DETECTION OF MACHINE-TRANSLATED CROATIAN–ENGLISH PHRASES: DO TRANSLATORS AGREE IN THEIR ASSESSMENT?

Mirjana Borucinsky
Margareta Čanžar

Faculty of Maritime Studies, University of Rijeka
Faculty of Humanities and Social Sciences, University of Rijeka
mirjana.borucinsky@uniri.hr
margaretacanzar@gmail.com

UDK: 81'25:004.42
DOI: 10.15291/csi.5069
Izvorni znanstveni članak
Primljen: 14. 10. 2025.
Prihvaćen za tisak: 26. 1. 2026.

This study examines the extent to which machine-translated and human-translated Croatian–English phrases can be distinguished by expert evaluators by analysing inter-rater agreement. For this purpose, a parallel corpus was compiled from four Croatian research articles in the fields of Education and Psychology, together with their human translations and machine translations produced using the online tool onlinedoctranslator.com. Two professional translators, native speakers of Croatian with extensive translation experience, were asked to classify selected English translations of Croatian phrases as either machine-translated or human-translated. The analysis shows a raw inter-rater agreement of 75 % for both machine-translated and human-translated phrase sets, indicating a moderate level of agreement between the raters. These results suggest that even expert translators do not consistently reach the same conclusions when judging the origin of short, decontextualised translated phrases. The findings support the view that, at the phrase level, machine-translated output is increasingly difficult to distinguish from human translation on the basis of surface linguistic features alone.

Keywords:
Croatian–English, human evaluation, inter-rater agreement, machine translation

1. Introduction

With the increasing integration of MT systems in everyday applications, assessing their quality and effectiveness has become critical, particularly for specific language pairs. As MT becomes increasingly fluent and natural, the distinction between machine-generated and human-produced translations is no longer self-evident. This development has raised new research questions within translation studies, shifting attention from traditional notions of translation quality to issues such as perception, detectability, and rater agreement.

Work done by Chichirau, Noord and Toral (2023) shows that advances in neural machine translation have made it increasingly difficult to distinguish machine-generated translations from human translations, even for classifiers trained specifically for this task. Using monolingual and multilingual language-model-based classifiers in multilingual scenarios, the authors demonstrate that while reliable discrimination is still possible, performance depends heavily on access to source-language information, document-level context, and multilingual training data. This growing indistinguishability is highly relevant for various NLP applications, corpus construction, and translation quality assessment (TQA), where the ability to identify machine-translated content is crucial for ensuring data reliability, avoiding evaluation bias, and maintaining the integrity of linguistic resources.

Rather than assessing translation quality directly, this study focuses on the ability of human evaluators to identify whether a translated phrase was produced by a machine or a human translator. Specifically, the paper investigates the extent to which professional translators agree in their judgments when classifying Croatian–English translated phrases as MT or HT. By applying inter-rater agreement measures, the study aims to provide an initial, exploratory contribution to research on MT detectability for a relatively under-researched language pair. The limitations of MT are especially pronounced for languages that are underrepresented or classified as low-resourced, such as Croatian. Due to the relatively small amount of available linguistic data, developing high-quality MT systems for such languages presents significant challenges. While MT systems have advanced considerably for widely spoken languages like English, German, or Spanish, this is not the case for under-resourced languages such as Croatian.

This paper builds on earlier work that examined acceptability and perception of MT among native and non-native speakers of English. The present study adopts a different methodological perspective by focusing exclusively on agreement between expert raters.

2. Theoretical background and previous work

Human translation has long been regarded as the gold standard in the translation industry due to its accuracy, cultural sensitivity, capacity to distinguish between various contexts and to deal with polysemy, connotations, implicatures, and similar phenomena. However, human translation is often costly and time-consuming. In response to these challenges, machine translation (MT) systems have emerged as a solution to facilitate communication and the transfer of meaning across languages. The term machine translation refers to the use of computers to automate (parts of) the process of translating from one language to another (cf. Dunđer 2021). While MT has made significant advancements since its first serious and large-scale implementation, it remains limited in its ability to replace human translators entirely, particularly when it comes to more complex concepts, creativity, and specific types of texts.

One of the key points of using MT systems are realistic expectations, and closely related to that is the evaluation of MT systems. "However, there is no general agreement on the assessment of translation quality, as there are no standardised criteria for (translation quality) assessment and the process itself is a complex one involving both linguistic and extra-linguistic aspects" (cf. Borucinsky, Kegalj and Zagorec 2025: 196). MT evaluation can be performed automatically, semi-automatically, and by human evaluators. For a more detailed overview of MT evaluation methods, cf. Borucinsky, Kegalj and Zagorec (2025).

It should be pointed out that human evaluation comes with its own set of challenges, as it is both time-consuming and expensive, and often subject to individual bias. The subjective nature of human judgement introduces variability into the evaluation process, making it difficult to establish a universally agreed-upon metric for what constitutes a good machine translation. Despite these challenges, human input is essential in improving MT systems. In TQA, raters who can rate individually, in groups, or in a crowd are used (Castilho et al. 2018). Ideally, more than one rater is involved in TQA tasks and the inter-rater agreement between assessors is then calculated (Chatzikoumi 2020). Inter-rater agreement plays an important role in human evaluation studies, as it provides an indication of how consistently evaluators apply given criteria. Previous research has demonstrated that even expert raters may diverge in their assessments, especially when evaluation tasks involve subjective judgments or binary classifications. Measures such as Cohen's kappa and Gwet's AC1 have therefore been widely adopted to account for chance agreement and to provide a more robust estimate of reliability. The challenge remains to develop evaluation frameworks that can effectively detect differences in quality between machine-generated translations and human-pro-

duced gold standard translations, while also minimising the subjectivity inherent in human judgement.

Despite these complexities, human translators play a crucial role in improving MT systems by providing feedback that helps refine both the systems themselves and the metrics used to evaluate them. Seljan, Brkić and Kučiš (2011: 331) point out that "evaluation from the user's perspective helps producers, but also examines the translation problems". They also add that evaluation of MT systems is usually performed for widely spoken languages, which means that under-resourced languages are neglected.

Previous research that has examined machine translation for Croatian across diverse domains focused on industry-specific texts (Dunđer 2020), business correspondence and speech-to-text translation pipelines (Seljan and Dunđer 2014), sociological-philosophical texts (Seljan and Dunđer 2015a), as well as literary language such as poetry (Dunđer et al. 2020; Seljan et al. 2020). These studies consistently show that domain-adapted MT systems outperform general-purpose systems, even when trained on relatively small in-domain corpora, and that acceptable levels of adequacy can be achieved in specialised domains. However, evaluations, particularly those involving human assessment, reveal persistent limitations in fluency, stylistic adequacy, and semantic precision, especially for creative and abstract text types such as poetry (Seljan et al. 2020). Moreover, several studies highlight a discrepancy between automatic evaluation metrics and human judgments of quality (Seljan and Dunđer 2015a; Seljan et al. 2015b), indicating that widely used automatic metrics fail to capture domain-specific, pragmatic, and stylistic aspects of translation quality. This shows that there is a research gap in machine translation evaluation, namely the absence of robust, domain-sensitive, and language-specific evaluation frameworks that combine human assessment with improved automatic methods, particularly for low-resource languages such as Croatian. In the context of Croatian–English machine translation, existing studies have primarily focused on automatic metrics and error analysis. For example, Borucinsky, Kegalj and Zagorec (2025) evaluate the quality of Croatian-to-English machine translation of administrative texts using a triangulated approach that combines corpus-based analysis, automatic metrics (BLEU and WER), error analysis tools, and human evaluation. Their results show that while Google Neural Machine Translation produces generally comprehensible and relatively fluent translations (mean fluency score 3.8/5), adequacy remains lower (3.5/5), with significant deviations from human reference translations reflected in a low BLEU score (0.35) and high WER (64.8 %). The study highlights persistent issues related to stylistic unnaturalness, terminological inconsistency, and semantic adequacy, demonstrating that automatic metrics alone fail to capture key quality aspects and reinforcing the need for combined human and automatic evalua-

tion, especially for morphologically rich languages such as Croatian. However, relatively little attention has been devoted to the question of MT detectability, and from the perspective of expert translators. This study addresses this gap by examining whether professional translators agree in their judgments when asked to classify translations as machine-generated or human-produced without access to source context.

This paper builds on findings from earlier work by one of the authors. In an anonymous survey conducted via Google Forms in June 2023, Čanžar (2024) examined the ability of 20 native and 20 non-native speakers of English to distinguish between machine- and human-translated phrases, as well as native speakers' evaluations of the acceptability of machine translation. Although a slight majority of native speakers (54 %) rated human-translated phrases as more acceptable, the relatively small difference suggests that machine-translated phrases are often perceived as sufficiently acceptable. This finding indicates a narrowing perceived quality gap between human and machine translation, at least at the phrase level. With regard to non-native speakers, the results show only a marginal advantage in identifying machine-translated phrases, as approximately 52.38 % of respondents succeeded in doing so, while 47.6 % were unable to distinguish between the two. This near-even distribution suggests that non-native speakers' ability to reliably differentiate between machine- and human-translated phrases is limited, raising questions about the robustness of such judgments and the factors influencing translation perception.

3. Research question

Based on the identified research gap in the evaluation of MT for the language pair Croatian-English and on the previous findings described in the section above, this study addresses the following research question:

To what extent do professional translators agree when classifying Croatian–English translated phrases as machine-translated or human-translated?

This study is explicitly limited to examining agreement patterns among raters.

3.1. Methodology

3.2. Materials

The initial phase of the study involved the compilation of corpora. For that purpose, three corpora were compiled: 1) a corpus consisting of four original texts written in Croatian counting 23,019 tokens, 750 sentences; 2) a corpus comprising the same

four texts translated from Croatian into English by a professional human translator with a total number of 28,661 tokens and 776 sentences; and 3) a corpus consisting of the four texts translated from Croatian into English using a neural machine translation system with 26,451 tokens and 757 sentences. The Croatian source texts were retrieved from the Portal of Croatian Scientific and Professional Journals, specifically from the 2022 issue of the Croatian Journal of Education. The selected articles address topics within the fields of Education and Psychology, including research on students' attitudes towards reading and assigned reading, gifted students with disabilities, and related issues. All texts are research articles written in a formal academic style and contain domain-specific terminology related to Education and Psychology. The machine-translated versions of the texts were produced using onlinedoctranslator.com, an online machine translation tool based on a neural MT engine, accessed via a web interface. No post-editing was applied.

3.3. Raters

Two independent raters participated in the study. Both are professional translators, native speakers of Croatian, with graduate degrees in English Studies and more than ten years of professional translation experience. The homogeneity of the rater group was intentional in this pilot phase, with the aim of minimising variation due to language proficiency.

3.4. Procedure

Sketch Engine (Kilgarriff et. al 2014) was used to extract keywords and terms from each corpus that will be used in the evaluation stage later on. The terms are presented in Table 1 and are also published in Čanžar (2024).

Table 1. Croatian phrases and their MT and HT English equivalents.

CROATIAN	MT	HT
daroviti učenik	gifted student	gifted student
lektira	school reading	reading assignment
učenik s teškoćama učenja	student with learning disabilities	student with learning disabilities
čitalačka pismenost	reading literacy	reading literacy

CROATIAN	MT	HT
darovita djeca	gifted children	gifted children
teškoća učenja	learning difficulty	learning disability
nastava lektire	teaching of reading	reading assignment class
teorija odgoja	theory of education	theory of upbringing
daroviti učenik s ADHD-om	gifted student with ADHD	gifted student with ADHD
književni tekst	literary text	literary text
školsko postignuće	school achievement	school achievement
procjena darovitosti	assessment of giftedness	assessment of giftedness
nastava lektire	reading class	reading class
kognitivna sposobnost	cognitive ability	cognitive ability
lektirni naslov	reading title	reading assignment title
nedaroviti učenik	non-gifted student	non-gifted student
program za darovite	program for gifted students	gifted program
samostalno čitanje	independent reading	independent reading
književno djelo	literary work	literary work
teškoća učenja	learning disability	learning difficulty
obrazovna praksa	educational practice	educational practice
kritičko čitanje	critical reading	critical reading
razumijevanje pročitanoga	comprehension of the read	reading comprehension
temeljna obrazovna kompetencija	fundamental educational competence	basic educational competence
teškoća u verbalnoj obradi	difficulty in verbal processing	difficulty in verbal processing
motrenje razumijevanja pročitanoga	observing the comprehension of the read	monitoring reading comprehension
neatraktivni lektirni naslov	unattractive reading title	unattractive reading assignment title
učenik s teškoćama čitanja	student with reading difficulties	student with reading disabilities
lektirni naslov	reading title	literary title

CROATIAN	MT	HT
čitalačka kompetencija	reading competency	reading competency
Okvir nacionalnog kurikuluma	Framework of the National Curriculum	National Curriculum Framework
pedagog	pedagogue	pedagogue
odgoj	education	upbringing
gimnazija	high school	grammar school
obrazovni sustav	educational system	educational system
konotativno čitanje	connotative reading	connotative reading
samoregulacija učenja	self-regulation of learning	self-regulated learning
strategija učenja	learning strategy	learning strategy
strategija poučavanja	teaching strategy	teaching strategy
učeničko zalaganje	student efforts	students' commitment
procjena učenika	students' assessment	pupils' assessment
odgojno-obrazovni ciljevi	educational goals	educational aims
ishodi	outcomes	outcomes
odgajatelj	educator	educator
odgojno-obrazovni plan	educational plan	education plan
dvostruko izuzetni učenici	doubly exceptional students	twice-exceptional students
završan razred	final grade	final year
roditeljski sastanak	parent meeting	parent-teacher conference

From this list presented in Table 1, 24 Croatian phrases that had different MT and HT translations were selected and presented to the raters in tabular form. For each phrase, raters were instructed to classify the translation either as machine translation (MT) or human translation (HT). No training session was conducted prior to the task, and raters completed the classification independently. The raters were given the following instructions:

> *The following table shows 24 Croatian ST phrases from the field of education and their English equivalents. Next to each phrase in English (column D) write* ***MT****, if you judge the phrase to have been translated by a machine and* ***HT*** *if you judge it to have been translated by a human translator.*

The raters were not told that the set contained equal numbers of MT and HT items. Inter-rater agreement was calculated for machine-translated and human-translated phrases using raw agreement, Cohen's kappa (κ), and Gwet's AC1. These measures were selected to account for chance agreement and to provide complementary perspectives on rater consistency.

3.5. Inter-rater agreement

Once the raters had judged the phrases either as MT or HT, we calculated the inter-rater agreement, which is crucial for ensuring the reliability and validity of MT assessments.

First, we calculated the raw agreement between the two raters:

Raw agreement is a useful first approximation of the reliability of the rating. However, we need to consider the fact that some agreement between the raters would have been achieved even if both raters coded the cases randomly (Brezina 2021). More sophisticated inter-rater agreement measures such as Cohen's Kappa (κ) or Gwet's (2014) AC1 therefore take this agreement by chance into account and subtract it from the raw agreement (cf. Brezina 2021). As Brezina (2021) points out, Cohen's κ has been traditionally used for nominal variables; however, recent studies recommend Gwet's AC1 statistic. Both Cohen's κ and Gwet's AC1 are based on the same equation, but they differ, however, in the estimation of the agreement by chance. Fleiss κ is commonly used for calculating agreement among multiple raters, so it was not applied to this study.

4. Results

Table 2 shows how raters judged the phrases.

Table 2. Rater evaluation of the selected Croatian-English phrases.

No.	Croatian	English (MT)	Rater 1	Rater 2	Cases of agreement	English (HT)	Rater 1	Rater 2	Cases of agreement
1.	lektira	school reading	MT	MT	AGREE	reading assignment	HT	HT	AGREE

No.	Croatian	English (MT)	Rater 1	Rater 2	Cases of agreement	English (HT)	Rater 1	Rater 2	Cases of agreement
2.	nastava lektire	teaching of reading	MT	MT	AGREE	reading assignment class	HT	HT	AGREE
3.	teorija odgoja	theory of education	HT	HT	AGREE	theory of upbringing	MT	MT	AGREE
4.	lektirni naslov	reading title	MT	MT	AGREE	reading assignment title	HT	HT	AGREE
5.	program za darovite	program for gifted students	MT	HT	DISAGREE	gifted program	HT	MT	DISAGREE
6.	teškoća učenja	learning disability	HT	HT	AGREE	learning difficulty	MT	MT	AGREE
7.	razumijevanje pročitanoga	comprehension of the read	MT	MT	AGREE	reading comprehension	HT	HT	AGREE
8.	temeljna obrazovna kompetencija	fundamental educational competence	MT	MT	AGREE	basic educational competence	HT	HT	AGREE
9.	motrenje razumijevanja pročitanoga	observing the comprehension of the read	MT	MT	AGREE	monitoring reading comprehension	HT	HT	AGREE
10.	neatraktivni lektirni naslov	unattractive reading title	MT	MT	AGREE	unattractive reading assignment title	HT	HT	AGREE

No.	Croatian	English (MT)	Rater 1	Rater 2	Cases of agreement	English (HT)	Rater 1	Rater 2	Cases of agreement
11.	učenik s teškoćama čitanja	student with reading difficulties	MT	MT	AGREE	student with reading disabilities	HT	HT	AGREE
12.	lektirni naslov	reading title	MT	HT	DISAGREE	literary title	HT	MT	DISAGREE
13.	Okvir nacionalnog kurikuluma	Framework of the National Curriculum	MT	MT	AGREE	National Curriculum Framework	HT	HT	AGREE
14.	pedagog	pedagogue	MT	MT	AGREE	pedagogue	MT	MT	AGREE
15.	odgoj	education	HT	HT	AGREE	upbringing	MT	MT	AGREE
16.	gimnazija	high school	MT	HT	DISAGREE	grammar school	HT	MT	DISAGREE
17.	samoregulacija učenja	self-regulation of learning	MT	MT	AGREE	self-regulated learning	HT	HT	AGREE
18.	učeničko zalaganje	student efforts	HT	MT	DISAGREE	students' commitment	MT	HT	DISAGREE
19.	procjena učenika	students' assessment	HT	HT	AGREE	pupils' assessment	MT	MT	AGREE
20.	odgojno-obrazovni ciljevi	educational goals	MT	HT	DISAGREE	educational aims	HT	MT	DISAGREE
21.	odgojno-obrazovni plan	educational plan	MT	HT	DISAGREE	education plan	HT	MT	DISAGREE

No.	Croatian	English (MT)	Rater 1	Rater 2	Cases of agreement	English (HT)	Rater 1	Rater 2	Cases of agreement
22.	dvostruko izuzetni učenici	doubly exceptional students	MT	MT	AGREE	twice-exceptional students	HT	HT	AGREE
23.	završni razred	final grade	MT	MT	AGREE	final year	HT	HT	AGREE
24.	roditeljski sastanak	parent meeting	MT	MT	AGREE	parent-teacher conference	HT	HT	AGREE

As can be seen from Table 2, when it comes to determining whether the phrases presented in column A are translated by a machine, rater 1 identified 19 out of 24 phrases correctly (79 %), rater 2 identified 15 out of 24 phrases (63 %) correctly as machine translated. In the second set, rater 1 identified 18 out of 24 (75 %) phrases correctly as translated by a human translator, and rater 2 identified 14 out of 24 (58 %) phrases correctly as human-translated. The raters agreed in 18 cases for both sets and disagreed in 6 cases. Hence, the raw agreement for both the machine-translated and human-translated phrases is 75 % (18/24; 18/24), indicating a consistent level of agreement among the raters

Cohen's Kappa (κ) for both the machine-translated and human-translated phrases is 0.5, whereas there is a slight difference in Gwet's AC1, suggesting moderate agreement for machine translations (AC1 = 0.574) and slightly lower agreement for human translations (AC1 = 0.550). The p values for machine-translated phrases (p = 0.003) and human-translated phrases (p = 0.005) indicate statistically significant results for AC1.

5. Discussion

The results show a consistent level of inter-rater agreement across both machine-translated and human-translated phrase sets. A raw agreement of 75 %, together with moderate values of Cohen's kappa (κ = 0.50) and Gwet's AC1 (0.574 for MT and 0.550 for HT), indicates that professional translators do not consistently reach the same conclusions when classifying short Croatian–English translated phrases. Notably, agreement levels were almost identical for machine-translated and human-trans-

lated phrases, which suggests that uncertainty is not limited to machine translation but is also present when evaluating human translations.

Disagreement occurred in six cases in each category. These phrases were terminologically plausible, regardless of whether they were produced by a machine or a human translator. This finding supports the view that surface terminological plausibility, which is often associated with professional translation quality, can no longer be considered a reliable indicator for detecting machine translation at the phrase level.

A closer examination of the disagreement cases shows that ambiguity often arises when both machine and human translations provide acceptable domain-specific solutions. For example, the Croatian term *gimnazija* was translated as high school in the machine-translated version and as grammar school in the human-translated version. Both translations are fluent and broadly acceptable in English, but they reflect different cultural mappings, namely generalisation versus institution-specific equivalence. Such cases are difficult to classify because neither option contains features that would clearly indicate machine or human authorship. Instead, successful classification depends on assumptions about the translator's familiarity with the educational systems of both cultures. This example also illustrates the complexity of the translation process, which involves not only linguistic competence but also extensive cultural and contextual knowledge.

A similar pattern can be observed in the translation of the phrase *program za darovite*, rendered as *program for gifted students* (MT) and *gifted program* (HT). The machine-translated version is longer and more explicit, while the human translation is more compact and idiomatic. However, contemporary MT systems increasingly produce such idiomatic noun compounds, and human translators may also opt for more explicit formulations depending on the intended audience. As a result, stylistic differences between MT and HT outputs are becoming less pronounced, which may contribute to rater disagreement.

Another illustrative example is *učeničko zalaganje*, translated as *student efforts* (MT) and *students' commitment* (HT). Both translations are acceptable in academic discourse, but they differ slightly in meaning, with the former emphasising observable behaviour and the latter focusing more on motivational or attitudinal aspects. Without broader context, such subtle semantic distinctions are difficult to use as a basis for origin detection, as both variants are terminologically appropriate and do not contain obvious errors.

In contrast, items that elicited agreement often included translations with atypical or pragmatically marked phrasing. Examples such as *comprehension of the read* and *observing the comprehension of the read* are grammatically well-formed, but

they sound unnatural in English academic usage. These cases suggest that raters may rely on noticeable local irregularities, such as collocational unnaturalness, awkward nominalisations, or overly literal renderings, rather than on systematic indicators of translation origin.

Clear agreement was also observed in cases involving conventionalised institutional terminology. For instance, *roditeljski sastanak* was translated as *parent meeting* in the machine-translated version and as *parent-teacher conference* in the human translation. Although parent meeting is grammatically correct, it lacks the institutional specificity typical of educational discourse in English. The human translation, by contrast, corresponds to an established term and provides clearer pragmatic anchoring. Such cases indicate that raters consider not only fluency, but also culturally embedded lexical conventions when determining whether a translation was produced by a machine or a human.

The findings further highlight the importance of distinguishing between detectability and translation quality. A translation may be difficult to identify as machine-generated while still containing subtle issues related to semantic precision, pragmatic appropriateness, or discourse coherence. These aspects cannot be captured through phrase-level classification alone. At the same time, human translations in specialised domains may adopt literal or conventionalised formulations that resemble machine output in terms of brevity or compositional structure. For this reason, moderate agreement in detection tasks should not be interpreted as evidence that machine translation has reached parity with human translation quality. Rather, it suggests that surface-level cues used for attribution are becoming less reliable.

The results also point to limitations inherent in binary detection tasks used in human evaluation studies. While inter-rater agreement statistics provide information about reliability, they do not explain why raters arrive at different judgments. Human evaluation is influenced by multiple factors, including professional experience, implicit norms, exposure to machine translation, expectations regarding linguistic cues associated with MT, and familiarity with the domain.

These observations are consistent with broader critiques of MT evaluation practices that rely on isolated units, simplified rating schemes, and limited rater training. In contexts where machine translation output is already fluent, evaluation results may be increasingly affected by individual interpretation, bias, or random variation. This highlights the need for clearer operational definitions and evaluation designs that separate questions of translation origin from questions of translation quality.

The present study deliberately focuses on short, decontextualised phrases, which allows for the examination of lexical and terminological choices. However, such an

approach also removes information that becomes available at higher levels of textual organisation. At the sentence level, evaluators can assess syntactic integration and local coherence, while at the document level they can observe terminological consistency, discourse relations, register stability, and long-range cohesion. Many differences between machine and human translation, particularly those related to discourse and pragmatics, may therefore remain undetected in phrase-level tasks.

This leads to two important implications. First, moderate agreement at the phrase level may underestimate raters' ability to identify translation origin when more context is available. Second, human-like performance in short segments does not necessarily imply human-like performance at the document level, where consistency and discourse management play a crucial role. Future research should therefore compare agreement patterns across different levels of context and examine whether access to broader textual information leads to higher agreement.

Finally, the disagreement cases demonstrate the value of qualitative follow-up analyses. Approaches such as error typology or linguistic feature analysis could help identify which translation properties contribute most to uncertainty, including culturally specific institutional terms (e.g., *gimnazija*), competing acceptable collocations (e.g., student *efforts* vs. *students' commitment*), or differing degrees of explicitation (e.g., *program for gifted students* vs. *gifted program*). Combining quantitative agreement measures with qualitative analysis would improve interpretability and provide a more nuanced understanding of translation evaluation.

6. Conclusion

The aim of this study was to find out to what extent professional translators agree when classifying Croatian–English translated phrases as machine-translated or human-translated. To answer this research question, inter-rater agreement was calculated. The results show moderate level of agreement suggesting that phrase-level detectability of machine translation is limited, even among expert evaluators.

The growing difficulty in distinguishing between machine and human translations reflects significant advances in machine translation technology, particularly with regard to terminological adequacy in specialised domains. However, the results should not be interpreted as evidence that machine translation is equivalent in quality to human translation. Instead, they point to a distinction between perceptual detectability and underlying translation quality, reinforcing the view that judgments about translation origin cannot replace comprehensive quality evaluation.

The study provides empirical evidence from an under-researched language pair and highlights the importance of inter-rater agreement analysis in research on machine translation detectability. In this context, the study contributes to ongoing discussions on human evaluation methodologies in machine translation research. The observed moderate inter-rater agreement supports existing critiques of overly simplified evaluation approaches and underlines the importance of carefully designed, multi-dimensional assessment frameworks (cf. Toral 2020).

Although limited in scope, this pilot study provides initial insights into the detectability of machine translation for the Croatian–English language pair and highlights the relevance of inter-rater agreement analysis in human evaluation research. Future studies that combine agreement measures with qualitative and context-sensitive analyses are likely to contribute to the development of more robust and informative translation evaluation practices, particularly for under-resourced languages.

As a pilot study, this research has several limitations. The number of raters was small, and the rater group was relatively homogeneous. The dataset was limited to short phrases from a single domain and presented without context, which does not reflect real-world translation conditions. Future studies should include a larger and more diverse group of raters (such as crowd workers), expand the dataset, incorporate contextualised text, and combine detection tasks with qualitative and quantitative quality assessment measures.

Declarations

The authors declare that this paper has not been published previously and is not under consideration elsewhere. Generative AI tools were used for linguistic refinement only; the authors take full responsibility for the content.

References

Borucinsky, Mirjana, Jana Kegalj and Dario Zagorec. 2025. "Translation Quality Evaluation of Croatian-to-English Machine-Translated Administrative Texts". *Journal of Language and Literary Studies* 51: 193–211.

Brezina, Vaclav. 2021. *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

Castilho, Sheila, Stephen Doherty, Federico Gaspari and Joss Moorkens. 2018. "Approaches to Human and Machine Translation Quality Assessment". In Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (Eds.). *Translation Quality Assessment: From Principles to Practice*. Heidelberg: Springer: 9–38.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone and Maria Gialama. 2017. "A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators". In *Proceedings of Machine Translation Summit XVI: Research Track*. Nagoya: Machine Translation Summit: 116–131.

Čanžar, Margareta. 2024. *Human Evaluation of Machine-Translated Texts*. Diplomski rad (MA thesis). Rijeka: University of Rijeka, Faculty of Humanities and Social Sciences.

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

Chichirau, Malina, Rik van Noord and Antonio Toral. 2023. "Automatic Discrimination of Human and Neural Machine Translation in Multilingual Scenarios". In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Tampere: European Association for Machine Translation: 217–226.

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

Dunđer, Ivan. 2021. "Pregled razvoja tehnologije automatskog strojnog prevođenja". *Polytechnic & Design* 9 (2): 90–100.

Dunđer, Ivan, Sanja Seljan and Marko Pavlovski. 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)*. Rijeka: MIPRO: 1034–1039.

Gwet, Kilem Li. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD: Advanced Analytics.

Online Doc Translator. n. d. "Online Doc Translator". (1 June 2024). *(online source)*

Seljan, Sanja and Ivan Dunđer. 2014. "Combined Automatic Speech Recognition

and Machine Translation in Business Correspondence Domain for English–Croatian". *International Journal of Computer, Information, Systems and Control Engineering* 8: 1069–1075.

Seljan, Sanja and Ivan Dunđer. 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)* 2: 128–131.

Seljan, Sanja and Ivan Dunđer. 2015b. "Machine Translation and Automatic Evaluation of English/Russian–Croatian". In *Proceedings of the International Conference "Corpus Linguistics – 2015" (CORPORA 2015)*: 72–79.

Seljan, Sanja, Marija Brkić and Vlasta Kučiš. 2011. "Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs". In Clive Billenness, Annette Hemera, Vladimir Mateljan, Mihaela Stančić, Hrvoje Stančić and Sanja Seljan (Eds.). *The Future of Information Sciences—Information Sciences and e-Society*. Zagreb: Department of Information Sciences: 331–344.

Seljan, Sanja, Marko Tucaković and Ivan Dunđer. 2015. "Human Evaluation of Online Machine Translation Services for English/Russian–Croatian". In *WorldCIST'15 – 3rd World Conference on Information Systems and Technologies*. *Advances in Intelligent Systems and Computing*: 1089–1098.

Seljan, Sanja, Ivan Dunđer and Angelina Gašpar. 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)*: 1329–1334.

Seljan, Sanja, Nikolina Škof Erdelja, Vlasta Kučiš, Ivan Dunđer and Mirjana Pejić Bach. 2020. "Quality Assurance in Computer-Assisted Translation in Business Environments". In *Natural Language Processing for Global and Local Business*. Hershey, PA: IGI Global: 247-270.

Seljan, Sanja, Ivan Dunđer and Marko Pavlovski. 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)*. Rijeka: MIPRO: 1040–1045.

Toral, Antonio. 2020. "Reassessing claims of human parity and super-human performance in machine translation at WMT 2019". In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*: 185–194.

Prepoznavanje strojno prevedenih engleskih izraza na hrvatski jezik: slažu li se prevoditelji u procjeni?

Mirjana Borucinsky
Margareta Čanžar

Sažetak

U ovom se istraživanju ispituje je li moguće razlikovati strojni prijevod od prijevoda koji je izradio čovjek na primjeru prijevoda hrvatskih izraza na engleski jezik. Analizira se u kojoj mjeri se stručni ocjenjivači slažu u procjeni prijevoda kao strojnoga prijevoda ili prijevoda koji je izradio čovjek. U tu svrhu sastavljen je korpus znanstvenih radova iz područja obrazovanja i psihologije, koji se sastoji od triju potkorpusa: 1. korpus izvornih tekstova na hrvatskome, 2. korpus prijevoda izvornih tekstova na engleski jezik koje je preveo čovjek, 3. korpus strojno prevedenih hrvatskih tekstova na engleski. U zadatku prepoznavanja dva profesionalna prevoditelja, izvorni govornici hrvatskoga, razvrstavala su engleske prijevode odabranih hrvatskih izraza kao strojne ili one koje je preveo čovjek. Slaganje među ocjenjivačima iznosilo je 75 % uz umjerenu podudarnost. Rezultati pokazuju da se prevoditelji ne slažu u potpunosti pri procjeni, što podupire stav da postaje teže razlikovati strojno prevedene izraze od prijevoda koje je izradio čovjek.

Ključne riječi:
hrvatsko-engleski prijevod, prepoznavanje strojnog prijevoda, procjena, suglasnost među ocjenjivačima