The Challenge of Translating Culture-Specific Items: Evaluating MT and LLMs Compared to Human Translators

Bojana Budimir
University of Belgrade, Falulty of Philology / Studentski trg 3
boj.budimir@gmail.com

Abstract

We evaluate state-of-the-art Large Language Models (LLM's) ChatGPT-4o, Gemini 1.5 Flash, and Google Translate, by focusing on the translation of culture-specific items (CSIs) between an underrepresented language pair: the Flemish variant of Dutch and Serbian. Using a corpus derived from three Flemish novels we analyze CSIs in three cultural domains: Material Culture, Proper Names, and Social Culture. Translation strategies are examined on a spectrum that goes from conservation to substitution. Quantitative analysis explores strategy distribution, while qualitative analysis investigates errors, linguistic accuracy, and cultural adaptation. Despite advancements, models struggle to balance cultural nuances with understandability for the target readers. Gemini aligns most closely with human translation strategies, while Google Translate shows significant limitations. These findings underscore the challenges of translating CSIs—particularly Proper Names—in low-resource languages and offer insights for improving machine translation models.

1 Introduction

Recent advancements in machine translation (MT) have significantly enhanced its quality and broadened its applicability, even in the domain of literary translation, an area often considered resistant to automation due to its reliance on nuance, creativity, and cultural context. Existing studies have reported varying levels of success for MT tools, with accuracy rates ranging from 44% (Fonteyne et al., 2020) to 20% (Webster et al., 2020). Several researchers have investigated the potential of machine translators pre-trained on literary texts (Matusov (2019); Kuzman et al. (2019), showing that tailored systems can improve automatic evaluation metrics for prose translations when compared to baseline models.

Beyond improvements in output quality, recent scholarship has also investigated how MT and computer-assisted translation (CAT) tools might be adapted to support the specific demands of literary translation. Hadley (2023), for instance, argues that these technologies can serve as productivity aids rather than replacements for human creativity. He identifies a range of functionalities, such as sentence length control, rhyme pattern identification, and syllable counting, that could be incorporated into CAT tools to assist translators working with poetry or stylistically marked texts. Similarly, Kolb and Miller (2022) provide empirical evidence that the tool PunCAT, designed to support the translation of puns, can stimulate and broaden the translator's pool of creative solutions, thus enhancing problem-solving in areas of high linguistic density and ambiguity.

In parallel, the impact of MT on translator creativity and reader experience has also been part of several studies. Their findings revealed that while human translations exhibit a higher degree of creativity, there is no statistically significant difference in the overall reading experience between human and post-edited machine translations (Guerberof-Arenas and Toral, 2020; Guerberof-Arenas and Toral, 2022). Furthermore, large language models (LLMs) have introduced tools capable of tackling complex tasks, such as creative writing (Gomez-Rodriguez and Williams, 2023) and poetry (Porter and Machery, 2024), expanding their potential applications.

Proceedings of Machine Translation Summit XX Volume 1, pages 455–467, June 23-27, 2025

The growing capabilities of MT tools and LLMs have led to their widespread use in various fields, including literary translation. According to a recent study conducted by the European Council of Literary Translators' Associations (CEATL, 2024), more than half (54%) of literary translators from 34 member countries occasionally use MT tools in their work, primarily for translating short passages or sentences (62%). Notably, some publishers have begun offering literary translators assignments to revise machine-translated texts, with approximately 7% of translators in Serbia reporting such requests (CEATL, 2024). This trend is further supported by publishers' emerging plans to release books translated entirely by artificial intelligence (Creamer, 2024). This trend emphasizes the growing importance of evaluating MT tools from a practical, user-oriented perspective.

Despite these advancements, one of the greatest challenges in both human and machine translation remains the accurate handling of culture-specific items (CSIs). These elements are particularly complex due to their dual role: they function within the narrative structure while carrying connotations and references to concepts often absent in the target culture. This duality makes CSIs a critical focus for evaluating MT systems. Understanding how MT tools and LLMs manage culturally bound elements provides valuable insights into their performance, particularly in literary translation, where maintaining the integrity of CSIs is crucial.

This need becomes especially apparent in light of findings by Daems (2022), who who studied the use and perceived usefulness of translation technologies by Dutch literary translators. Her research shows that many literary translators consider MT and CAT tools largely inadequate for capturing essential literary features such as style, humor, irony, and metaphor, as well as broader aspects such as context and cultural background. These perceptions underscore the persistent gap between current technological capabilities and the nuanced demands of literary translation. Therefore, examining the treatment of CSIs by MT and LLM systems not only provides a means to evaluate current performance but also reveals areas in need of targeted development, contributing to the creation of more culturally aware and context-sensitive translation technologies.

This study also addresses the challenges posed by low-resource languages, such as Serbian, which lack sufficient training data for MT systems. Serbian ranks among the least technologically developed European languages, alongside Maltese, Irish, Luxembourgish, and Bosnian, as highlighted by the ELE (European Language Equality) project (Srebnik, 2023). Furthermore, in the field of machine translation research, studies on low-resource languages often focus on their pairing with dominant global languages, such as English. By examining an underrepresented language pair, this research contributes to a deeper understanding of MT performance in less-studied linguistic contexts and offers insight into existing gaps that can inform future improvements in AI tool development.

This article investigates how contemporary MT tools, including ChatGPT-4o, Gemini 1.5 Flash, and Google Translate, handle CSIs, and how their use of translation strategies compares to those of human translators. In this study, the term machine translation (MT) is used as an umbrella term encompassing both neural machine translation (NMT) systems and large language models (LLMs). These tools were chosen to facilitate a comparison between NMT and LLM-based approaches. NMT systems, such as Google Translate and DeepL, rely on large-scale parallel corpora and are expensive to develop and maintain, which limits their coverage of less-resourced language pairs. At the time of writing, Google Translate is the only major NMT service that supports translation between Dutch and Serbian, restricting access to high-quality NMT for Serbian-speaking users. In contrast, LLMs are trained on vast multilingual datasets, including monolingual and non-parallel corpora, which allows them to perform translations across a broader range of language pairs, even in low-resource scenarios. Their growing adoption by professional translators, as indicated in recent surveys such as the CEATL report (2024), further underscores their relevance to translation practice.

By analyzing strategy distribution, mistranslation rates, and error patterns across cultural categories, the study evaluates the differences between these models, identifies which tool aligns most closely

with human translation, and highlights the most common types of errors in each approach.

Related work

The term culture-specific item (CSI) was introduced by Franco Aixelá to describe elements in a text that may pose challenges for translators due to their function or connotation, particularly when the referenced phenomenon does not exist in the target culture or holds a different intertextual status in the readers' cultural framework (Aixelá, 1996). Alongside this term, translation studies have proposed a variety of other terms to describe this phenomenon, such as realia (Grit, 2010; Leppihalme, 2001), cultural words (Newmark, 1988), cultural references (Olk, 2013), and cultureme (Katan, 2009). Leppihalme (2001) defines realia as lexical elements that refer to the real world outside language, making them extralinguistic phenomena. She argues that these references can lead to what she terms a cultural bump, a situation in which the reader of the target text encounters difficulty understanding a culture-bound reference because it has no equivalent in their own cultural context (Leppihalme, 1997).

Due to their extralinguistic and culture-bound nature, CSIs require the translator to possess not only bilingual proficiency but also a high degree of bicultural competence. The successful translation of CSIs demands encyclopedic cultural knowledge as well as creativity in identifying appropriate solutions. Translators can draw upon a range of strategies and procedures to address these challenges. From a macro perspective, they may choose to preserve the original CSI, a strategy associated with foreignization, or to adapt it for the target audience through domestication (Venuti, 1995). At the micro level, various procedures exist for rendering CSIs within the text, and numerous taxonomies have been developed specifically for this purpose (Aixelá, 1996; Leppihalme, 2001; Grit, 2010; Olk, 2013). However, as Olk (2013) points out, no single taxonomy can be considered universally applicable; the selection of a particular model often depends on factors such as the objectives of the study and the language pair involved. This is also the case in the present study, where a specific taxonomy was chosen based on these methodological considerations.

The selection of an appropriate translation strategy for CSIs is influenced by numerous factors. Scholars such as Newmark (1988), Aixelá (1996), and Grit (2010) have identified patterns in the application of strategies based on the type of CSI. In addition to the inherent characteristics of CSIs, and textual features such as canonization, markedness and relevance, supratextual, textual, and intratextual parameters play a crucial role in strategy selection. Supratextual parameters include linguistic norms, reader expectations, and publisher policies, while the function of the CSI within the text is considered intertextual (Aixelá, 1996).

While much of the research on CSIs has focused on human translation, there is growing interest in how MT tools handle CSIs. Yao et al. (2024) addressed the challenges MT systems face when translating culturally specific content. They introduced a culturally aware machine translation (CAMT) parallel corpus enriched with CSI annotations and proposed a novel evaluation metric to assess translation understandability using GPT-4. This research highlights the potential of LLMs to handle complex cultural elements while revealing areas where improvements are needed. Similarly, Pudjiati et al. (2021) explored the role of post-editing in improving machine-translated CSIs from Indonesian into English. Their findings underscore the limitations of MT systems in handling figurative language and culturally nuanced terms, emphasizing the importance of human intervention in achieving semantic accuracy and cultural fidelity.

Proper names, a subset of CSIs, have also received significant attention in MT research. Hurskainen (2013) examined the challenges associated with translating proper names, highlighting the role of tagging, rule-based disambiguation, and probability measures in resolving ambiguity. This study demonstrated how linguistic and contextual rules can improve MT accuracy when handling proper names, particularly those with dual meanings or capitalization issues.

All the above-mentioned studies collectively support the present research by providing theoretical and practical insights into the complexities of CSI translation and the evolving role of MT. By focusing on low-resource

language and evaluating specific MT models, this study builds on prior work to address gaps in understanding how machine and human translators handle CSIs across cultural categories.

2 Methodology

This study is based on the research into the translation of CSIs in Flemish literature (Budimir, 2021), extending its scope to include a comparative analysis of machine translation and human translation strategies. Specifically, it investigates the translation of CSIs across three cultural categories—Material Culture (MC), Proper Names (PN), and Social Culture (SC)—as defined by Newmark (1988). Material Culture (MC) includes items such as food, drink, and towns/housing, reflecting tangible aspects of everyday life. Proper Names (PN) encompass street names, brand names, and HoReCa (hotel, restaurant, and café) names, which often require specific adaptation to the target culture. Social Culture (SC) comprises job titles, sports and games, and leisure activities, highlighting culturally embedded practices and societal roles. These categories were selected due to the clear divergence in translation strategies applied to each. As demonstrated in Budimir (2021), translators predominantly employed orthographic adaptation and literal translation when rendering PNs, thereby opting for the conservation of these elements. In contrast, description and localization were more frequently used for items related to MC and SC, where translators tended to adapt the elements to the expectations of target readers. To further support this categorization, a Chi-Square test was conducted to assess whether there were significant differences in the strategies used by the human translators within each category. The test revealed no statistically significant differences in the distribution of translation strategies between the translators ($p = 0.066$ for PN, $p = 0.256$ for MC, and $p = 0.438$ for SC). This outcome supports the assumption that it is primarily the nature of the CSI—and not the individual translator—that influences strategic choices, thereby reinforcing the relevance of these categories for analyzing patterns of translation behavior across different cultural domains.

The research adopts a mixed-method approach. Quantitative analysis examines the distribution of translation strategies employed by machine translation models and human translators, while qualitative analysis explores errors, linguistic accuracy, and cultural nuances. The methodology comprises three phases: (1) corpus formation, (2) extraction of translation equivalents, and (3) classification of translation strategies.

2.1 Corpus Formation

The corpus for this study is derived from an existing dataset of six Flemish novels. For this research, excerpts were selected from three culturally rich novels from the original corpus: *Het verdriet van België* (*The Sorrow of Belgium*) by Hugo Claus, translated into Serbian by Ivana Šćepanović and published in 2000; *De komst van Joachim Stiller* (*The Coming of Joachim Stiller*) by Hubert Lampo; and *De helaasheid der dingen* (*The Misfortunates*) by Dimitri Verhulst. The latter two were translated by Jelica Novaković-Lopušina in 1992 and 2015, respectively. These two translators are among the most productive and prominent figures working in the field of literary translation from Dutch to Serbian.

The corpus formation process began with a predefined list of 197 CSIs, from the previous research serving as the primary units of analysis. Instances of CSIs were identified within a parallel corpus organized in an Excel sheet. Sentences containing CSIs, along with preceding and following sentences, were extracted to provide contextual information. This process resulted in a corpus containing 246 sentences, 7,087 words and 43,421 characters.

Given the character limitations of machine translation models—Google Translate (5,000 characters) and ChatGPT (4,096 characters), the corpus was divided into chunks of approximately 600 words. Consistent chunks were used across all models (ChatGPT, Gemini, and Google Translate) to ensure comparability. ChatGPT produced two translation variants: ChatGPT (1), without the search option, and ChatGPT (2), with the search option. Translations were generated on November 12th 2024 using a standardized zero-shot prompt: "Translate this text from Dutch to Serbian." The use of a simple, zero-shot prompt was intentional, as the goal of the study was to evaluate the baseline performance of two LLMs and an NMT system when translating CSIs.

Strategies	Description
Repetition (R)	The CSI is kept in its original form. For Serbian, this may include adding inflectional suffixes to align with the target language's grammatical rules. For example: In de Volkskring - U *De Volkskring-u*.
Orthographic Adaptation (OA)	The CSI is adapted to reflect its pronunciation in the target language, following Serbian orthographic conventions. For example: Scheldewindeke - *Sheldevindeke.*
Combination of strategies (COM)	The CSI is either retained in its original or adapted form and supplemented with additional information, such as a classifier, an explanation integrated into the text, or a footnote. For example: De Leie - reka Leja [the river Leie].
Literal Translation (LT)	A word-for-word translation of a concept that may be unfamiliar in the target culture, preserving the source language's structure as closely as possible. For example: Het Hoekske - *Ćošak* [The Corner].
Description (D)	The CSI is replaced by a descriptive phrase or explanation to convey its meaning or function in the target language. For example: Glas-in-lood - *Okna u raznobojnom staklu* [pane with colorful glass].
Generalization (G)	The CSI is replaced with a neutral or broader reference that lacks cultural specificity. For example: Boterkoek - *Pecivo* [Pastry].
Localization (L)	The CSI is replaced with a reference specific to the target culture, making it more familiar to the target audience. For example: Hutsepot - *Čušpajz*.
Mistranslation (Mis)	Errors in translation, including incorrect orthographic adaptation, grammatical or semantic inaccuracies, or the use of non-existent words. For example: Vogelpik - *Kljucanje ptica* [Birds pecking].

Table 1: Adapted Taxonomy of Translation Strategies for Rendering CSIs. (Budimir 2021)

Varying the prompts would have introduced an additional variable, potentially influencing the outcome and making cross-model comparisons (NMT and LLMs) less meaningful.

It is important to note that dividing the corpus into chunks may disrupt the narrative flow, potentially limiting the models' ability to fully comprehend and translate context-dependent CSIs. Despite this limitation, the approach ensures consistency and accommodates the technical constraints of the models.

2.2 Translation Equivalents and Strategies

Translation equivalents were extracted in an Excel sheet, paired with the original CSI and the human translation from the previous study. On average, 206 equivalents per model were identified, as multiple translations of the same CSI were recorded.

The translations were categorized by the author, an experienced translator and researcher in the field of translation studies, using a taxonomy adapted from Budimir (2021), which includes the following strategies: Repetition (R), Orthographic Adaptation (OA), Combination of Strategies (COM), Literal Translation (LT), Description (D), Generalization (G), Localization (L), and Mistranslation (Mis). Table 1 provides a detailed description of these strategies. The taxonomy facilitates a granular analysis, capturing the diversity of approaches employed by the models and enabling meaningful comparisons with human translations.

It is necessary to point out that the classification process is inherently subjective. As the categorization was performed by a single annotator, the results may reflect one individual's interpretive biases. While the taxonomy provides clear guidelines, subjective judgment is often required to determine the most appropriate category for each translation equivalent. This limitation is particularly relevant for studies involving smaller language pairs, where finding annotators can be challenging. Future studies could mitigate this limitation by employing multiple annotators and calculating inter-annotator agreement to enhance reliability and validity.

3 Results

3.1 General Overview

A visual illustration of the distribution of translation strategies employed by the MT tools (Google Translate, Gemini, ChatGPT (1), and

ChatGPT (2)), as well as by the human translator is presented in Figure 1. A Chi-Square test was conducted to assess differences in translation strategies among the models, revealing significant variation ($\chi^2 = 100.83$, $p < 0.05$, dof = 24). These results indicate that each model exhibits distinct approaches to handling CSIs.

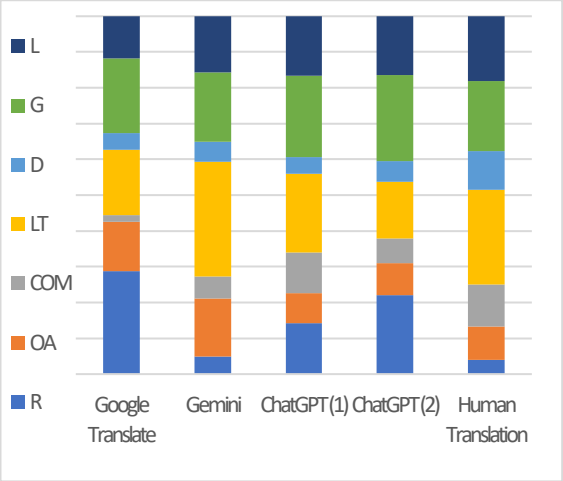

Figure 1: Distribution of Strategies across Models and Human Translators.

Mistranslation (Mis) rates (Table 2) highlight differences in model accuracy. Google Translate exhibits the highest error rate, with 56 instances (26.8%) of mistranslation, reflecting substantial challenges in handling CSIs. This result aligns with the findings of Yao et al. (2024), which demonstrate the superior ability of LLMs over NMT systems in managing CSIs. The relatively high error rate of ChatGPT (2), however, can be attributed to a specific translation issue: the omission of the case suffix in retained CSIs. This issue will be further discussed in the section 3.2.

Excluding mistranslations, the distribution of correct strategies provides insights into the strengths and weaknesses of each model. Google Translate relies heavily on Repetition (R) (21.1%), indicating a tendency to preserve CSIs in their original form without adaptation. This approach often contradicts Serbian norms, where orthographic adaptation is preferred. In contrast, ChatGPT (both versions) exhibits the strongest reliance on Generalization (G) and Localization (L), reflecting an effort to adapt cultural references for the target audience. ChatGPT (1)'s frequent use of the Combination of Strategies (COM) underscores its capacity to enhance contextual clarity, while Google Translate and Gemini rely more on straightforward strategies, offering limited additional explanation.

Among machine translation models, Gemini demonstrates the most balanced distribution of strategies. It effectively integrates Literal Translation (LT), Orthographic Adaptation (OA), Generalization (G) and Localization (L), suggesting a more adaptable approach. This balance mirrors the diversity observed in human translation more closely than in either Google Translate or ChatGPT, which exhibit a narrower range of strategies. Statistical metrics support this conclusion, as demonstrated by measuring Euclidean distance—a method commonly used to evaluate similarity between categorical data—between each model and human translation as the baseline. Gemini exhibits the smallest Euclidean distance from human translation (0.122), followed by ChatGPT (1) (0.133). ChatGPT (2) (0.227) and Google Translate (0.297) display greater divergence.

The analysis of the distribution of strategies across cultural categories (Figure 2) offers additional insights. ChatGPT (2) and Gemini produce the highest number of errors in the Proper Names (PN) category, indicating significant challenges in adapting names to Serbian linguistic norms. In contrast, ChatGPT (1) and Google Translate show the most errors in the Social Culture (SC) category, suggesting difficulties in handling references to job titles, leisure activities, and institutions. These results highlight how different models struggle with specific cultural categories, reflecting varying capabilities in adapting to cultural and linguistic nuances. Google Translate continues to demonstrate a relatively consistent struggle across all categories, underlining its limited cultural sensitivity.

For Material Culture (MC) references, all machine translation models frequently rely on Generalization and Literal Translation. Human Translation, by contrast, employs Generalization (28%) alongside a stronger preference for Description (24%) and Literal Translation (18.7%).

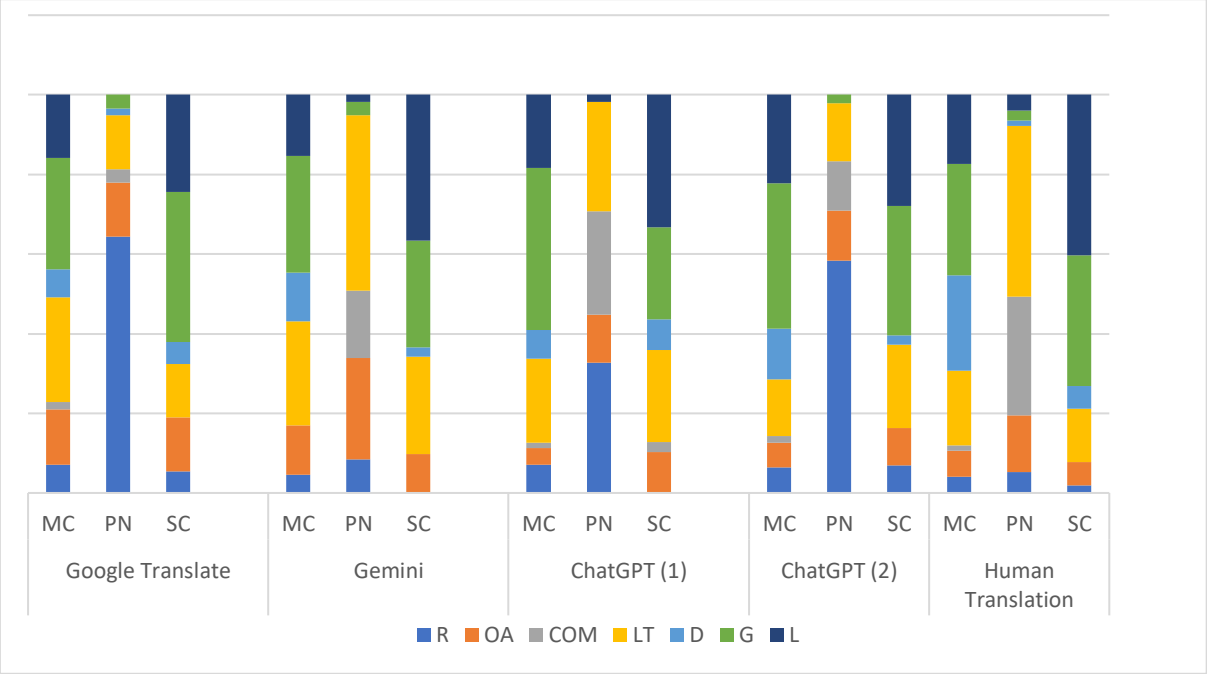

Figure 2: Distribution of Strategies across Cultural Categories and Models.

For Proper Names (PN), Repetition is predominant in Google Translate, while Gemini and ChatGPT favor Literal Translation. Human Translation demonstrates a preference for Literal Translation (42.9%) and the Combination of Strategies (29.9%), reflecting its emphasis on adapting to the target culture and providing contextual information. In the case of Social Culture, Generalization and Localization are dominant strategies for both machine models and Human Translation. However, machine models employ Localization less frequently (e.g. 16.5% in ChatGPT (2)) than Human Translation (40.4%).

Human Translation consistently prioritizes Localization and Generalization for Social and Material Culture, while favoring Literal Translation for Proper Names, thus demonstrating a clear preference for culturally adaptive strategies. In contrast, machine models show less consistency and rely more heavily on Generalization and Literal Translation, particularly in more challenging categories.

3.2 Error Analysis

Semantic errors were the most prevalent errors across all models (Figure 3), reflecting the significant challenges these systems face with the polysemy of multi-word expressions and exocentric compound words, which CSIs often comprise. Insights from studies on polysemous words emphasize the importance of contextual dependency in resolving such errors. Machine translation often fails to disambiguate polysemous terms and uses primary meanings without considering context (Ohuoba et al., 2024). For instance, Google Translate rendered *jarige kaas* as *rođendanski sir* [birthday cheese], incorrectly interpreting *jarig* as "birthday" rather than its actual meaning in this context, "aged"—the correct translation being "aged cheese". Similarly, *zure spekken* was mistranslated as *kisele slanine*

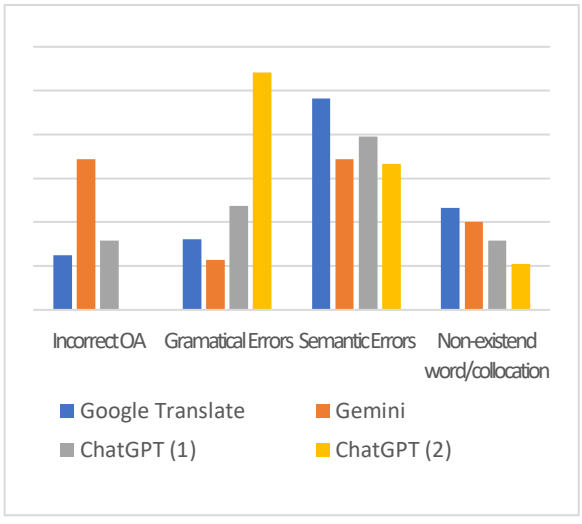

Figure 3: Distribution of Error Occurrences (%) across Models.

[sour bacon], where *spekken* (a type of soft candy) was wrongly interpreted as "bacon".

After Google Translate (23.3%), Gemini exhibited the highest frequency of invented words and collocations (20%), with examples including *takmičenje u kašetanju* for *kaatwedstrijd* [a sort of ball sport] and *kafić Korabljavanje* for *cafe De Scheepvaart* [cafe Shipping]. These outputs suggest a tendency to generate nonsensical and non-existent terms in Serbian, reflecting lexical gaps in the model's training data and its resorting to hallucination. While ChatGPT produced fewer non-existent terms, similar issues were observed, indicating room for improvement in vocabulary alignment with Serbian norms.

Grammatical errors were particularly pronounced in ChatGPT (2), with a notable 54.2% rate, largely due to challenges with Serbian case endings, agreement, and word order (16 out of 26 errors in this category). For example, *Grote Markt* [Main Square] was translated without the proper locative case ending (*na Grote Markt* instead of *Markt-u*), and brand names were often repeated without morphological adaptation, as in *flaša Bols likera* instead of *flaša Bolsovog likera* [bottle of Bols liqueur]. One prominent error in Gemini involved the incorrect use of the suffix *-ski* instead of *-ov* when forming adjectives from people's names. For instance, *Snellaertstraat* was incorrectly adapted as *Snerlatska ulica* instead of the correct *Snelartova ulica*.These errors disrupt the syntactic and morphological coherence of the output, diminishing its overall accuracy.

Incorrect orthographic adaptation (OA) was another common issue, especially in Gemini (34.3%) and ChatGPT (1) (15.8%). For example, place and street names were often inconsistently adapted, violating Serbian orthographic norms. For example, the diphthong *ui* is adapted as *u* in *Oostduinkerke* and as *iu* in *Korte Gasthuisstraat.* The correct form should be *aj*. In contrast, ChatGPT (2) avoided such errors entirely, due to the predominant use of Repetition.

As highlighted in the Prolex study on French, Serbian, and Bulgarian (Maurel et al., 2007), rich inflectional systems require proper names to be adapted across cases (e.g., nominative, genitive, dative), which adds complexity to translation tasks. Serbian proper names exhibit multiple inflectional forms, underscoring the need for MT systems to incorporate morphological rules effectively.

The influence of English was most apparent in Google Translate's outputs. For example, the café name *Het Hoekske* was translated into Serbian with the English definite article "the," resulting in *The Hoekske*. Similarly, the term *bloedworst*, a type of blood sausage commonly found in Flemish cuisine, was incorrectly rendered as *crni puding* [black pudding], borrowing the literal English term, which does not align with typical Serbian culinary terminology. Another case is the translation of *schorseneer* (a root vegetable known as salsify) where the English term salsify was transferred directly into Serbian. These examples illustrate the system's reliance on English as an intermediary language, which can distort meaning and reduce the cultural and linguistic accuracy of the target text. These findings are consistent with the challenges described by Ohuoba et al. (2024), where English's dominance as a high-resource language often skews translations for low-resourced languages, by introducing cultural mismatches and semantic inaccuracies.

Some translations included lexical forms from closely related languages such as Croatian or Slovenian, such as *vrtuljak* for *paardenmolen* (a Flemish term for "carousel") and *pivovarna* for brouwerij (brewery). While these forms may be intelligible to Serbian speakers due to the linguistic similarities among South Slavic languages, they are less common or non-standard in Serbian. This highlights potential inconsistencies in the models' adaptation to regional language norms and raises questions about the influence of neighboring languages on machine-generated output.

4 Discussion

The present analysis of translation strategies and error patterns highlights the key challenges of machine translation of CSIs and especially into morphologically complex languages like Serbian. When compared to human translation, machine translation exhibits significant problems in handling CSIs.

The study of human translation strategies reveals clear and consistent patterns when dealing with proper names, including street names, brand names, and HoReCa terminology (Budimir, 2021). Orthographic adaptation is commonly applied when the name includes people's names, particularly historical figures, well-known fictional or real people, or toponyms. In contrast, literal translation is used when the name consists of nouns and adjectives. This systematic approach ensures that cultural and semantic nuances are preserved in the target text while maintaining the readability and cultural familiarity for the audience. Machine translation, however, fails to follow such patterns, often producing random and inconsistent results.

Another notable issue is the predominant use of repetition without adding contextual information. While repetition can sometimes suffice when the meaning of the CSI can be inferred from context, this is often not the case. The preservation of the communicative function is a crucial aspect of translating CSIs (Ivir, 2003). Simply repeating a term without adaptation or explanation can fail to convey its intended cultural significance, leaving the target audience disconnected from the original message. For instance, retaining pastries such as *mastellen* and *pistoletten* in their original forms does not evoke any cultural or semantic associations apart from the act of eating. This approach neglects the cultural connotations and traditional significance attached to these items in the source culture. Such CSIs require additional strategies, such as adaptation or explanatory supplementation, to ensure that their cultural and communicative essence is effectively conveyed (Ivir, 2003). Without added context, the meaning and significance of these items are lost to the Serbian audience, reducing the overall effectiveness of the translation (Hlebec, 2009).

5 Conclusion

This study has highlighted several key findings regarding the performance of machine translation (MT) systems in translating culture-specific items (CSIs) between Flemish Dutch and Serbian. First, while models such as Gemini and ChatGPT demonstrate a promising use of generalization and localization strategies for material and social culture CSIs, they often fail to apply nuanced approaches required for complex or less common CSIs. Proper names, in particular, pose significant challenges due to the rich inflectional demands of Serbian and the need for orthographic adaptation.

From a strategy perspective, the analysis reveals that Gemini exhibits the most balanced distribution of approaches, incorporating literal translation, orthographic adaptation, generalization, and localization more effectively than other models. Nevertheless, even Gemini struggles with systematic cultural adaptation and fails to match the nuanced strategies consistently employed in human translation. ChatGPT's use of the combination strategies shows potential for improving contextual clarity, yet its tendency to omit morphological adaptations in Serbian remains a limitation. Meanwhile, Google Translate, while heavily reliant on repetition, exhibits the highest error rates and demonstrates limited cultural sensitivity in handling CSIs.

These findings underscore the irreplaceable role of human translators in effectively handling CSIs, particularly in literary and culturally rich texts. Human translators not only bring cultural and contextual understanding to the task, but also excel at preserving the communicative function of CSIs, a dimension often overlooked by MT systems. For example, while MT models tend to rely on repetition or overgeneralization, human translators adapt CSIs dynamically, ensuring that their cultural essence and intended meanings resonate with the target audience.

Furthermore, the implications of this study extend to translator training and workflow design. As MT systems become more prevalent, human translators are increasingly assuming roles as post-editors. This shift emphasizes the importance of equipping translators with the skills needed to identify and address the shortcomings of MT outputs, such as the failure to capture cultural nuances or apply morphological adaptations. By integrating human expertise with MT capabilities, translation workflows can achieve greater efficiency while preserving linguistic and cultural fidelity.

Recent studies have increasingly emphasized the potential of CAT and MT tools to enhance translator efficiency, particularly when dealing with complex or culture-bound elements that require extensive background research and

strategic decision-making. Hadley (2023) highlights how such tools can alleviate cognitive load by supporting specific aspects of literary translation, such as managing sentence length, rhythm, or poetic form. Similarly, Kolb and Miller (2022) demonstrate that specialized tools like PunCAT can aid in resolving linguistically dense challenges, such as puns, by expanding the translator's pool of potential solutions. These developments suggest promising avenues for future tool design.

In the context of CSI translation, where human translators often invest significant time in interpreting meaning and selecting appropriate strategies, MT systems could be further adapted to present a range of contextually informed suggestions. Experimenting with prompt engineering, designed to generate multiple culturally and linguistically relevant options for each CSI, may prove especially beneficial in supporting informed and efficient human decision-making. In this regard, hybrid human-machine approaches and the development of culturally aware translation tools are crucial. Yao et al. (2024) provide a compelling framework for advancing MT by integrating cultural databases and CSI annotations, as well as introducing innovative metrics to evaluate cultural and contextual fidelity. Building on such approaches could significantly enhance MT performance, particularly for texts with rich cultural content.

One concrete avenue for such improvement involves addressing the persistent errors in orthographic adaptation, particularly when translating into morphologically rich languages like Serbian. These errors could be mitigated through the integration of language-specific orthographic rules, culturally adapted name databases, and targeted post-editing support within LLM systems. Such refinements, combined with the insight and flexibility of human translators, would allow for more accurate and culturally resonant translations of proper names and other CSIs.

Limitations

It should be noted that the present study is limited by its relatively small dataset of 197 analyzed CSIs, which restricts the generalizability of its conclusions. Additionally, the reliance on a single annotator introduces potential subjectivity in strategy classifications. To address these limitations, future research should analyze larger datasets, employ multiple annotators for improved reliability, and explore different datasets, including other low-resource languages, to test the consistency of observed patterns. Investigating the impact of varying prompts for large language models (LLMs) and experimenting with hybrid approaches that combine machine translation and human post-editing could further enhance the understanding and handling of culturally nuanced content. Such advancements would contribute to more robust cultural adaptation and contextual modeling in MT systems, aligning them more closely with human translation standards.

Ethics Statement

The corpus used in this research has been utilized exclusively for academic and research purposes, in compliance with copyright and ethical guidelines. All texts within the corpus have been accessed and processed solely to analyze translation strategies and linguistic phenomena as part of this study.

The corpus is securely stored on a private computer and is not accessible on any online platform or public repository. No part of the corpus has been shared, distributed, or made available beyond the scope of this research.

For the purposes of analysis, the texts were processed as loose, decontextualized sentences to focus on specific translation patterns and strategies. This approach ensures that the study adheres to ethical research standards while minimizing potential risks associated with handling copyrighted material in its entirety.

Researchers interested in the corpus for academic purposes may request access by contacting the author directly, subject to appropriate ethical and copyright considerations.

Sustainability Statement

The estimated energy usage for this study was negligible, as it involved text processing and analysis rather than high-resource training or large-scale inference tasks. As such, the environmental impact of the research is minimal.

References

Primary sources

Hugo Claus. 1983. *Het verdriet van België*. De Bezige Bij, Amsterdam.

Hugo Klaus. 2000. *Tuga Belgije T.1 Tuga*. Prometej, Novi Sad.

Hubert Lampo. 1960. *De komst van Joachim Stiller*. Meulenhoff, Amsterdam.

Hubert Lampo. 1992. *Dolazak Joahima Štilera*. Luta, Beograd.

Dimitri Verhulst. 2006. *De helaasheid der dingen*. Contact, Amsterdam.

Dimitri Verhulst. 2015. *Zaludnost življenja*. Clio, Beograd.

Secondary sources

Javier Franco Aixelà. 1996. Culture-Specific Items in Translation. In R. Álvarez and M. C. Vidal, editors, *Translation, Power, Subversion.* Multilingual Matters, Philadelphia, pages 52–78.

Ana Guerberof-Arenas, and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas, and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. *Translation Spaces*, 11(2): 184-212.

Бојана Будимир. 2021. Културноспецифични елементи из фламанске културе у преводу на српски језик [*Culture-Specific Items from Flemish Culture in Serbian Translation*]. D.Phil. dissertation, University of Belgrade, Faculty of Philology, Belgrade, Serbia.

CEATL. 2024. AI Survey for individual translators. https://www.ceatl.eu/wp-content/uploads/2024/04/CEATL_AI_survey_for_members.pdf.

Ella Creamer. 2024. Dutch publisher to use AI to translate 'limited number of books' into English. *Guardian*.

Joke Daems. 2022. Dutch literary translators' use and perceived usefulness of technology: The role of awareness and attitude. In *Using technologies for creative-text translation*. Routledge, pages 40-65.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. *Language Resources and Evaluation*, 3790–3798. https://biblio.ugent.be/publication/8662553/file/8662566.pdf

Diederik Grit. 2010. De vertaling van realia. In T. Naaijkens, editor, *Denken over vertalen*. Vantilt, Nijmegen, pages 189–196.

Vladimir Ivir. 2003. Translation of Culture and Culture of Translation. *SRAZ XLVII-XLVIII***:** 117–126.

James Luke Hadley. 2023. MT and CAT: Challenges, Irrelevancies, or Opportunities for Literary Translation?. In *Computer-assisted Literary translation*. Routledge, pages 91-105.

Борис Хлебец. 2009. *Опита начела превођења* [*General Principles of Translating*]. Београдска књига, Београд.

Arvi Hurskainen. 2013. Handling proper names in Machine Translation. Technical Report 12.

David Katan 2009. Translation as intercultural communication. In J. Munday, editor, *The Routledge Companion to Translation Studies.* Routledge, London/New York, pages 74–92.

Waltraud Kolb, and Tristan Miller. 2022. Human–computer interaction in pun translation. In *Using technologies for creative-text translation*. Routledge, pages 66-88.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland. European Association for Machine Translation.

Ritva Leppihalme. 1997. *Culture bumps: An empirical approach to the translation of allusions*. Multilingual Matters.

Ritva Leppihalme. 2001. Translation strategies for realia. In *Mission, vision, strategies, and values: a celebration of translator training and translation studies in Kouvola*. Helsinki University Press, pages 139-148.

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

D. Maurel, D. Vitas, C. Krstev, S. Koeva. 2007. Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian. In A. Dziadkiewicz and I. Thomas, editors, *Bulag - Bulletin de Linguistique Appliquée et Générale, Les langues slaves et le français : approches formelles dans les études contrastives*, No. 32, pages 55–72, Presses Universitaires de Franche Comté, Besançon.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, New York/London.

Adaeze Ohuoba, Serge Sharoff, and Callum Walker. 2024. Quantifying the Contribution of MWEs and Polysemy in Translation Errors for English–Igbo MT. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 537–547, Sheffield, UK. European Association for Machine Translation (EAMT).

Harald Martin Olk. 2013. Cultural references in translation: a framework for quantitative translation analysis. *Perspectives,* 21(3): 344-357

Brian Porter and Édouard Machery. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Dental science reports*, 14(1).

Danti Pudjiati, Ninuk Lustyantie, Ifan Iskandar, and Tira Nur Fitria. 2022. Post-editing of machine translation: Creating a better translation of cultural specific terms. *Language Circle: Journal of Language and Literature*, 17(1): 61-73.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Anita Srebnik. 2023. Het taaltechnologische landschap van het Nederlands in een meertalig Europa. *Internationale Neerlandistiek*, 61(3): 217–241.

Lawrence Venuti. 1995. The Translator's Invisibility: A History of Translation. Routledge, London/New York.

Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics (Basel)*, 7(3):32.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking Machine Translation with Cultural Awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

A Additional Data

Table 2 presents a detailed overview of the distribution of translation strategies and mistranslation employed by MT models (Google Translate, Gemini, ChatGPT (1), ChatGPT (2)) and human translation.

	Google Translate	**Gemini**	**ChatGPT (1)**	**ChatGPT (2)**	**Human Translation**
Repetition (R)	44 (28.8%)	8 (4.8%)	24 (14.3%)	35 (22.2%)	8 (3.9%)
Orthographic Adaptation (OA)	21 (13.7)	27 (16.4%)	14 (16.4%)	14 (8.9%)	19 (9.3%)
Combination of Strategies (COM)	3 (2%)	10 (6.1%)	19 (6.1%)	11 (7%)	24 (11.8%)
Literal Translation (LT)	28 (18.3%)	53 (32.1%)	37 (32.1%)	25 (15.8%)	54 (26.5%)
Description (D)	7 (4.6%)	9 (5.5%)	8 (5.5%)	9 (5.7%)	22 (10.8%)
Generalization (G)	32 (20.9%)	32 (19.4%)	38 (19.4%)	38 (24.1%)	40 (19.6%)
Localization (L)	18 (11.8%)	26 (15.8%)	28 (15.8%)	26 (16.5%)	37 (18.1%)
Total	153	165	168	158	204
Mistranslation (Mis)	56 (26.8%)	35 (17.5%)	38 (18.4%)	48 (23.3)	1 (0.5%)

Table 2: Distribution of Strategies and Mistranslation across Models and Human Translators.