languages). At least, consistently rendered terminology will be the result of PE. Grammatical problems will be reduced in repeatable contexts.

As carried out, the research will fructify with a professional database within legal translations. Some new patterns will be retrieved from the research used to elaborate a teaching programme for university courses in PE.

Disclosure of Interests. Author A has received research grants from*Narodowe Centrum Nauki (NCN)*, Poland.

References

1. Angelelli, C. V.: Using a rubric to assess translation ability. In: Testing and assessment in translation and interpreting studies,pp. 13-49. (2009)
2. Biel, Ł.: Postedycja tłumaczeń maszynowych. Lingua Legis, 1(29), pp. 11-34 (2021)
3. Brunette, L.: Towards a terminology for translation quality assessment: A comparison of TQA practices. The Translator 6(2), pp. 169–182 (2000)
4. Bühler, K.: Theory of language. The representational function of language. John Benjamins Publishing Company, Amsterdam (1990)
5. Colina, S.: Translation quality evaluation: Empirical evidence for a functionalist approach. The translator 14(1), pp. 97-134. (2008)
6. Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Germann, U.: The MateCat tool. In: COLING (Demos), pp. 129-132. (2014)
7. Forcada, M. L.: Machine translation today. In: Handbookof translation studies, 1,, pp. 215-223. John Benjamins Publishing Conpany, Amsterdam(2010)
8. Han, L., G., JF Jones, A. F. Smeaton: Translation quality assessment: A brief survey on manual and automatic methods. In: arXiv preprint arXiv:2105.03311 (2021)
9. Halliday, M. A. K.: Language as Social Semiotic. Edward Arnold, London (1978)
10. Halliday, M.A.K.: An Introduction to Functional Grammar. Edward Arnold, London (1985)
11. Hatim, B.: Translation quality assessment: Setting and maintaining a trend. The Translator 4 (1), pp. 91-100 (1998)
12. House, J.: Translation Quality Assessment: A Model Re-visited. Narr, Tübingen(1997)
13. House, J.: Translation Quality Assessment: Past and Present. Routledge, New York (2015)
14. Nord, Ch.:Translating as a Purposeful Activity. Functionalist Approaches Explained. St. Jerome, Manchester (1997)
15. Mossop, B.: Revising and editing for translators, 2nd edn. St. Jerome Publishing, Manchester (2007)
16. Maučec, M. S., Donaj, G.: Machine translation and the evaluation of its quality. In: Recent Trends in Computational Intelligence, pp. 143-162. IntechOpen (2020)
17. Pym, A., Torres-Simón, E.: Is automation changing the translation profession?International Journal of the Sociology of Language 2021(270), pp. 39-57, (2021)
18. Ramos, P. F.: Quality assurance in legal translation: Evaluating process, competence and product in the pursuit of adequacy. International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 28, pp. 11-30 (2015)
19. Reiss, K.: Moglichkeiten und Grenzen der Ubersetzungskritik. Hueber, Munchen (1971)

20. Reiss, K., Vermeer, H. J., Nord, C., Dudenhöfer, M.: Towards a General Theory of Translational Action: Skopos Theory Explained. 1st edn. Routledge Taylor & Francis Group, London (2015)
21. Vermeer, H. J.: Ein Rahmen für eine allgemeine Translationstheorie. LebendeSprachen,*23*(3), pp. 99-102 (1978)
22. Źrałka, E.: Quality assessment and post-editing of Google Translate output in specialised legal translation. In: Approaches on Machine Translation: Quality Assessment, Acceptance and Language Trends, Translating and Translanguaging in Multilingual Contexts. Amsterdam, John Benjamins (unpublished – publication planned 2024)