# What Makes Machine-Translated Poetry Look Bad? A Human Error Classification Analysis

**Ivan Dunđer, Sanja Seljan, Marko Pavlovski**
Faculty of Humanities and Social Sciences, University of Zagreb
Department of Information and Communication Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
ivandunder@gmail.com, sanja.seljan@ffzg.hr, mpavlovs1987@gmail.com

***Abstract****. Human translation of literary works is a profession as ancient as the appearance of writing with a very long tradition. However, with the ever-growing improvements of technologies in the field of machine translation, the possibilities of adopting machine-translated literary texts are in constant progress. Nevertheless, human evaluation of machine-translated text is a necessity. The aim of this paper is to examine what makes machine translations appear bad in terms of error types that occur within such translations. Therefore, a human error classification analysis on a machine-translated text corpus in the domain of poetry is conducted. This study could be used for improving the methodology for machine translation evaluation of literary texts.*

**Keywords.** machine translation, quality evaluation, human analysis, error classification, natural language processing, information and communication sciences

## 1 Introduction and Motivation

The importance of high-quality translations, of oftentimes complex literary works of art, is a tradition intertwined with the beginnings of literacy. In the modern age automatic machine translations are becoming increasingly actual and accepted due to numerous factors, which include among others, their cost and time effectiveness, and these characteristics are applicable to literary texts as well.

However, empirical studies have shown machine-translated texts to be of lower quality than human translations, and to require additional post-editing, i.e. manual error correction after machine translation. A task of substantial importance in this post-editing process is conducting a human error classification analysis, in which several error types, also known as error classes or error categories, must be identified and examined.

The aim of this paper is to analyse errors in the machine-translated text, more specifically, on an example of a corpus containing work of art in the domain of poetry, and with special emphasis on error types that mostly contribute to the worse public perception of machine translation quality in general.

The error type analysis in this paper is based on the Dynamic Framework Quality (DQF) error classification methodology. It is used to verify earlier findings of the comparison of machine translation system performances when translating literature. In this way, the overall effectiveness of applying DQF for machine translation quality evaluation in the domain of poetry is also investigated.

In this paper this is achieved by conducting an evaluation experiment on a data set consisting of a collection of poems written by a notable Croatian poet, Delimir Rešicki. Here the authors focus on the evaluation and impact of the most dominant error types, which are identified as part of the human evaluation task. Such an approach can be understood as an important part within the translation quality assurance (QA) process, since machine translation results should always be addressed differently, depending on the level of sufficient translation quality, user-specific requirements, translation scenarios etc.

The motivation for this research is accentuated in the series of actions that involve evaluating and differentiating the dominant error types that are found in the process of machine translation error evaluation, which is conducted on the part of a human evaluator. This research will show the various problems a human evaluator encountered when assessing the quality of machine-translated works of art, in this case verses from the domain of poetry.

The authors presume that some of the defined error types are more prevalent when compared to others during this machine translation quality assessment, due to the very nature of poetry and literary texts in general.

This evaluation methodology could be applied in the context of education, first and foremost in higher education with the aim of teaching students the intricacies of e.g. the style or other characteristics of a literary work of art that is to be analysed. Students from the technical and non-technical sciences could be demonstrated the impact of different error types on the comprehension of translated texts.

Furthermore, the research of error classification categories applied to the aforementioned type of text could give a deeper insight into the inner workings of automatic machine translation technology, and its challenges and drawbacks.

Also, by concentrating on particular error types, one could implement necessary machine translation system enhancements and upgrades, employ domain adaptation techniques, perform additional fine-tuning etc., which are all required for the application of this methodology in an optimised and cost-effective manner.

## 2 Related Work

While automatic quality evaluation metrics are important and inevitable tools for rapid development of machine translation systems, they are only a substitution for human assessment of machine translation quality (Popović, 2020).

This human assessment tends to be automated through approaches that allow automatic classification of machine translation errors, but they still cannot provide the same detailed granularity as manual, i.e. human error classification. However, they enable valuable estimation of machine translation errors and better understanding of the analysed machine translation system in a short time and on a large scale (Popović & Vilar, 2019).

Evaluation of machine translation quality is necessarily a subjective process because it involves human judgments. Classification of errors can allow these judgments to be made in a more consistent and systematic manner (Flanagan, 1994).

New approaches to formal machine translation assessment and the corresponding tools are in development. Graduates who encounter them early in the classroom will be better prepared to allocate their time, self-assess their output, and revise that of others. Thus equipped, professional translators could assert their standing by using the relevant quality assurance procedures (García, 2014).

The company TAUS from The Netherlands launched the Dynamic Quality Framework (DQF) in 2011, and since then TAUS has applied different quality criteria and methods for machine translation evaluation – such as adequacy, fluency, productivity testing and machine translation ranking.

TAUS has also provided the means to compare results to previous projects and to minimise subjectivity by using a standardised evaluation workflow. However, benchmarking to satisfy the user needs and to provide the right quality level for each user is still work in progress. In order to develop and improve translation quality, one needs to measure quality constantly and consistently (Görög, 2014a).

Multidimensional Quality Metrics (MQM) can provide thorough insights about diverse translation issues and errors on different levels of granularity, up to the word or phrase level as input for systematic approaches to overcome various translation quality barriers. Like the common practice of post-editing, it requires tedious manual work that will hopefully become less labour-intensive in the future through, at least partial, automation (Burchardt & Lommel, 2014).

One author accentuates that in Fall of 2014 the Dutch company TAUS and the German Research Center for Artificial Intelligence (Deutsches Forschungszentrum für Künstliche Intelligenz, DFKI) started the harmonisation of Dynamic Quality Framework (DQF) and Multidimensional Quality Metrics (MQM) with the aim of bridging the gap between the definitions and specifications of the two models (Görög, 2014b). The idea was to integrate the reference frameworks into a combined typology (Rivera-Trigueros, 2021).

One recent study showed that more than half of the analysed papers on machine translation included error classification and analysis, a fundamental aspect for identifying flaws and improving the performance of machine translation systems and, in addition, over half of these works carried out this analysis employing standardisation frameworks such as DQF and MQM (Rivera-Trigueros, 2021).

Another research listed crucial error typologies and industry standards mentioning DQF and MQM in machine translation evaluation, and highlighted their special role in the post-editing process (Nunziatini & Marg, 2020).

A paper applied both DQF and MQM for error classification in a comparative evaluation of different machine translation systems – a statistical machine translation system, a generic neural machine translation system and a tailored neural machine translation system (Stasimioti et al., 2020).

A case study has recently been conducted on user expectations towards machine translations using MQM and DQF. According to the paper, user expectations are crucial in translation, including processes in machine translation since they may help predict user interventions, such as pre- and post-editing. This study argues that users' past experiences, expectations and (dis)confirmation of expectations steer human evaluation of machine translation (Heinisch & Lušicky, 2019).

When it comes to the field of translation studies, translation quality cedes to be perceived as a monolithic and single construct with a multidimensional approach and with different price points related to how the translation is produced (Jiménez-Crespo, 2017).

One research concludes that logic would indicate that, as machine translation improves, the post-editing task will become "easier" for translators, i.e. translators would be able to process more words in less time and with a lower technical and cognitive effort. However, this will vary depending on the machine translation system, the language combination, and the domain. Statistical machine translation models have been customised and long implemented in the language industry, with adjustments for specific workflows taking into consideration customers, content and language combinations etc. (Guerberof Arenas, 2019).

According to one author, post-editing is only now becoming a common task for a translator, and research into many different aspects of post-editing such as cognitive effort, editing time, or whether it is similar to translation or not, has skyrocketed (Dede, 2019).

Although many researchers have demonstrated that post-editing differs from translation in many ways (O'Brien, 2002; Rico & Torrejón, 2012), others have suggested that the features of a post-editing task depend on many factors, such as the text type, the machine translation system used, the language pair, and the competence of the translator and post-editor (Aranberri, 2017).

One paper denotes that it is important to know that all error typologies are based on observation, and that all successful evaluation methods were developed under unique circumstances (Lengyel, 2011).

Referencing the "shift towards the use of more explicit error classification and analysis", another paper affirms that error profiles are typically *ad hoc* categorisations and specific to individual machine translation research projects, thus limiting their general usability for research or comparability with human translation results (Lommel et al., 2014).

Neural machine translation has been utilised for literature translation and many challenges and potential improvements have been identified in a recent study (Matusov, 2019). The paper suggests that longer sentences are translated well in terms of syntactic structure, so that the necessary post-editing is often local and minor. Furthermore, up to 30% of evaluated segments, mostly short sentences, were considered acceptable and might only require proof-reading by a monolingual editor of the target language.

When it comes to applying human machine translation error classification, extensive research has also been done for the Croatian language (Dunđer, 2020; Dunđer, 2015). Other automatic and human quality evaluation trials including the Croatian language have also been presented in several papers (Seljan & Dunđer, 2014; Seljan et al., 2015; Seljan & Dunđer, 2015a; Seljan & Dunđer, 2015b).

# 3 Research

This research represents a continuation of a series of studies that deal with applying and evaluating of automatic machine translation of poetry-related corpora including the Croatian language.

The data set used in this research does also derive from two earlier experiments that considered quantitative and qualitative machine translation quality aspects (Dunđer et al., 2020; Seljan et al., 2020).

The data set that is used in all of the experiments contains 14 poems written originally in Croatian and in short verses by a notable contemporary poet – Delimir Rešicki – and the translations of his poems in German that were conducted by Klaus Detlef Olof and Alida Bremer, two professional literary translators. All poems were crawled from the internet, more specifically, from an international web platform for publishing contemporary poetry, which is called "Lyrikline" (https://www.lyrikline.org/en/poems/rosa-mystica-13248). In total, the corpus consisted of 532 sentences/segments in Croatian, and 532 in German.

The term "segments" refers to text chunks, i.e. text lines that do not end with a sentence delimiter. The data set was pre-processed and prepared for machine translation mostly by using Python and Perl. This step included various tasks, such as file format conversion, adjusting proper character encoding, stripping of text formatting and styles, data cleaning with regular expressions, tokenisation, converting to parallel corpus format, human error inspection etc.

Once the data set was ready, the authors decided to use it as the input for two freely available online machine translation systems, i.e. web services for generating machine translations for the Croatian-German and German-Croatian language pairs, i.e. for both translation directions: Google Translate (https://translate.google.com) and Yandex.Translate (https://translate.yandex.com/).

Google Translate serves as a benchmark today and is able to produce high-quality machine translations. Both Google Translate and Yandex.Translate support the Croatian language and perform relatively well in terms of automatic metrics and human quality evaluation (Dunđer, 2015).

Afterwards a quantitative automatic machine translation quality evaluation was conducted by using four different automatic metrics on the corpus level: BLEU, METEOR, RIBES and CharacTER.

After all the automatic metrics were calculated, the authors decided to analyse the qualitative aspects of the resulting machine translations with the help of three human evaluators who are native speakers of the Croatian language, and who were told to rate machine translations with respect to the original reference (source) texts and by considering two quality criteria – adequacy and fluency. Human evaluations were made only for the German-Croatian language direction, but on the entire machine-translated data set. In other words, evaluators had to grade all of the 532 machine translations per each machine translation system. The idea of this study was to manually explore the performances of the machine translation systems in the literature domain.

According to the findings of the first and second study, Google Translate achieved better results when compared to Yandex.Translate in terms of automatic machine translation quality metrics and human evaluation with focus on adequacy and fluency, which was an important departure point for the third study, as they indicate that DQF might also identify Google

Translate as the system that is more suitable for machine translation in the literature domain.

Now, when it comes to the third study, i.e. this particular research, in order to conduct an error typology analysis and to quantify the different types of machine translation errors, out of the 532 German-Croatian machine translations, 100 were extracted randomly per each system. Those 100 sentences/segments were same for both Google Translate and Yandex.Translate.

Then, one skilled human evaluator with formal education in language and literature studies was presented the 100 extracted reference, i.e. source sentences/segments and the corresponding 100 machine translations generated by both machine translation systems in a tabular format.

The applied machine translation error classification methodology has been taken from an analytic metric created by TAUS, which is called Dynamic Quality Framework (DQF). A detailed presentation of the different error types (error classes or error categories) along with the corresponding descriptions are given in Table 1.

**Table 1.** Machine translation error classification.

| Error class | Description |
|---|---|
| Accuracy | Incorrect interpretation of source text – mistranslation |
| | Incorrect/misunderstanding of technical concept |
| | Ambiguous translation |
| | Omission (essential element in the source text missing in the translation) |
| | Addition (unnecessary elements in the translation not originally present in the source text) |
| | 100% match not well translated or not appropriate for context |
| | Untranslated text |
| Language | Grammar – syntax: non-compliance with target language rules |
| | Punctuation: non-compliance with target language rules |
| | Spelling: errors, accents, capital letters |
| Terminology | Non-compliance with company terminology |
| | Non-compliance with 3rd party or product/application terminology |
| | Inconsistent |
| Style | Non-compliance with company style guides |
| | Inconsistent with other reference material |
| | Inconsistent within text |
| | Literal translation |
| | Awkward syntax |
| | Unidiomatic use of target language |
| | Tone |
| | Ambiguous translation |
| Country standards | Dates |
| | Units of measurement |
| | Currency |
| | Delimiters |
| | Addresses |
| | Phone numbers |
| | Zip codes |
| | Shortcut keys |
| | Cultural references |
| | Tone |
| | … |

Table 1 shows that there are five main error classes in the Dynamic Quality Framework (DQF): accuracy, language, terminology, style and country standards, whereas each error class contains more fine-grained subelements.

Beforehand, the authors presumed that some of the error classification elements within the defined error categories would be more present than others, since this quality assessment method was applied on a data set from the domain of poetry. For example, non-compliance with company terminology from the terminology category, and country standards should be virtually inexistent, whereas errors such as incorrect interpretation of source text – mistranslation, omissions, additions, awkward syntax, grammar, spelling and punctuation errors should be present.

## 4 Results and Discussion

The results of the human evaluation with regard to the defined error classes, as presented in Table 2, show that in total more than 620 errors were found in the machine translations generated by Google Translate and Yandex.Translate.

**Table 2.** Results of human evaluation with regard to error classes.

| **Error class** | **Machine translation system** | |
|---|---|---|
| | Google Translate | Yandex.Translate |
| Accuracy | 176<br>(average: 2.12) | 202<br>(average: 2.43) |
| Language | 74<br>(average: 0.90) | 77<br>(average: 0.93) |
| Terminology | 6<br>(average: 0.07) | 6<br>(average: 0.07) |
| Style | 39<br>(average: 0.47) | 42<br>(average: 0.51) |

| | | |
|---|---|---|
| Country standards | 0<br>(average: 0.00) | 0<br>(average: 0.00) |
| **Total** | 295 | 327 |

295 errors were identified in machine translations generated by the Google Translate system, whereas 32 errors more were found in the output of the Yandex.Translate system. Out of the almost 300 detected errors made by Google Translate 176 were accuracy errors (ca. 60%), followed by language errors, style errors and terminology errors (only ca. 2%). Similarly, 62% out of all Yandex.Translate errors were also accuracy errors, followed by language, style and terminology errors. Country standard errors were, as expected, not found in either machine translation outputs. In average, more than 2 accuracy errors appeared in every sentence/segment, and less than 1 language error. This might be caused due to the difficulties with processing Croatian, a relatively complex and morphologically rich language.

Although it is evident that accuracy errors were mostly represented in both Google Translate and Yandex.Translate output, the Yandex.Translate system still generated more errors in every category (except terminology).

The authors would like to note that for 17 sentences/segments (per machine translation system output) the human evaluator did not count the different errors, as there were significant and obvious mismatches in alignments between source text and generated machine translations, so there was no point to additionally penalise the machine translation quality. Some of the segments (verses) were just not correctly aligned due to the artistic license of professional human translations. This means that some of the translations were relatively freely translated, while some other, although correct, just appeared in later verses and therefore caused misalignments. However, this was expected to some degree.

This research has shown that, when such a quality assessment method is applied on a literary text, some error classification categories are more frequent. The aforementioned error categories that were present in the poetry-related machine translations were most dominantly accuracy, language and style. Terminology errors occurred only a few times, whereas the evaluator did not find any country standards errors.

On the other hand, the evaluator encountered problems differentiating several defining elements in the separate error classification categories, such as the incorrect interpretation of source text – mistranslation, which is categorised under the accuracy category, in comparison to syntax or the non-compliance with target language rules (grammar in general), which is categorised under the language category.

For example, the original sentence/segment “tko će već jednom toj curici” was translated by Google Translate as “tko će napokon ovu djevojku” and by Yandex.Translate as “tko će konačno postati ova djevojka”, and here the evaluator decided to categorise the omission of the word “već” as a language error, i.e. an error of syntax (original: “tko će već jednom”, machine translation by Google Translate: “tko će napokon”, Yandex.Translate: “tko će konačno”), although syntax is broadly defined in TAUS’ DQF as “Grammar – syntax: non-compliance with target language rules”, as opposed to omission, a definition within the accuracy category, which is described as an essential element in the source text which is missing in the translation.

Also, another difficulty was the differentiation between omission and unnecessary elements in the translation not originally present in the source text (addition) when juxtaposed with the omission and/or addition of accents, capital letters and errors in general (spelling). For example, the original sentence/segment “S neba , na polja u proljeće” was translated by Google Translate as “od neba do polja .” and by Yandex.Translate as “od neba do polja .” where the language error is the omission of punctuation, i.e. the missing comma (“,”), which is identified by the evaluator as a language error and not a punctuation error because the error was the omission of punctuation in the translations.

The most challenging task according to the evaluator was evaluating style, which as an error classification category is defined as: non-compliance with company style guides, inconsistent with other reference material, inconsistent within text, literal translation, awkward syntax, unidiomatic use of target language, tone and ambiguous translation. These definitions do not precisely match style as a literary device, which is simply defined as the way an author writes. When applied during human classification of machine translation errors on the example of a literary text, this category seemed somewhat abstract and vague to the evaluator.

Additionally, “ambiguous translation” is a definition shared by both the accuracy and style categories, which caused some overlap and subjective decision-making on the appropriate category to which a machine translation error belongs. For example, the sentence/segment “othodi u umoran mi san djevojče” was translated identically by both machine translation systems – Google Translate and Yandex.Translate – as “dolazi djevojka s najljepšim od svih imena”, and here the evaluator decided to count:

- 5 accuracy errors: “othodi” – “dolazi”: 100% match not well translated or not appropriate for context; “u”, “umoran”, “mi”, “san”: untranslated text
- and one style error (“djevojče” – “djevojka”).

Nonetheless, even though some error categories did overlap, and their assignment was left to the evaluator’s subjective judgment, this decision-making process was consistent, i.e. the same error types were consistently assigned to the same subjectively established error category. For example, the evaluator decided to categorise ambiguous translation errors as accuracy and not style, and every ambiguous translation error was assigned to the accuracy category.

This was a serious hindrance to the evaluation process. Anyway, once the preferred category for ambiguous errors was established, carrying out the evaluation was consistent.

However, this decision influenced the results of this research by eventually showing that accuracy as a machine translation error category had the most notable impact among the error classification categories within TAUS' Dynamic Quality Framework (DQF).

In conclusion, because of the aforementioned decision, accuracy had the most visible impact as an error classification category with an observed error count for Google Translate of 176, and an observed error count for Yandex.Translate of 202. The second most frequent error category was language, with an observed error count of 74 for Google Translate and an observed error count of 77 for Yandex.Translate. The third category by error count was style, with an observed error count of 39 and 42 for Google Translate, and Yandex.Translate respectively. The fourth category was terminology (6 errors per machine translation system), whereas the fifth and last category by error count was "country standards", which had no observed errors (error count of 0) for both Google Translate and Yandex.Translate.

The results for this particular error type show that it is not a suitable error classification category in this specific literary text based evaluation environment, and should possibly be omitted in future poetry-related machine translation research, since other error categories such as accuracy or style could cover "country standards" errors that in theory might occur but which are less likely.

Overall, the results show that Google Translate had a lesser error count in comparison to Yandex.Translate. Specifically, on the matter of machine translation error classification of style, the TAUS Dynamic Quality Framework (DQF) was suitable to a certain point, while some of the categories were not appropriate for the literary environment. This leads to the conclusion that in future research a more adequate definition of error categories relevant to the domain of literary texts should be applied.

# 5 Suggestions for Future Research

The results of the human evaluation with regard to the error classes presented in this paper show that the DQF error classification methodology might not be completely appropriate for evaluating machine translations of literature, since the evaluator in this research encountered problems of differentiating between different types of errors, e.g. difference between mistranslation and syntax, omission and unnecessary elements; style definition; ambiguous translation, leaving the evaluator to his/her subjective judgement. This emphasises once more the difficulty of evaluating machine translations of literary works. Hence, modifying the DQF error classification methodology for evaluating machine translations of literary works and examining it according to evaluators feedback should be investigated more closely in future research.

More extensive human evaluation on a larger data set should be done, also for the Croatian-German language direction, and possibly in combination with a more detailed evaluation framework, such as the renowned Multidimensional Quality Metrics (MQM), which is expanded in such a way that it allows for encompassing user-specific error categories to be applied in specific environments.

A specially built crowdsourcing platform that makes use of the gamification principle could be upgraded, i.e. expanded so that users with background in natural language processing could be motivated to perform machine translation error classification (Jaworski et al., 2017). The poetry-related data set could be annotated statistically or linguistically to facilitate the human machine translation quality assessment (Seljan et al., 2013).

Utilising word embeddings in form of word vectors could reveal interesting concept-related and semantic relationships between different constituent parts (Dunđer & Pavlovski, 2019b). A sentiment analysis could detect the overall affective states in the poetry-related data set as well (Dunđer & Pavlovski, 2019a).

Frequency analyses of word occurrences and their corresponding distributions (Pavlovski & Dunđer, 2018), key word extraction (Seljan et al., 2014; Dunđer et al., 2015), concordances (Jaworski et al., 2021; Dunđer & Pavlovski, 2018) and context analyses (Dunđer et al., 2020) might also expose interesting author-specific writing styles and literary elements.

A usability analysis of combining speech synthesis (Dunđer, 2013) and machine-translated texts from the domain of poetry, subsequent quality assessments (Seljan & Dunđer, 2013) and domain-specific evaluation trials (Dunđer et al., 2013) could also explore the possibilities of uniting diverse disruptive technologies and the humanities (in this case, data sets from the domain of poetry), for maximising the advantages of the various natural language processing methodologies and tools for Croatian and for various purposes.

# 6 Conclusion

This research is the third study conducted by the authors on machine translation quality when applied to the domain of poetry and including the Croatian and German language. The first study dealt with utilising automatic quality metrics, the second with the human evaluation of adequacy and fluency in machine-translated poetry, whereas the third with the impact of specific machine translation errors on the quality perception, as thoroughly presented and discussed in this paper.

The results of this research show that the most prominent error category in total was accuracy, which amounted to two thirds of all machine translation errors identified by a human evaluator, followed by the language category. These two error categories had the most effect on the machine translation quality.

Regarding the TAUS Dynamic Quality Framework (DQF) error classification methodology, the authors have demonstrated that for analysing literary works of art, such as poetry, human evaluators should be conducting evaluation by employing user-specific and customisable evaluation frameworks.

Finally, the authors conclude that after conducting and completing three separate studies on different aspects of machine translation quality, out of the two selected machine translation systems, Google Translate was better suited for the automatic translation of poetry, and this claim is confirmed by all three aforementioned studies.

# References

Aranberri, N. (2017). What Do Professional Translators Do when Post-Editing for the First Time? First Insight into the Spanish-Basque Language Pair. *Hermes – Journal of Language and Communication in Business*, 56, p. 22. doi: https://doi.org/10.7146/hjlcb.v0i56.97235

Burchardt, A., & Lommel, A. (2014). *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. Preparation and Launch of a Large-scale Action for Quality Translation Technology (QTLaunchPad), project report (p. 19).

Dede, V. (2019). Does a formal post-editing training affect the performance of novice post-editors? An experimental study, preprint (p. 16). doi: 10.13140/RG.2.2.23578.08643

Dunđer, I. (2013). CroSS: Croatian Speech Synthesizer - design and implementation. In *Proceedings of the 16th International Multiconference INFORMATION SOCIETY - IS 2013 / Collaboration, Software and Services in Information Society (CSS'2013)*, vol. A (pp. 257–260).

Dunđer, I. (2020). Machine Translation System for the Industry Domain and Croatian Language. *Journal of Information and Organizational Sciences (JIOS)*, 44(1), 33–50.

Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene* (*Statistical machine translation and computational domain adaptation*), doctoral dissertation. Zagreb: University of Zagreb, p. 281.

Dunđer, I., & Pavlovski, M. (2018). Computational Concordance Analysis of Fictional Literary Work. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018)* (pp. 0644–0648).

Dunđer, I., & Pavlovski, M. (2019a). Behind the Dystopian Sentiment: a Sentiment Analysis of George Orwell's 1984. In *Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019)* (pp. 0685–0690).

Dunđer, I., & Pavlovski, M. (2019b). Through the Limits of Newspeak: an Analysis of the Vector Representation of Words in George Orwell's 1984. In *Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019)* (pp. 0691–0696).

Dunđer, I., Pavlovski, M., & Seljan, S. (2020). Computational Analysis of a Literary Work in the Context of Its Spatiality. In *Proceedings of the World Conference on Information Systems and Technologies (WorldCIST 2020): Trends and Innovations in Information Systems and Technologies* / Advances in Intelligent Systems and Computing book series (AISC, volume 1159) (pp. 252–261).

Dunđer, I., Seljan, S., & Arambašić, M. (2013). Domain-Specific Evaluation of Croatian Speech Synthesis in CALL. In *Proceedings of the 7th European Computing Conference (ECC '13) – Recent Advances in Information Science (Recent Advances in Computer Engineering Series 13) / Language and Text Processing* (pp. 142–147).

Dunđer, I., Seljan, S., & Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. In *Proceedings of the 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020)* (pp. 1034–1039).

Dunđer, I., Seljan, S., & Stančić, H. (2015). Koncept automatske klasifikacije registraturnoga i arhivskoga gradiva (The concept of the automatic classification of the registry and archival records). In *Proceedings of the 48. savjetovanje hrvatskih arhivista (HAD) / Zaštita arhivskoga gradiva u nastajanju* (pp. 195–211). Hrvatsko arhivističko društvo.

Flanagan, M. (1994). Error classification for MT evaluation. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-1994)* (pp. 65–72). Association for Machine Translation in the Americas.

García, I. (2014). Training quality evaluators. *Revista Tradumàtica*, 12, 430–436.

Görög, A. (2014a). Quality Evaluation Today: the Dynamic Quality Framework. In *Proceedings of* *the Translating and The Computer 36 Conference* (pp. 155–164). The International Association for Advancement in Language Technology.

Görög, A. (2014b). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Revista Tradumàtica*, 12, 443–454.

Guerberof Arenas, A. (2019). Pre-editing and post-editing. In: E. Angelone, M. Ehrensberger-Dow, Maureen, & G. Massey, Gary (Eds.) *The Bloomsbury Companion to Language Industry Studies* (p. 42). Bloomsbury Academic, London.

Heinisch, B., & Lušicky, V. (2019). User expectations towards machine translation: A case study. In *Proceedings of the Machine Translation Summit XVII* (*MTSummit 2019)*, vol. 2: Translator, Project and User Tracks (pp. 42–48). European Association for Machine Translation.

Jaworski, R., Dunđer, I., & Seljan, S. (2021). Usability Analysis of the Concordia Tool Applying Novel Concordance Searching. In *Proceedings of the International Conference on Information Technology & Systems (ICITS 2021): Advances in Intelligent Systems and Computing* (2021) (pp. 128–138).

Jaworski, R., Seljan, S., & Dunđer, I. (2017). Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour. In *Proceedings of the 8th Language & Technology Conference – Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 332–336).

Jiménez-Crespo, M. A. (2017). How much would you like to pay? Reframing and expanding the notion of translation quality through crowdsourcing and volunteer approaches. *Perspectives - Studies in Translation Theory and Practice*, 25(3), 478–491.

Lengyel, I. (2011). Translation Quality Assessment at the Industrial Level: Methods for Professional Translation Quality Assessment. In: Horváth, I. (Ed.). *The Modern Translator and Interpreter* (pp. 123–138). Eötvös University Press, Budapest.

Lommel, A., Burchardt, A., Popović, M., Harris, K., Avramidis, E., & Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)* (pp. 165–172). European Association for Machine Translation.

Matusov, E. (2019). The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation* (pp. 10–19). European Association for Machine Translation.

Nunziatini, M., & Marg, L. (2020). Machine Translation Post-Editing Levels: Breaking Away from the Tradition and Delivering a Tailored Service. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 309–318). European Association for Machine Translation.

O'Brien, S. (2002). Teaching Post-Editing: A Proposal for Course Content. In *Proceedings* of *the 6th EAMT Workshop "Teaching machine translation" (EAMT 2002)* (p. 8). European Association for Machine Translation.

Pavlovski, M., & Dunđer, I. (2018). Is Big Brother Watching You? A Computational Analysis of Frequencies of Dystopian Terminology in George Orwell's 1984. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018)* (pp. 0638–0643).

Popović, M. (2020). Informative Manual Evaluation of Machine Translation Output. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5059–5069).

Popović, M., & Vilar, D. (2019). Automatic error classification with multiple error labels. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track (EAMT)* (pp. 87–95).

Rico, C., & Torrejón, E. (2012). Skills and Profile of the New Role of the Translator as MT Post-editor. *Revista Tradumàtica*, 10, 166–178.

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation. https://doi.org/10.1007/s10579-021-09537-5. Springer.

Seljan, S., & Dunđer, I. (2013). Automatic Word-Level Evaluation and Error Analysis of Formant Speech Synthesis for Croatian. In *Proceedings of the 4th European Conference of Computer Science (ECCS '13) – Recent Advances in Information Science (Recent Advances in Computer Engineering Series 17) / Image, Speech and Signal Processing* (pp. 172–178).

Seljan, S., & Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. In *Proceedings of the International Conference on Embedded Systems and Intelligent Technology (ICESIT 2014)* / International Journal of Computer, Information, Systems and Control Engineering, 8(11) (pp. 1069–1075). World Academy of Science, Engineering and Technology.

Seljan, S., & Dunđer, I. (2015a). Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain. In

*Proceedings of the 10th Iberian Conference on Information Systems and Technologies (CISTI'2015),* vol. 2, (pp. 128–131). Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI).

Seljan, S., & Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. In *Proceedings of the International Conference "Corpus Linguistics – 2015" (CORPORA 2015)* (pp. 72–79). St. Petersburg State University, Russian Academy of Sciences, & Herzen State Pedagogical University of Russia.

Seljan, S., Dunđer, I., & Gašpar, A. (2013). From Digitisation Process to Terminological Digital Resources. In *Proceedings of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013)* (pp. 1329–1334).

Seljan, S., Dunđer, I., & Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. *In Proceedings of the 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2020)* (pp. 1040–1045).

Seljan, S., Dunđer, I., & Stančić, H. (2014). Extracting Terminology by Language Independent Methods. In *Proceedings of the 2nd International Conference on Translation and Interpreting Studies "Translation Studies and Translation Practice" (TRANSLATA II)* / Peter Lang series "Forum Translationswissenschaft", vol. 19 (pp. 141–147). Peter Lang Publishing.

Seljan, S., Tucaković, M., & Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. In *Proceedings of the 3rd World Conference on Information Systems and Technologies (WorldCIST'15)* / Advances in Intelligent Systems and Computing – New Contributions in Information Systems and Technologies (pp. 1089–1098). Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI).

Stasimioti, M., Sosoni, V., Kermanidis, K., & Mouratidis, D. (2020). Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 441–450). European Association for Machine Translation.