TRANSLATION QUALITY EVALUATION OF CROATIAN-TO-ENGLISH MACHINE-TRANSLATED ADMINISTRATIVE TEXTS

*Mirjana **Borucinsky**, University of Rijeka Faculty of Maritime Studies, Rijeka, Croatia, mirjana.borucinsky@uniri.hr*

*Jana **Kegalj**, University of Rijeka Faculty of Maritime Studies, Rijeka, Croatia, jana.kegalj@uniri.hr*

*Dario **Zagorec**, DZ Translations, obrt za prevođenje, vl. Dario Zagorec, Varaždin, Croatia, Ulica braće Radić 6, zagorec.dario@gmail.com*

Original scientific paper
DOI: 10.31902/fll.51.2025.10
UDC: 811.111'322.4:811.163.42

Abstract: A once heavily flawed method of translation, machine translation (MT) has improved and continues to improve every day. One of the major issues during its development was its quality and how to measure and assess it objectively. This paper presents an attempt to apply a triangulation of translation quality assessment (TQA) methods: automatic and human assessment, corpus-based analysis and error analysis, on the translation of a sample of administrative texts from Croatian into English, to provide a comprehensive view of the said translation, compare the results of different methods and identify areas of greatest discrepancy in machine-generated translation. The texts were translated using Google Translate (GT), which relies on the currently dominant neural model that significantly reduces translation errors when compared to the phrase-based model, and it is expected to deal with issues such as congruence (i.e. agreement) and inflection better than other systems. This is of special importance for morphologically rich languages such as Croatian. Even though literature on MT and the evaluation of MT is abundant, this paper aims to contribute to research of an under-resourced Slavic language, Croatian.

Keywords: machine translation, translation quality assessment, administrative texts, Croatian, English

1. Introduction

The debate about the efficiency of machine translation (MT) vs human translation (HT) is ongoing, particularly in light of recent developments in artificial intelligence and the rise of ChatGPT. Although it is generally accepted that human translation still outperforms machine-generated translation, due to many advantages such as speed, flexibility, cost-efficiency, etc., online translation tools have become increasingly popular in recent years, also for less common languages such as Croatian, and especially in cases where high-quality translations are not required, but rather the quintessence of a paragraph, a website, a conference or product information (cf. Seljan et al. 331).

The current state of machine translation is such that most machine-translated texts require extensive post-editing by a human editor, as the translated texts are often riddled with errors that make the texts confusing and often even comical. In addition, MT programmes do not perform many translation operations applied by human translators, such as sentence separation, function and/or category shifts, explicitation, modulation and paraphrasing. The length, information flow and structure of machine-translated texts are more similar to the source text than to a text translated by a human (Ahrenberg 26). The result is a text that feels strange and robotic, that cannot be accepted as it is, and that needs to be revised by a human.

MT can, among other things, be used to gain insight into existing problems in the translation and improvement of MT software. Therefore, evaluating MT is important for researchers, product developers and users alike (cf. Hovy et al. 1). A number of researchers (e.g. Gonzàlez and Giménez 77; Graham et al. 1183; Bentivogli et al. 62) have highlighted the crucial importance of evaluating MT, as it is not only used to compare different systems, but also to identify a system's weaknesses and improve it (Gonzàlez and Giménez 77). The quality of the output of MT has recently been assessed in the context of fitness-for-purpose models (cf. Way 16; Jiménez-Crespo 73), which require that the translation is appropriate for the purpose or audience for which it is intended. The literature on MT and the evaluation of MT is vast. However, research focusing on morphologically rich languages such as Croatian is scarce (cf. Tadić; Simeon; Seljan et al.; Pavlović; Ljubas; Klubička et al). The research has shown that rich morphology and relatively free word order cause particular difficulties for MT.

This study is an attempt to evaluate the quality of machine-translated administrative texts for the Croatian-English language pair and to identify the areas of greatest weaknesses of the MT tool. The findings can be used for further improvement of MT systems for

Croatian-to-English translation and might benefit MT developers in improving MT fluency and adequacy.

2. An overview of contemporary MT models

The shift from rule-based to Statistical Machine Translation (SMT) has been significant in the development of MT. SMT methods are usually phrase-based and have been used extensively since 2002 (Wu et al. 2). In 2007, Google Online Translator was made available, which relied more on the statistical and less on the rule-based approach. SMT searches for existing translations and calculates, among other things, the probability that a word is translated with another word (cf. Ahrenberg and Merkel 42). In recent years, however, a new type of MT has been developed: Neural Machine Translation (NMT) (cf. Kalchbrenner and Blunsom 1701; Cho et al. 1724; Sutskever et al. 3104; Bahdanau et al. 1; Srivastava et al.; Goldberg). In SMT, the translation model and the language model are trained separately and combined during decoding (Koehn), while NMT learns a single large neural network that inputs a sentence and outputs a translation (Yuan and Briscoe 380). According to Wu et al. (1), NMT "has the potential to overcome many of the weaknesses of conventional phrase-based translation systems" and was expected to handle problems such as congruence and inflexion better than other systems (Castilho et al. 110-11), which is important for morphologically rich languages such as Croatian. According to Wu et al. (19), Google Neural Machine Translation (GNMT) "reduces translation errors by an average of 60% compared to Google's initial phrase-based production system".

In a study involving the English-Croatian language pair, Klubička et al. (121) found that "the best-performing system (neural) reduces the errors produced by the worst system (pure phrase-based) by more than half (54%)". However, NMT systems sometimes do not translate all parts of the input sentence (Wu et. al. 2) and the output data is considered less accurate than that obtained by SMT. Toral and Sánchez-Cartagena (1069) found that NMT shows a significant drop in translation quality for sentences longer than 40 words. In a study on Croatian and Swedish translations, Ljubas (88-89) found that the output data for GNMT is less accurate than the one based on a statistical model, and that NMT for the Croatian-Swedish language pair has more untranslated words than the previous version of GT for the same language pair, but performs better on morphology. Similarly, Vieira (325) reports that research has shown that phrase-based statistical machine translation performs better when it comes to conveying meaning, but NMT performs better when it comes to fluency.

Moreover, training an NMT system on a large translation dataset requires a lot of time and computational resources. Due to the large number of parameters used in NMT systems, they are generally much slower than phrase-based systems. Another important element is the amount of input data and the accuracy of the data used for training, which is even more difficult for under-resourced languages like Croatian. It is also worth mentioning that most MT systems are developed for large languages such as English, German, Chinese, etc., while research for smaller languages such as Croatian is always somewhat sparse.

3. An overview of MT evaluation methods

When using MT, expectations must be realistic to avoid misunderstandings (Kit and Wong, 302). Closely related to expectations of the MT is its evaluation. However, there is no general agreement on the assessment of translation quality, as there are no standardised criteria for assessment and the process itself is a complex one involving both linguistic and extra-linguistic aspects.

In recent years, numerous evaluation techniques have been developed and used for all types of MT systems. Although assessment techniques have been classified as either automated or human (manual) metrics, in practice, they are not so easy to distinguish. Automated assessment uses either human translations or human annotations, and what is done automatically is a calculation; on the other hand, human assessment implies human intervention during the evaluation phase, but also includes the use of computerised tools and automated processes. In between these two categories are the so-called semi-automated metrics.

As human translation evaluation is costly, subjective and slow, various automatic scores have been developed (Dorr et al. 805). Automatic evaluation measures system performance and identifies weaknesses (cf. Agarwal & Lavie 115-16; Giménez and Màrquez 77, Denkowski & Lavie 6-7) using various language-independent algorithms. Automatic metrics generally compare the output of MT with a reference translation produced by a human translator. The best-known systems are BLEU (cf. Papineni et al.) and NIST (cf. Doddington). One of the oldest scores, Word Error Rate (WER), which is based on Levenshtein distance, operates at the word level and observes word sequences, the number of substitutions, deletions, insertions and correct words (cf. Nießen et al. 40; Mauser et al. 3090; Han 5; Castilho et al. 17). Since this does not allow for word reordering, other metrics such as Position-Independent Word Error rate (PER, Tillmann et al., 2669), Sentence Error Rate (SER, Tomás et al. 28) and Translation Error Rate (TER, Snover et al. 223-24)

have been developed to address the problem. The metric BLEU goes beyond individual words and measures quality at the n-gram level. As it has shown good correlation with human evaluations, it was also used in this study to assess the quality of MT translation from Croatian into English. In the last decade, a number of new generation automatic metrics have been proposed that outperform BLEU, such as MaxSim (Chan and Ng 55), ULC (Gimenez and Marquez 218), RTE (Padó et al. 182), posBleu (Popović and Ney 29) and chrF (Popović). However, these still cannot fully cover some linguistic phenomena such as synonymy and paraphrasing (Dorr et al. 870).

In order to obtain metrics that provide results closer to human evaluation results, a qualitative evaluation of different linguistic phenomena in combination with statistical approaches is required. Human evaluations of MT take into account various aspects of translation such as adequacy, fidelity and fluency of translation (Hovy 1-2; White and O'Connell 136; White et al. 196-97). Although Hovy (6) argues that human evaluation approaches are expensive and slow, and some research (cf. Snover et al 229) has shown low correlation between annotators, the software developed for evaluating MT is trained on data provided by human evaluators/annotators.

The metrics most commonly used by human evaluators are fluency and adequacy. In fluency, the output of MT is judged on whether it is fluent, regardless of its fidelity to the input, whereas in adequacy the evaluator judges whether the output contains the essential information of the input (cf. Snover et al. 223; Dorr et al. 804). In assessing fluency, evaluators or raters can be monolingual, while in assessing adequacy they should preferably be bilingual. It is also possible to use both professional and amateur raters in Translation Quality Assessment (TQA), (Castilho et al. 9), who can rate individually, in groups or in a crowd. Scores are usually ranked on a scale and then averaged into a single score for a given output. Although the reliability of the method is debatable, the assessment of semantic adequacy by human annotators is still very useful (Dorr et al. 804). Ideally, more than one assessor is involved in TQA tasks and the inter-rater agreement between assessors is then calculated (Chatzikoumi 158).

In this study, the automatic evaluation scores BLEU and WER were combined with human evaluation carried out by five evaluators, who rated randomly selected sentences of the machine-translated texts on a Likert scale from 1 to 5. The texts were also checked for errors by three error checking tools which classified the errors into several predefined categories and provided an overall score for the translated text.

4. Methodology

This paper takes a somewhat unorthodox approach to the analysis of machine-translated texts, as it applies several different analytical tools to obtain a comprehensive view of the translated text. The initial analysis uses corpus tools to compare corpora and gain insight into their internal composition. To this end, three corpora have been compiled:

1. A corpus of original texts (i.e. source texts, ST) in Croatian, counting 21547 tokens,

2. A corpus of machine-translated texts from Croatian ST into English target texts (TT), counting 19380 tokens,

3. A corpus of English TT translated by a human translator, as a reference corpus, with 19639 tokens.

The compilation of the corpora was register-oriented. Both the source texts and the human-translated texts all originated from the website of the Ministry of Science and Education of the Republic of Croatia [1]. The texts are administrative in nature and contain the government's recommendations for organising the working day of students in distance education which were implemented in Croatia in 2020 during the COVID-19 epidemic. In addition, the texts contain guidelines for assessment and grading in a virtual environment, guidelines for distance education for primary and secondary schools, and information about meetings and conferences attended by Croatian politicians, also written in administrative style.

The administrative register in Croatian has its own lexico-grammatical features, the most important of which are nominalisation, objectivity, accuracy, clarity, conciseness, stylistically unmarked language structures, formulaicity and established macrostructure (cf. Silić; Frančić et al.). English administrative texts share similar features (cf. Swales), which makes the texts relatively easy to translate for MT systems, in contrast to literary texts or texts of everyday language. The latter use many colloquial expressions, which makes it most difficult for MT systems to translate them correctly. The texts were translated by Google Translate (GT) using onlinedoctranslator.com, as it is the most accessible and user-friendly system. Moreover, it can translate longer texts or entire documents instantly.

The three corpora were compared in terms of their general characteristics, such as the number of words and the distribution of parts of speech (POS), in order to identify possible differences between them. The Sketch Engine platform (Kilgarriff et al.) was used for this

[1] mzo.gov.hr

purpose. Statistical measures such as lexical density, average sentence length and lexical richness were also used to examine the corpora and gain useful insights into their features.

The next step was to use the automatic metric BLEU, which compares n-grams of human-generated translation, which is considered to be the gold standard, with n-grams from MT. As mentioned earlier, this metric was used because it has shown good correlation with human judgements in several studies (cf. Papineni et al 318). Besides that, the WER score was used to compare the results.

In the following step of the research, a combination of automatic and human judgement was used. First, the MT translation was checked using the quality assurance and terminology checking tools Xbench[2], ProWritingAid [3] and Grammarly [4] . These applications check completeness, consistency, terminology and spelling, and provide feedback on the quality of the translation.

The final part of the research included five evaluators, native speakers of Croatian with a university degree in English, a PhD in linguistics, and at least ten years' experience in translation and language teaching. Based on their ratings, the inter-rater agreement was calculated as an indicator of reliability. The results of this evaluation were then compared with the results of the automatic MT quality assessment.

5. Results and discussion

5.1. Corpus analysis

In the first step of the analysis, the three corpora were compared in terms of the number of words and the distribution of parts of speech (POS), to see if there were major differences between them. The results are shown in Table 1.

Table 1. Differences in the number of words and distribution of POS between texts

[2] ApSIC XBench 3.0 Build 1546, https://www.xbench.net/index.php
[3] ProWritingAid - https://prowritingaid.com
[4] Grammarly - https://grammarly.com

	Corpus 1 Source texts (Croatian) ST	Corpus 2 Machine-translated target texts (English) MT	Corpus 3 Human-translated target texts (English) HT
Total words	17570	17126	17450
Nouns	6450	5356	5562
Verbs	2370	2692	2817
Adjectives	2037	1534	1609
Adverbs	566	670	450
Pronouns	1108	562	652
Conjunctions	1643	934	876
Prepositions	2508	2587	2624
Numeral	191	138	167

Both MT and HT translations have more words overall compared to the ST, which can be attributed to language-intrinsic differences, such as the use of articles in the TL. The ST also contains significantly more pronouns, conjunctions and adjectives, but fewer verbs and prepositions. Another interesting fact is the difference in the number of nouns and verbs in Croatian and English texts. The administrative genre is known for its tendency towards nominalisation, which is particularly pronounced in Croatian. This is also reflected in the number of adverbs and adjectives, as adverbs describe actions and are therefore associated with verbs, while adjectives describe nouns. In the case of adverbs, there is a discrepancy between MT and HT, whereby the MT corpus shows a greater frequency of adverbs. This could indicate possible discrepancies in the style of MT, but further insights are necessary to make this conclusion.

The difference in the number of pronouns can be attributed to two trends. One is the generic property of impersonality, which is expressed by the passive voice in English, and the other is the deictic function of pronouns in Croatian, which is not reflected in the same way in the English text. The biggest differences between MT and HT into English are in the number of nouns, which is larger in the human translation and could be attributed to the interference of SL, and in the number of conjunctions, which is larger in MT and contributes to a greater cohesion of the text. The greater frequency of conjunctions in ST might indicate more explicit syntactic relations between sentences and clauses in Croatian.

The results of the corpus were also used for statistical calculations that provide information about lexical density, such as the type-token ratio (TTR), the Halliday lexical words per sentence ratio (LDHal), the noun-verb ratio (N/V), the average sentence length and the lexical to functional words ratio (L/F), as shown in Table 2.

Table 2. Statistical indicators from the corpus

	Corpus 1 ST	Corpus 2 MT	Corpus 3 HT
TTR	0.82	0.88	0.89
AVS	30.88	31.9	30.6
N/V	2.72	1.99	1.97
L/F	1.77	1.84	1.94
LDHal	14	14	14

The results show similar trends in terms of TTR and LDHal, while the differences occurred in average sentence length, with MT texts having the longest sentences, which is common in English administrative texts. In terms of the N/V ratio, Croatian ST shows the greatest tendency towards nominalisation, while both English texts show similar results with the N/V ratio indicating a larger number of verbs in relation to nouns. This indicates that there was no interference from the SL in this segment, but that the texts conformed to the norms of the TL. Considering that there are far more resources for English on which machine translation can rely, this level of compliance to target language structure and norms can be expected. Another difference is in the ratio between lexical and functional words, which is highest in the human-translated texts. This could indicate a tendency towards explicitation of the grammatical relationship within the sentence. In general, it can be seen that the MT texts have similar characteristics to the human-translated texts.

5.2. Automatic evaluation

The second part of the study involved the calculation of the automatic metric evaluation scores BLEU (Bilingual Evaluation Understudy) and WER (Word Error Rate). In this study, the BLEU score is calculated by comparing the machine-generated translation with the reference human translation, which is considered the “gold standard”. In this study, the result of the BLEU score calculated in Python was 0.35. This is a relatively low value indicating that machine translation has significant differences from the reference translation, i.e. it is not very

accurate when compared to the gold standard. However, it still falls into the category of understandable translations.

The WER value is relatively high, 64.8%, and indicates a significant deviation between the machine translation and the human translation, as it indicates that almost 65% of the words in MT differ from the words in HT. This indicates that the machine translation is rather inaccurate and does not remain faithful to the original text. There are some possible reasons for the high WER score, such as that the MT system does not understand the semantic nuances and context well, that there are some grammatical and syntactic errors due to wrong word order or subject-verb agreement, that the MT system does not handle ambiguity well, that there were some domain-specific words, or that there are some inherent limitations in the MT system.

In any case, these two values indicate a lower quality of the output of the MT system compared to the human-generated translation, but one which can still be considered understandable.

5.3. Error analysis

The following part of the study involved the analysis of errors using commercial products such as ProWritingAid, Grammarly and XBench. These tools detected some problems with the grammar and style of the machine-generated translation and flagged them accordingly.

ProWritingAid identified style as particularly problematic, with a passive index of 34, with 87 hidden verbs and 31 repeated sentence starters. The report gave an overall score of 66/100; specifically, 74/100 for grammar, 80/100 for spelling and 44/100 for style. This is in line with the findings of the corpus analysis. ProWritingAid found the three biggest problems to be: (i) low readability, indicating that the text is difficult to read, (ii) a high "glue index", indicating a large number of filler words in the text, (iii) many long sentences, making the text difficult to read (there are 203 long sentences in the text out of a total of 797 sentences), which is in line with the results of corpus analysis. In terms of readability, ProWritingAid identified 137 very difficult-to-read paragraphs, which is about one third of the total number of paragraphs identified, and 27 slightly difficult-to-read paragraphs.

Grammarly analysed the text in terms of several aspects, namely correctness, clarity, engagement and delivery. However, in this context, Grammarly often missed many errors, especially in terms of correctness, i.e. spelling, grammar and punctuation errors, and often identified correct sentences as errors. As far as clarity is concerned, the overall score is positive, which means that the machine-generated translation serves its purpose, i.e. that it conveys a sufficiently accurate meaning in

the target language. Grammarly has identified 316 clarity issues in the corpus, making the corpus overall "a bit unclear". In the "Engagement" category, Grammarly highlighted frequently overused words or words that appear repeatedly in the text and identified 224 engagement problems in the corpus, grading this aspect as "a bit bland". When it comes to the "Delivery" category, only 5 problems were identified, making delivery just right. Among the parts of speech, lexical, morphosyntactic and stylistic errors dominate the corpus, as expected, accounting for 89 of all errors. Morphosyntactic errors are due to the differences between English and Croatian, i.e. English, unlike Croatian, has a relatively fixed word order and Croatian, unlike English, is inflectionally rich. Untranslated words are not that frequent in the corpus, but omissions of text segments are, which is consistent with the findings of Ljubas (2018). Some inconsistent translations were found, probably due to the absence or inconsistency of input data.

In addition to the discrepancies identified by ProWritingAid and Grammarly, the XBench programme identified 17 key term mismatches, as well as word repetitions, numeric and alphanumeric mismatches and untranslated segments. Since these key terms cannot be considered rare or complex, the inconsistencies found could be due to the system not understanding the semantic differences, the system's inability to resolve ambiguities, a lack of training data that may not be representative of this particular domain, or a lack of context. Clearly, the fact that the translation direction was from a low-resourced language to a highly resourced language had no influence in this case, and it could suggest that the quality might depend on the resources for both languages, although this claim requires further study.

5.4. Human evaluation

The following part of the study involved the evaluation by five raters. The raters rated 55 randomly selected sentences from the MT on a scale of 1 to 5, according to values described in Table 3 (cf. Sanders et al., 756). The raters were given an explanation of the grading scale. The grading scale shown in Table 3 was adjusted and revised for this study according to Koehn and Munz (107) and Waddington (22).

Table 3. Grading scale

Grade	Adequacy	Fluency	Degree
5	All meaning: complete transfer of information, minor revisions required	Flawless English: reads like a piece originally written in English	Successful
4	Most meaning: almost complete transfer, one or two insignificant inaccuracies	Good English: seems like originally written in English to a larger extent; some lexical, grammatical or spelling errors	Almost completely successful
3	Much meaning: general idea is transferred but there are lapses in accuracy; needs considerable revision	Non-native English: reads like a translation; considerable number of lexical, grammatical or spelling errors	Adequate
2	Little meaning: serious inaccuracies; thorough revision required	Disfluent English: the sentence reads completely as a translation, sounds foreign	Inadequate
1	None: totally inadequate and inconsistent	Incomprehensible: sentence is not intelligible	Totally inadequate

The raters rated the sentences according to the fluency and adequacy of the translation. Table 4 shows a summary of the results,

where P0 indicates raw agreement, Pe indicates expected purely random agreement and Fleiss' kappa value to measure the agreement among multiple raters. The results were obtained by using an online statistical tool developed by Lancaster University[5].

Table 4. Results of calculated inter-rater agreement for fluency and accuracy

	P0	Pe	Fleiss' kappa
Fluency	35.56%	0.21	0.12
Adequacy	30.56%	0.14	0.08

In terms of fluency, the evaluators gave the overall fluency an average score of 3.82, with 36% of the ratings being in agreement. The Pe of 0.21 indicates moderate agreement among the raters, with a p-value of < 0.001, indicating that the agreement is statistically significant, i.e. that it did not occur merely by chance. Fleiss' kappa takes both P0 and Pe into account and can range from -1, indicating agreement worse than chance, to 1, indicating perfect agreement. In this case, the value of 0.12 indicates a slight agreement among the raters beyond what can be expected by chance. The calculated p-value (0.001) indicates that the result is statistically significant.

For adequacy, the mean score of raters was lower than for fluency, 3.5. The raw agreement of 31% is slightly lower than the raw agreement for fluency, and the Pe of 0.14 indicates a low level of agreement among raters. The p-value (0.001) indicates that the results are statistically significant. The Fleiss' kappa of 0.08 indicates slight agreement among the raters beyond the expected random agreement. The calculated p-value in this case (0.015) shows that the results are statistically significant. The low agreement indicates that there is significant inconsistency in the ratings beyond what would be expected by chance, but the p-values indicate that the calculated agreement did not arise by chance alone.

Despite the slight agreement among the raters, it can be concluded that the MT translation is fluent to a certain degree, but it does not preserve the source meaning adequately.

[5] http://corpora.lancs.ac.uk/stats/toolbox.php

6. Conclusion

In this paper, the authors conducted an evaluation of MT translation of administrative texts from Croatian into English based on a triangulation of existing evaluation methods in order to gain deeper insight into the nature of errors, consistency and coherence in the structural level of translation and to provide a comprehensive evaluation of translation quality based on different evaluation methods (automatic and human assessment, corpus analysis and error detection). The corpus analysis revealed similarities between MT and HT, i.e. the gold standard in the general corpus data, such as the distribution of parts of speech, type-token ratio, N/V ratio. MT had slightly longer sentences than HT, while the ratio of lexical to functional words in MT showed somewhat fewer functional words than in HT. The automatic evaluation measures, BLEU and WER, showed a significant deviation of MT from HT, but still grading MT as comprehensible. In a further step, MT was evaluated by three programmes which identified various errors in MT. They pointed to problems with style, clarity and key terminology. The MT was also evaluated by human raters, who gave it an average score of 3.8 for fluency and 3.5 for adequacy, with moderate inter-rater agreement for fluency and low agreement for adequacy.

In general, all the evaluation methods used showed an understandable translation, but with significant problems with style and key terminology. Corpus analysis showed that MT lacks the underlying structural rules of the target language, which possibly influenced fluency, but the other methods of analysis (e.g. the higher score on fluency and lower score on accuracy in human evaluation) showed significant problems with adequacy. The reasons for this could be traced to the SL interference and differences between the two languages, the different terminology in the different sources and insufficient input data for MT systems. Since MT models rely on existing data and perform worse when fewer resources are available, there is a need to provide them with more data and conduct further research in this area, especially for smaller languages such as Croatian. As suggested by Ljubas (89), there is still a great need for more concrete proposals to improve MT systems for Croatian.

Despite the limitations in the study in terms of a small corpus limited to administrative texts and the fact that the raters were not native speakers of English, the triangulation of methods provided a more comprehensive insight into machine translation quality and pointed to areas for further improvement – mainly style and generic features. Further studies could focus on larger corpora, different genres,

different translation directions and different language pairs for more insight.

Works cited:

Agarwal, Abhaya and Alon Lavie. "METEOR, M-BLEU and M-TER: Evaluation Metrics for High Correlation with Human Rankings of Machine Translation Output." *Proceedings of the ACL 2008 Workshop on Statistical Machine Translation*. Ed. Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, Cameron Shaw Fordyce. Columbus, Ohio: Association for Computational Linguistics, 2008. 115-118.

Ahrenberg, Lars and Magnus Merkel. "Correspondence Measures for MT Evaluation." *Proceedings of the LREC 2000 Workshop on Evaluation of Machine Translation*. Ed. Bente Maegaard, Athens, Greece: LREC, 2000. 41-46.

Ahrenberg, Lars. "Comparing Machine Translation and Human Translation: A Case Study." *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*. Ed. Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, Stephan Vogel. Shoumen, Bulgaria: Association for Computational Linguistics, 2017. 21-28.

Bahdanau, Dzmitry, Cho, KyungHyun and Bengio, Yoshua. "Neural machine translation by jointly learning to align and translate." *3rd International Conference on Learning Representations*. Ed. Yoshua Bengio, Yann LeCun. San Diego, CA, USA: ICLR, 2015.

Bentivogli, Luisa, Cettolo, Mauro, Federico, Marcello and Christian Federmann. "Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment." *Proceedings of the 15th International Workshop on Spoken Language Translation.* Ed. Marco Turchi, Jan Niehues, Marcello Frederico. Bruges, Belgium: International Conference on Spoken Language Translation, 2018. 62-69.

Castilho, Sheila, Moorkens, Joss, Gaspari, Federico, Calixto, Iacer, Tinsley, John and Andy Way. "Is Neural Machine Translation the New State of the Art?" *The Prague Bulletin of Mathematical Linguistics* 108.1 (2017): 109-120.

Castilho, Sheila, Doherty, Stephen, Gaspari, Federico, and Joss Moorkens. "Approaches to Human and Machine Translation Quality Assessment." *Translation Quality Assessment: From Principles to Practice.* Ed. Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty. Springer, 2018. 9-38.

Chan Yee Seng and Ng Hwee Tou. "MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation." *Proceedings of ACL-08: HLT*. Ed. Johanna D. Moore, Simone Teufel, James Allan, Sadaoki Furui. Columbus, Ohio, USA: Association for Computational Linguistics, June 2008. 55-62.

Chatzikoumi, Eirini. "How to evaluate machine translation: A review of automated and human metrics." *Natural Language Engineering*, 26.2 (2020): 137–161. <https://doi.org/10.1017/S1351324919000469>.

Cho, Kyunghyun, van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger and Yoshua Bengio. "Learning phrase representations using RNN encoder decoder for statistical machine translation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. Alessandro Moschitti, Bo Pang, and Walter Daelemans. Qatar: Association for Computational Linguistics, 2014. 1724–1734. <https://doi.org/ 10.3115/v1/D14-1179>.

Denkowski, Michael J. and Alon Lavie. "Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks." *Proceedings of AMTA*, Denver, Colorado, USA: Association for Machine Translation in the Americas, 2010.

Doddington, George. "Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrence Statistics." *Proceedings of the 2nd International Conference on Human Language Technology Research*. Ed. Mitchell Marcus. San Francisco: Morgan Kaufmann Publishers, 2002. 138–145.

Dorr, Bonnie, Snover, Matt and Nitin Madnani. "Part 5: Machine translation evaluation." *DARPA GALE program report*. Ed. Bonnie Dorr, 2009. 801-894.

Frančić, Anđela, Hudeček, Lana, and Milica Mihaljević. *Normativnost i višefunkcionalnost u hrvatskome standardnom jeziku*. Zagreb: Hrvatska sveučilišna naklada. 2006.

Giménez, Jesus and Lluís Màrquez. "Linguistic Measures for Automatic Machine Translation Evaluation." *Machine Translation* 24.3/4 (2011): 209-40.

Giménez, Jesus and Lluís Màrquez. "A smorgasbord of features for automatic MT evaluation." *Proceedings of the third workshop on Statistical Machine Translation*. Ed. Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, Cameron Shaw Fordyce. Columbus, Ohio: Association for Computational Linguistics, 2008. 195-198.

Giménez, Jesus and Lluís Màrquez. "Asiya: An open toolkit for automatic machine translation (meta-) evaluation." *The Prague Bulletin of Mathematical Linguistics* 94 (2010): 77-86.

Goldberg, Yoav. *Neural Network Methods in Natural Language Processing: Synthesis Lectures on Human Language Technologies*. Springer. 2017.

Graham, Yvette, Mathur, Nitika and Timothy Baldwin. "Accurate Evaluation of Segment-level Machine Translation Metrics." *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. Ed. Rada Mihalcea, Joyce Chai, Anoop Sarkar. Denver, Colorado: Association for Computational Linguistics, 2015. 1183–1191.

Han, Lifeng. "Machine Translation Evaluation Resources and Methods: A Survey." *arXiv e-prints*, arXiv:1605.04515. 2016.

Hovy, Eduard H., King, Maghi and Andrei Popescu-Belis. "An Introduction to MT Evaluation." *Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*, Workshop at the LREC 2002 Conference, Las Palmas, Canary Islands, Spain. 2002. <https://mt-archive.net/00/LREC-2002-WS-MTEval.pdf>. Accessed 12 November 2024.

Jiménez-Crespo, Miguel A. "Crowdsourcing and Translation Quality: Novel Approaches in the Language Industry and Translation Studies." *Translation Quality Assessment: From Principles to Practice*. Ed. Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty. Springer, 2018. 69-94.

Kalchbrenner, Nal and Phil Blunsom. "Recurrent continuous translation models." *Proceedings of EMNLP*. Ed. David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, Steven Bethard. Seattle, USA: Association for Computational Linguistics, 2013. 1700-1709.

Kilgarriff, Adam, Rychlý, Pavel, Smrž, Pavel and David Tugwell. "The Sketch Engine." Itri-04-08 *Information Technology Research Institute Technical Report Series*, 2004. 105-116.

Kit, Chunyu and Tak Ming Wong. "Comparative Evaluation of Online Machine Translation Systems with Legal Texts." *Law Library Journal* 100.2 (2008): 299-321.

Klubička Filip, Toral, Antonio and Victor M. Sánchez-Cartagena. "Fine-grained human evaluation of neural versus phrase- based machine translation." *The Prague Bull Math Linguist* 108 (2017): 121-132.

Koehn, Philipp and Christof Monz. "Manual and automatic evaluation of machine translation between European languages." *Proceedings of the Workshop on Statistical Machine Translation*. Ed. Philipp Koehn, Christof Monz. New York City: Association for Computational Linguistics, 2006. 102-121.

Koehn, Philipp. *Statistical Machine Translation*. Cambridge: Cambridge University Press. 2010.

Levenshtein, Vladimir I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady*, 10 (1966): 707–710.

Ljubas, Sandra. "Analiza pogrešaka u strojnim prijevodima sa švedskog na hrvatski." *Hieronymous* 4 (2017): 28-64.

Ljubas, Sandra. "Prijelaz sa statističkog na neuronski model: usporedba strojnih prijevoda sa švedskoga na hrvatski jezik." *Hieronymus* 5 (2018): 72-91

Mauser, Arne, Hasan, Saša and Hermann Ney. "Automatic Evaluation Measures for Statistical Machine Translation System Optimization." *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC*. Ed. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Daniel Tapias. Marrakech, Morocco: European Language Resources Association (ELRA), 2008.

Nießen, Sonja, Och Franz Josef, Leusch, Gregor and Hermann Ney. "An evaluation tool for machine translation: fast evaluation for MT research." *Proceedings of the second international conference on language resources and evaluation*. Ed. M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhauer. Athens: European Language Resources Association (ELRA), 2000. 39–45.

Padó, Sebastian, Cer, Daniel, Galley, Michel, Jurafsky, Dand and Christopher D. Manning. "Measuring machine translation quality as semantic

equivalence: a metric based on entailment features." *Mach Transl* 23.2–3 (2009):181–193.

Pavlović, Nataša. "Strojno i konvencionalno prevođenje s engleskoga na hrvatski: usporedba pogrešaka." *Jezik kao predmet proučavanja i jezik kao predmet poučavanja*. Ed. Diana Stolac and Anastazija Vlastelić. Zagreb: Central Europe and CALS, 2017. 279-295.

Papineni, Kishore, Roukos, Salim, Ward, Todd and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Ed. Pierre Isabelle, Eugene Charniak, Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. 311-318.

Popović, Maja and Hermann Ney. "Syntax-oriented evaluation measures for machine translation output." *Proceedings of the fourth workshop on Statistical Machine Translation* (StatMT '09). Ed. Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder. Athens, Greece: Association for Computational Linguistics, 2009. 29-32.

Popović, Maja. "ChrF: character n-gram F-score for automatic MT evaluation." *Proceedings of the 10th workshop on Statistical Machine Translation* (WMT-15). Ed. Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina. Lisbon, Portugal: Association for Computational Linguistics, 2015. 392-395.

Sanders, Gregory, Przybocki, Mark, Madnani, Nitin and Matthew Snover. "Human Subjective Judgements." *Handbook of Natural Language Processing and Machine Translation*. Ed. Olive Joseph, Caitlin Christianson and John McCary. Springer, 2011. 750-758.

Seljan, Sanja, Brkić, Marija and Vlasta Kučiš. "Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs." *Information Sciences and E-Society*. Ed. Clive Billenness, Annette Hemera, Vladimir Mateljan, Zorica Banek, Mihaela Stančić, Hrvoje and Sanja Seljan. Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2011. 331-344.

Silić, Josip. *Funkcionalni stilovi hrvatskoga jezika*. Zagreb: Disput, 2005.

Simeon, Ivana. *Vrednovanje strojnoga prevođenja*, PhD thesis. Zagreb: Filozofski fakultet, 2008.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea and John Makhoul. "A Study of Translation Edit Rate with Targeted Human Annotation." *Proceedings of the 7th conference of the Association for Machine Translation in the Americas: Visions for the future of Machine Translation*. Cambridge, Massachusetts: Association for Machine Translation in the Americas, 2006. 223-231.

Snover, Matthew, Madnani, Nitin, Dorr, Bonnie and Richard Schwartz. "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric." *Proceedings of the Fourth Workshop on Statistical Machine Translation.* Ed. Chris Callison-Burch, Chris, Koehn, Philipp, Monz, Christof

and Josh Schroeder. Athens, Greece: Association for Computational Linguistics, 2009. 259–268.

Srivastava, Rupesh Kumar, Greff, Klaus and Schmidhuber, Jürgen. "Highway networks." *CoRR*, <https://arxiv.org/pdf/1505.00387>. 2015.

Sutskever, Ilya, Vinyals, Oriol and Quoc V Le. "Sequence to sequence learning with neural networks." *Advances in Neural Information Processing Systems*. Ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger. Montreal, Quebec, Canada: NIPS, 2014. 3104–3112.

Swales, John M. "A genre-based approach to language across the curriculum." *Language Across the Curriculum*. Ed. M. L. Tickoo. Singapore: SEAMEO Regional Language Centre, 1986.

Tadić, Marko. *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex Libris. 2003.

Tillmann, Christoph, Vogel, Stefan, Ney, Hermann, Sawaf, Hasan and Alex Zubiaga. "Accelerated DP-based search for statistical translation." *Proceedings of the 5th European conference on Speech Communication and Technology* (EuroSpeech '97). Ed. G. Kokkinakis, N. Fakotakis, E. Dermatas. Rhodes, Greece: European Speech Communication Association, 1997. 2667-2670.

Tomás, Jesús, Mas Josep Àngel, and Francisco Casacuberta. "A Quantitative Method for Machine Translation Evaluation." *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?* Ed. Katerina Pastra. Columbus, Ohio: Association for Computational Linguistics, 2003. 27-34.

Toral, Antonio and Victor M Sánchez-Cartagena. "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions." *Proceedings of the 15th Conference of the European chapter of the association for computational linguistics*. Ed. Mirella Lapata, Phil Blunsom, Alexander Koller. Valencia, Spain: Association for Computational Linguistics, 2017. 1063-1073.

Vieira, Lucas Nunes. "Post-Editing of Machine Translation." *The Routledge Handbook of Translation and Technology*. Ed. M. O'Hagan. Routledge, 2019. 319-335.

Waddington, Christopher. "Should Translations be Assessed Holistically or through error analysis?" *Hermes, Journal of Linguistics* 26 (2001): 15-37.

Way, Andy. "Quality Expectations of Machine Translation." *Translation Quality Assessment: From Principles to Practice*. Ed. Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty. Springer, 2018. 159-178.

White, John S. and Theresa O'Connell. "Evaluation in the ARPA Machine Translation Program: 1993 Methodology." *Proceedings of the ARPA HLT Workshop*. Ed. Madeleine Bates. Plainsboro, NJ: Association for Computational Linguistics, 1994. 135-140. <https://doi.org/10.3115/1075812.1075840>.

White, John S., O'Connell, Theresa and O'Mara, Francis E. "Evaluation Methodologies in the ARPA Machine Translation Initiative." *Proceedings of AIPA95*. Automatic Information Processing Association Steering Group,

1995. <https://www.aclweb.org/anthology/H94-1024.pdf>. Accessed 11 September 2024.

Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Quoc, Le V. and Norouzi Mohammad. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv* preprint arXiv:1609.08144, 2016.

Yuan, Zheng and Ted Briscoe. "Grammatical error correction using neural machine translation." *Conference Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. Kevin Knight, Ani Nenkova, Owen Rambow. San Diego, California: Association for Computational Linguistics, 2016. 380-386. <http://doi.org/10.18653/v1/N16-1042>.

ZUSAMMENFASSUNG: BEWERTUNG DER QUALITÄT DER MASCHINELLEN ÜBERSETZUNG ADMINISTRATIVER TEXTE AUS DEM KROATISCHEN INS ENGLISCHE

Maschinelle Übersetzung, einst eine fehlerhafte Methode, ist inzwischen deutlich besser geworden und wird kontinuierlich optimiert. Eine große Herausforderung besteht darin, die Qualität der maschinellen Übersetzungen zu bewerten. In diesem Beitrag wird eine Triangulation von Methoden zur Bewertung der Übersetzungsqualität vorgenommen: automatische und menschliche Beurteilung, korpusbasierte Analyse und Analyse der Übersetzungsfehler. Dabei wurden administrative Texte aus dem Kroatischen ins Englische in Betracht gezogen. Ziel ist es, einen umfassenden Überblick über die Qualität dieser Übersetzungen zu bekommen, die Ergebnisse der verschiedenen Bewertungsmethoden zu vergleichen und die Bereiche mit den größten Diskrepanzen in maschineller Übersetzungen zu identifizieren. Die Übersetzungen wurden mit *Google Translate* (GT) erstellt, das auf dem derzeit vorherrschenden neuronalen Modell basiert, welches Übersetzungsfehler im Vergleich zum phrasenbasierten Modell signifikant reduziert. Es wird erwartet, dass dieses Modell in der Lage ist, Probleme wie Kongruenz und Flexion besser zu lösen als andere Systeme. Dies ist insbesondere für morphologisch reiche Sprachen wie Kroatisch von Bedeutung. Obwohl die Literatur zur maschinellen Übersetzung und deren Bewertung umfangreich ist, soll diese Arbeit zur Forschung der unterrepräsentierten slawischen Sprache beitragen.

Schlüsselwörter: maschinelle Übersetzung, Bewertung der Übersetzungsqualität, administrative Texte, Kroatisch, Englisch