Does the perceived source of a translation (NMT vs. HT) impact student revision quality for news and literary texts?

Xiaoye Li
Hunan University / Changsha, China
Ghent University / Ghent, Belgium
lixiaoye2022@hnu.edu.cn

Joke Daems
Ghent University / Ghent, Belgium
joke.daems@ugent.be

Abstract

With quality improvements in neural machine translation (NMT), scholars have argued that human translation revision and MT post-editing are becoming more alike, which would have implications for translator training. This exploratory study contributes to this growing body of work by exploring the ability of 16 student translators (ZH-EN) to distinguish between NMT and human translation (HT) for news text and literary text and analyses how text type and student perceptions influence their subsequent revision process. We found that participants were reasonably adept at distinguishing between NMT and HT, particularly for literary text. Participants' revision quality was dependent on the text type and the perceived source of translation. The findings also highlight student translators' limited competence in revision and post-editing, emphasizing the need to integrate NMT, revision, and post-editing into translation training programmes.

1 Introduction

The rise of neural machine translation (NMT) has led to paradigm shifts in translation research and education. Claims of NMT quality reaching 'human parity' (Hassan et al., 2018), suggest blurring boundaries between MT and human translation (HT). While this claim has been contested (Läubli et al., 2020; Poibeau, 2022), it has had consequences for text types considered to be suitable for MT and for the conceptualisation of revision and post-editing (PE). With quality improvements over earlier MT paradigms, a growing body of work has explored the potential of NMT for literary translation (Matusov, 2019; Toral and Way, 2018), and some book publishers are actively integrating post-editing into their workflows (Creamer, 2024).

From a theoretical perspective, the evolution in MT quality sparked a debate on the fundamental differences between post-editing (the improvement of machine-translated text) and revision (the improvement of human-translated text), and whether they have essentially become the same task (Do Carmo and Moorkens, 2020). Such distinctions (or lack thereof) are important from a translation training perspective, as students need to develop the necessary skills to thrive in the translation industry. If post-editing and revision are fundamentally different tasks, students need to receive training explicitly tailored to both (Robert et al., 2024). It has been suggested that being able to identify differences between HT and NMT output is a key component of MT literacy (De Clercq et al., 2021). An additional factor is the potential lack of transparency in the translation workflow itself. When receiving a revision assignment, translators might not always be made aware of the actual provenance (machine or human) of a translation, and they will still be required to produce a final text fit for publication.

In order to improve our understanding of (perceived) differences and similarities between NMT and HT, this study evaluates the ability of student translators to distinguish between both for two distinct text types: news text and literary text. Students were asked to clarify their choices and to edit the text to produce a final product of publishable quality. In this paper, we answer the following research questions:

RQ1. Can students distinguish between NMT-translated or human-translated texts for different text types?

RQ2. Which factors do students take into account when determining the source of translations and do they consider different factors for different text types?

RQ3. Does the source of a translation and whether or not it is correctly identified influence the

Proceedings of the Second Workshop on Creative-text Translation and Technology (CTT), pages 14–26, June 24, 2025

changes students make? That is, will the number of revision changes and the revision quality be higher when participants correctly identify the source of the translation?

2 Related research

2.1 NMT for different text types

NMT represents a new paradigm in the field of machine translation (MT), demonstrating substantial improvements in translation quality (Koponen et al., 2021). Existing studies have shown the promising performance of NMT for technical text (Wang and Wang, 2021), yet its effectiveness for creative texts remains unclear. Creative texts, shaped by human creativity and reliant on aesthetics, traditionally include genres like novels, poetry, plays, and comics (Hadley et al., 2022). Although news texts are informational in nature, they still offer a certain degree of creative space, especially when conveying cultural nuances and subtleties. Due to the need for rapid information delivery (Yang et al., 2023), NMT has been adopted in news translation (Krüger, 2022; Dai and Liu, 2024). In comparison to news texts, literary texts require a higher level of creativity, and there is an ongoing debate about NMT's applicability to literary works (Daems, 2022; Guerberof-Arenas et al., 2024; Rothwell et al., 2023). Studies showed that around 30-44% of sentences in NMT outputs for literary texts do not contain any errors (Matusov, 2019; Tezcan et al., 2019; Toral and Way, 2018) and that there is a great variability in error types depending on the language pair, with morphosyntactic errors being particularly common for French-to-Croatian, and literal translations and lexical errors being common for English-to-French translations (Petrak et al., 2022).

2.2 Differences between NMT and HT

While quality is a crucial component to judge the potential of NMT for different text types, the impact of technology on language as a whole is a topic of discussion as well. Research into so-called Machine Translationese has shown that the use of MT leads to a loss of linguistic, morphological and syntactic diversity compared to original texts or human translations (Vanmassenhove et al., 2019, 2021; De Clercq et al., 2021; Tezcan et al., 2019; Bizzoni et al., 2020; Sizov et al., 2024). Understanding the impact of MT on language becomes particularly important in the context of literary translation, where creativity is a key component. For classic novels translated from English into Dutch, for example, NMT was shown to have a lower level of lexical diversity than HT, and HT showed more syntactic variability, whereas NMT output generally followed the source text structure (Webster et al., 2020). Research on literary translation for English-Turkish (Şahin and Gürses, 2019) and English-Catalan and English-Dutch (Guerberof-Arenas and Toral, 2022) revealed that NMT negatively impacts creativity, which could also influence the readers' experience. Idioms and manipulated multiword expressions can pose a particular challenge to NMT compared to human translators (Corpas Pastor and Noriega-Santiáñez, 2024).

2.3 Translation revision and post-editing

The ISO 17100 (2015) explicitly states that translation services should ensure that translations are revised. It underscores the critical role of revision in enhancing the quality of human-translated texts. Similarly, post-editing is an essential step in ensuring the final quality of NMT-translated texts. Numerous researchers have engaged in theoretical discussions on translation revision and post-editing (Do Carmo and Moorkens, 2020; Nitzke et al., 2019; Robert et al., 2017; Scocchera, 2020).

Theoretically, some scholars argue that post-editing can be regarded as a form of revision, as it may be carried out monolingually without access to the source text (Krings, 2001; Schwartz, 2014). This view is further supported by evidence that translators spend more time pausing than typing during post-editing tasks (Koehn, 2009; Ortiz-Martínez et al., 2016).

However, this viewpoint has been contested. Jakobsen (2018) notes that while technology blurs the boundaries between translation, revision, and post-editing, fundamental differences between human and machine translation distinguish the latter two. Similarly, do Carmo and Moorkens (2020) caution that perceiving post-editing as a form of revision might lead to its undervaluation and influence pedagogical practices. Girletti (2022) found that corporate translators in Switzerland regard revision and post-editing as separate tasks, despite similarities in reading strategies. Evidence from a literary translation workflow in which an MT text was first post-edited and then revised indeed shows that the kinds of changes introduced in both processes are very different (Macken et al., 2022), with post-editing actions mostly focusing on the

correction of MT errors, and revision actions focusing on preferential changes (explicitation, stylistics changes, coherence markers).

The debate about the differences and similarities between translation revision and post-editing has an impact on translator training, and the (presumed) necessary competences for both. Some scholars have proposed different competence frameworks for revision and post-editing, respectively (Pym, 2013; Rico Pérez and Torrejón, 2012; Robert et al., 2017; Scocchera, 2020; Konttinen et al., 2020). It suggests that earlier theoretical discussions are still not settled, and previous conclusions are still being examined through practical research.

To date, few studies have empirically verified the differences between translation revision and post-editing, and the data analysed in existing studies have been limited to translations between European languages, such as English-Dutch (Daems and Macken, 2020) and Dutch-French (Robert et al., 2023). Daems and Macken (2020) conducted a study in which professional translators were asked to revise or post-edit a given source text, with some 'revisors' being given an MT source text, and some 'post-editors' being given a HT, without the translators being aware of this deception. The study showed that most changes were made to the MT text when participants thought they were revising a HT, and that this also led to the greatest quality improvements. Robert, Schrijver and Ureel (2023) focused on competences in translation trainees to establish if there are differences between translation, translation revision, and post-editing. They concluded that revision is different from post-editing as students performed worse for the post-editing task, although the 'problem detection' competence was found to be shared across both tasks.

Therefore, this study aims to examine the perceived similarities and differences between NMT and HT, and to determine whether the perceived source of translations affect the revision quality across news and literary texts. Moreover, it will be the first to explore the relationship between translation revision and post-editing across languages that are more distant and linguistically remote from each other, i.e. Chinese and English.

3 Methods

3.1 Participants

Participants were 16 students enrolled in a Translation and Interpreting Master program from Hunan university in China. They were native Chinese speakers (L1) with English as their second language (L2). All participants had passed the Test for English Majors at Band 8, a standardized English proficiency exam in Chinese universities, ensuring a high level of English proficiency. Translation into English is a core competence in Chinese translator education due to market demands. Participation in the experiment was optional as part of the students' translation course, with participation leading to extra course credit. The experiment was approved by the Ethics Committee of the School of Foreign Languages at Hunan University. All participants signed an Informed Consent form before the experiment.

To minimize the impact of English proficiency on the experimental results, we used R Studio to run a greedy algorithm that divided participants into two groups: Group A and Group B. The greedy algorithm is used to optimize resource allocation with minimal variance between groups (Korte and Vygen, 2018), and we used it to ensure that the overall English proficiency of the two groups was as balanced as possible (see Table 1).

Indicators	Group A	Group B
Participants	8	8
EN proficiency	86.00	86.13
Age	21.70	21.50
Task 1	HT of prose	NMT of prose
Task 2	NMT of news	HT of news

Table 1: Information of Group A and Group B

3.2 Materials

3.2.1 Source texts

When selecting the source text (ST), we comply with the following requirements. First, the total length of the ST should be kept within a specified range to prevent participant fatigue from excessively long texts. Second, the average sentence length of the ST should be regulated, as NMT performs better with shorter sentences (Moorkens et al., 2018). This ensured that NMT outputs were not overly refined, which could reduce the need for post-editing (Daems et al., 2017). Third, text complexity should be evaluated via objective and subjective methods. The objective evaluation method could take into account linguistic features such as sentence structure (Hvelplund, 2011). The subjective evaluation method involves translator experts to assess the readability, comprehensibility, and

translatability of texts (Zheng et al., 2020).

Based on these requirements, We firstly selected four source texts (two from a news text and two from a literary text), ensuring consistency in the number of Chinese characters and sentences. Then, four translators with over five years of experience evaluated the text complexity across readability, comprehensibility, and translatability using a Likert scale (1 = easy, 5 = difficult).

After evaluation, we selected two Chinese source texts (ST 1 and ST 2) for the formal experiments. ST 1 was drawn from a news text titled 'The Belt and Road Initiative: A Key Pillar of the Global Community of Shared Future' published by The State Council Information Office of the People's Republic of China. ST 2 was excerpted from a Chinese literary text titled 'Time Is Life' by Shiqiu Liang, a well-known Chinese writer. It conveyed the author's thoughts on time and life, characterized by rich rhetorical devices. After evaluation, ST 1 and ST 2 predominantly consist of complex sentences such as progressive compound sentences or sequential compound sentences. Both texts share a similar number of Chinese characters, average sentence length, and closely aligned results in subjective evaluations (see Table 2). These results suggest that the two texts are comparable in terms of text complexity.

Indicators	ST 1	ST 2
Text genre	news text	literary text
Number of characters	131	123
Number of sentences	4	4
Avg. sentence length	32.75	30.75
Readability	3.50	3.55
Comprehensibility	3.00	3.10
Translatability	3.85	3.80

Table 2: Information of ST 1 and ST 2

3.2.2 Sources of translations

The source texts were translated from scratch by a senior undergraduate student majoring in English. The student was instructed to translate without using NMT or consult any online resources.

To simulate real-world scenarios where translators assess translation outputs from various NMT systems, we translated the STs using several widely accessible and reliable NMT systems, including Google Translate, DeepL, Baidu Translate, and Youdao Translate.

The quality of these HT and NMT outputs was then evaluated by two professional translators, with more than 10 years of translation experience, using Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) alongside reference translations. The reference translation of the literary text was by Peiji Zhang, a renowned Chinese translator, while the news translation was adopted from the China State Council Information Office. Following Carl and Baez (2019), translation errors could be classified into two categories: 'critical' errors (score of 2) and 'minor' errors (score of 1).

Based on the evaluation results, we selected the translations by DeepL for PE tasks, as they were the most similar to HT in terms of the number of errors. This was done to prevent the imbalance in the number of errors between NMT and HT from affecting the results. Since the number of errors in the NMT-translated literary text was one more than in the HT-translated literary text, We then corrected one minor fluency error in the DeepL translation to ensure consistency in both the number of errors and the error severity weight between the human-translated and NMT-translated texts. After the adjustment, the final translations contained 8 errors with a total error severity weight of 14. A full overview of error types for each text and translation source can be seen in Appendix A.

3.3 Experiment procedure

To initiate the experiment, we provided participants with a task brief, post-editing guidelines (Nitzke and Hansen-Schirra, 2021), and general revision guidelines (Moorkens et al., 2018), followed by the distribution of the experimental tasks. These supplementary materials can be found online at `https://osf.io/ubc8x/?view_only=f11dc93134004f01a029b207415b02d2`. All participants installed EVCapture, a screen recording software, on their laptops and completed a warm-up exercise to practice revision and post-editing.

During the formal experiments, participants were instructed to complete two tasks (Task 1 and Task 2). We used a between-subjects design, with each participant receiving one ST in its HT version and the other in its NMT version (participants were not informed of the origin of the translations). The details of group division are shown in Table 1. The current design was implemented to minimize potential biases that could arise from asking participants to identify both HT and NMT of the same ST. Such a task might have influenced their judgment

and revision quality in two ways: First, participants could have compared the two translations while making their assessment. Second, having access to both translations might offer helpful hints during the revision or post-editing process.

For each task, participants were instructed to identify the source of translations, provide the rationale behind their perceptions, revise the translations, and explain the reasons for their changes. There was no time limitation and participants were permitted to use online resources. Given that English translations of the STs were available at the time, we instructed participants not to consult any English versions during the experiment. Furthermore, all participants' translation processes was recorded by EVCapture. Upon completing the tasks, participants needed to save their revised documents and screen recording files.

3.4 Data collection and evaluation

This study utilized Microsoft Word documents to gather participants' perceived sources of translations, their criteria for judgment, revised translations, and the reasons for their revision changes. The accuracy rate of participants' perceived translation sources was calculated based on the perception results documented in the Word files.

The coding process for the judgment criteria was conducted across four scenarios: NMT of news text, HT of news text, NMT of literary text, and HT of literary text. Each scenario was classified as either 'consistent' or 'inconsistent', depending on whether the perceived source aligned with the actual source. 'Consistent' refers to instances where participants' perceptions matched the true source, whereas 'inconsistent' denotes cases where there was a discrepancy. In light of participants' comments, thematic keywords for each judgment criterion were identified, with care taken to minimize overlap between themes. Two main thematic keywords were identified: all comments could be classified as either relating to translation strategies or to errors. For instance, P01 noted that 'the translation is too literal' in the NMT of news text, which was categorized under the thematic keyword 'translation strategies'. Similarly, P01's comments in the HT of literary text, 'this translation even includes a spelling error,' was categorized under 'translation errors'. The frequency of each thematic keyword was then collected to identify the primary criteria participants used to assess the source of the translations across different scenarios.

For revision changes, six types of changes proposed by Robert et al. (2018) was categorised into two groups. The first group consists of changes made to translation segments with errors, which were further divided into 'necessary changes', 'missed necessary changes', and 'underrevisions'. The second group encompasses changes made to error-free translation segments, classified as 'overrevisions' (or error introduction), 'hyperrevisions' (or unnecessary changes), and 'improvements'. Two professional translators with more than 10 years of translation experience were invited to classify the revision changes based on these categories and were compensated for their time. Any disagreements in classification were resolved through discussion with the authors of this study.

Revision quality was calculated using the formula proposed by Daems and Macken (2020), focusing on three key metrics: 'necessary changes', 'overrevisions', and 'total number of errors'. Changes for critical errors, which greatly impact quality, were assigned a severity weight of '2', while those for minor errors, with less impact, were assigned '1'. The revision quality is calculated using the following formula:

$$\text{Revision Quality} = \frac{(NC \times SW) - (OR \times SW)}{TNE \times SW}$$

where:

- NC = Necessary Changes
- OR = Overrevisions
- SW = Severity Weight
- TNE = Total Number of Errors

3.5 Data analysis

Descriptive statistics were calculated for the accuracy rate of perceived sources of translations and revision changes, with the mean values for each reported. Inferential statistics were performed on revision quality. Specifically, the Shapiro-Wilk test was used to assess the data's normality due to the small sample size. The data are considered to follow a normal distribution if the p-values are greater than 0.05 (Thode, 2002). When normality is confirmed, an independent samples t-test will be employed to investigate differences between the two groups; otherwise, a Mann-Whitney U test will be applied. These analyses were conducted using

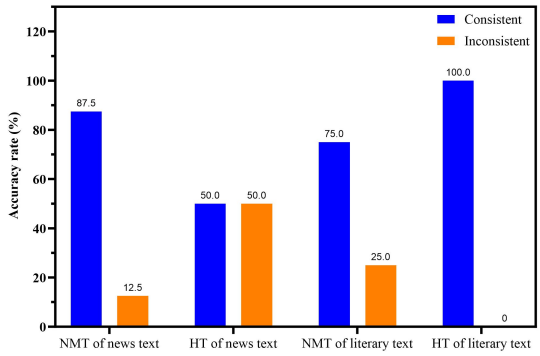

Figure 1: Accuracy rate of participants' perceptions. *Consistent* = perception matches actual source, *Inconsistent* = perception does not match actual source.

SPSS 26.0. Furthermore, GPower 3.1, a tool for statistical power analyses, was used to compute Cohen's d, a widely recognized measure of effect size. An effect size of 0.2, 0.5, or 0.8 corresponds to small, medium, or large effects, respectively (Cohen, 1988). Sawilowsky (2009) further revised the rules of thumb for effect sizes to define 0.01 (very small), 1.2 (very large), and 2.0 (huge).

4 Results

4.1 Accuracy rate of perceived sources of translations

The accuracy rate of participants' perceived sources of translations was examined across four different scenarios: NMT of news text, HT of news text, NMT of literary text, and HT of literary text (see Figure 1). The findings suggest that participants were more accurate in identifying the source of NMT-translated news text compared to human-translated news text. However, the accuracy rate for human-translated literary text is higher than that for NMT-translated literary text. It could be seen that that participants were better at recognizing the source of NMT-translated news text and human-translated literary text.

4.2 Criteria for determining sources of translations

The criteria were investigated via participants' comments across four scenarios in consistent and inconsistent situations. For NMT of news text, 87.5% of participants identified the output as machine-translated based on literal translation strategies. Most participants noted that the translation was 'overly literal' or that the sentence structures 'closely followed the ST, lacking flexibility in segmentation and cohesion.' 50.0% of participants attributed their judgment to a terminology error. For instance, they highlighted the mistranslation of the term'一带一路' ('the Belt and Road Initiative') as 'the Belt and Road', further suggesting NMT involvement. However, 12.5% of participants argued that the translation was produced by a human translator and claimed, 'Since NMT typically ensures terminology accuracy, this terminology error suggests the involvement of a human translator.'

For HT of news text, 50% of participants attributed the translation to a human translator due primarily to free translation strategies, including omission, word class shifts and meaning-based sentence restructuring. For instance, they believed that omission is a strategy commonly employed by human translators to address Chinese parallelism, where the same idea is conveyed through two similar phrases. In this translation, the parallelism '彼此隔绝、闭关锁国' was translated simply as 'isolation', suggesting the involvement of a human translator. Moreover, the term '受益者' ('the beneficiary') was transformed from a noun in Chinese into a verb phrase in English ('benefit from'). Their perception was further reinforced by the division of the third Chinese sentence into two separate English sentences in the translation. However, 50.0% of participants argued the translation was NMT-produced, noting its 'awkward and unnatural flow' or stating that most of the text is 'word-for-word', 'closely mirroring the word order in the ST', and 'failing to convey the progressive meaning.'

For NMT of literary text, 75.0% of participants attributed the output to NMT due to its overly literal translation strategies. For instance, the Chinese phrase '钟表上的秒针一下一下的移动' ('a watch or clock ticking away the seconds') was translated literally as 'the second hand on the clock move one by one.' Other awkward translations, such as 'look at the calendars hanging on the wall that can be torn off one by one,' further reinforced their perception that the translation was NMT-generated. However, 25.0% of participants argued the translation was human-produced. For instance, P16 noted that 'the overly wordy translation of the first three Chinese sentences reflected a human translator's interpretation of the ST.'

For HT of literary text, all participants correctly identified its source, as it featured free translation strategies like improved logical cohesion and omissions. For example, the phrase '再看看墙上挂着的日历' (literally 'and look at the calendars hang-

ing on the wall') was translated as 'likewise, as for the calender on the wall.' Similarly, the translation of '因为时间即生命' (literally 'because time is life') as 'after all, time is life' demonstrated a nuanced understanding of logical flow, an ability often regarded as unique to human translators. Participants also believed that a human translator is better equipped to integrate the original meanings and preserve the style of the ST by employing omission strategy, compared to NMT systems. It was interesting to note that only P01 identified the spelling error 'calender' in the human-translated literary text, a technical error that is common in HT but rare in NMT outputs.

4.3 Revision changes

The average number of revision changes was investigated across four scenarios (see Figure 2). For translation segments with errors, participants made the most necessary changes to NMT of literary text, followed by NMT of news text, HT of news text, and HT of literary text. The average number of underrevisions was similar for HT of news text, NMT of literary text, and HT of literary text, while no underrevisions were observed in NMT of news text. These results suggested participants were most effective at detecting and correcting errors in NMT-translated literary text and least effective in HT-translated literary text.

For error-free translation segments, participants introduced more errors in HT of literary text than in the other three scenarios. Additionally, hyperrevisions were most frequent in HT of news text, followed by NMT of literary text, NMT of news text, and HT of literary text. Participants achieved the most improvements to HT of news text and the fewest to HT of literary text. Overall, participants made many changes to error-free segments across all scenarios, with the highest number of changes occurring in HT of news text.

We further compared the average number of revision changes across four scenarios in consistent and inconsistent situations (see Figure 3). For NMT of news text, participants detected and corrected more errors in inconsistent scenarios than in consistent ones. In contrast, for HT of news text and NMT of literary text, participants made more necessary changes and underrevisions when their perception of the source matched the actual origin. The results suggested that participants tended to detect and correct more errors when they believed the news text was human-translated and when they perceived the literary text as NMT-generated.

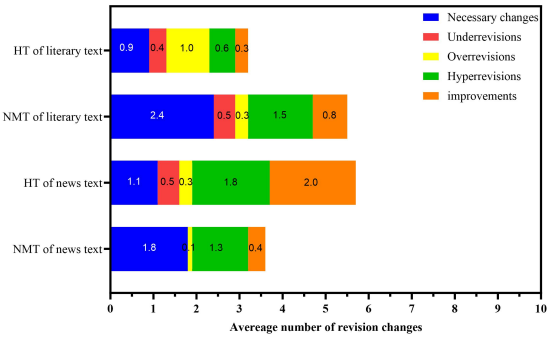

Figure 2: Average number of revision changes across four scenarios

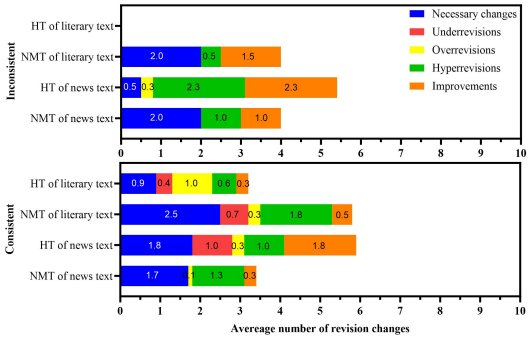

Figure 3: Average number of revision changes across four scenarios in consistent and inconsistent situations

4.4 Revision quality

We first compared the revision quality between the different sources of translation for the same text type. As shown in Figure 4, the revision quality in post-editing NMT of news text ranged from 0.00 to 0.64 (M = 0.21, SD = 0.19), while for revising HT of news text, it ranged from -0.14 to 0.29 (M = 0.12, SD = 0.13). No significant difference in revision quality was found between those two scenarios ($t(12.28) = 1.180$, $p = 0.26$, Cohen's $d = 0.59$). However, post-editing NMT of literary text ranged from 0.14 to 0.57 (M = 0.30, SD = 0.16), while revising HT of literary text ranged from -0.14 to 0.14 (M = 0.01, SD = 0.09). The independent samples t-test showed that the revision quality in post-editing NMT of literary text was significantly higher than that in revising HT of literary text ($t(14) = 4.44$, $p = 0.001 < 0.01$, Cohen's $d = 2.22$).

We also compared the revision quality between different text types within the same translation source. To be specific, post-editing NMT of news text showed lower revision quality than post-editing

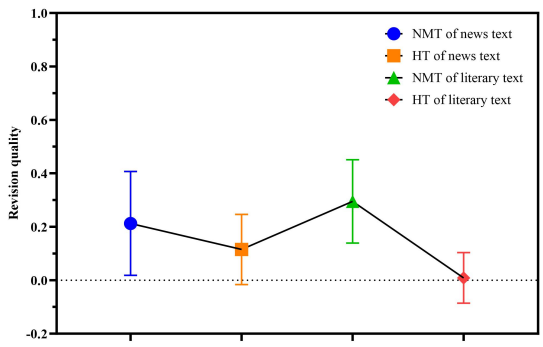

Figure 4: Revision quality in four scenarios

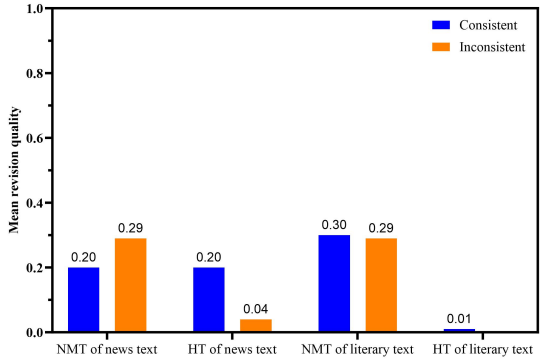

Figure 5: Mean revision quality across four scenarios in consistent and inconsistent situations

NMT of literary text, though the difference was not significant (t(14) = -0.94, p = 0.37, Cohen's d = 0.47). Similarly, revising HT of news text had higher revision quality than revising HT of literary text, but the difference was not significant (t(14) = 1.86, p = 0.09, Cohen's d = 0.09).

Due to the difference in the number of participants between the 'consistent' and 'inconsistent' situations, We only compared the mean revision quality across four scenarios in those two situations (see Figure 5). For post-editing NMT of news text, the revision quality in the consistent situation (M = 0.20) was lower compared to the inconsistent one (M = 0.29). However, the revision quality of revising HT of news text in the consistent situation (M = 0.20) was higher than that in the inconsistent situation (M = 0.04). Regarding post-editing NMT of literary text, the revision quality in the consistent situation (M = 0.30) was slightly higher than that in the inconsistent (M = 0.29). The lowest revision quality score occurred when revising HT of literary text (M = 0.01) in cases where participants' perceptions matched the actual source. No inconsistencies were found in HT of literary text.

5 Discussion

5.1 Accuracy rate of perceived sources of translations

This study aimed to investigate the ability of student translators to distinguish between NMT and HT for two distinct text types: news text and literary text. Research on earlier MT paradigms suggests that readers and translators are not always able to reliably distinguish between human and machine translations (He et al., 2010; Vasconcellos and Bostad, 2002). Our findings indicate that participants were quite adept at identifying NMT-translated texts, but for human-translated texts, the type of text appeared to play a significant role.

To be specific, all participants successfully recognized the human-translated literary text, while this was not the case for news text. One plausible reason is that, when translating texts with a higher degree of creativity, human translators can leverage their linguistic competence and personal aesthetic tendency (Chen, 2011) to translate more freely. NMT, however, often struggles with sentence comprehension in literary texts, failing to grasp the underlying semantic connotations (Moneus and Sahari, 2024). As a result, human-translated literary texts are relatively easy to distinguish.

Regarding the difficulty in identifying human-translated news text, it may stem from the nature of news text itself. News text, as a type of informative texts, is embodied by the large amount of terminology in a certain area (Reiss and Rhodes, 2014; Dai and Liu, 2024). Although it is crucial to ensure cultural elements in news texts align with the background of the target audience, information accuracy remains the top priority. Furthermore, Chinese news texts feature clear and concise sentence structures that align with English grammar, meaning that neither human translators nor NMT need to heavily adjust sentence structures when translating into English. Thus, it can be challenging to distinguish between HT and NMT, as both news translations often show similarities in vocabulary and sentence structure.

5.2 Criteria for determining sources of translations

For the criteria for determining the sources of translations, our findings indicated that participants relied heavily on translation strategies and, to a lesser extent, on translation errors. Specifically, participants judging NMT primarily noted literal trans-

lation strategies (particularly repetitive sentence structures and disjointed logical flow), as well as terminology errors. Those judging HT, however, reported free translation strategies (including omission, word class shifts, meaning-based sentence restructuring, and adjustments for logical cohesion), and technical errors (such as spelling errors).

Although these two criteria enable most participants to distinguish between NMT and HT, a small group including those who achieved high revision quality were misled by certain misconceptions. First, some believed that NMT systems show remarkable accuracy in terminology, but NMT and other AI technologies may generate mistranslations due to untimely updates to the databases (Zhu et al., 2024) or restrictions for preserving the privacy of sensitive information (Das et al., 2025). This suggests that terminology errors are not exclusive to HT, and terminology translation in NMT deserves particular attention during PE.

Second, NMT systems adopt an 'interpretive' translation strategy, producing overly wordy translations that resemble HT. It might confuse participants with a limited understanding of the differences between NMT and HT, leading them to mistakenly identify such translations as human-translated. These findings are in line with work on student conceptualisations of MT and HT, showing that students do not clearly distinguish between both or understand how MT works (Salmi et al., 2023). It is therefore important to integrate NMT into translator training programs and systematically enhance students' knowledge and critical awareness of both NMT and HT, including their respective strengths, limitations, and suitable application contexts (Li et al., 2025).

5.3 Revision changes

The revision changes to translation segments with errors were analysed first. The findings indicated that participants detected and corrected more errors in NMT than in HT, regardless of text types. A possible explanation is that, although the number of errors and their severity were comparable between NMT and HT in this study, the type of errors might influence participants' revision changes. Critical and accuracy errors are more evident than minor and fluency errors (Carl and Báez, 2019), leading to more opportunities for detection and correction. Thus, participants may have found NMT easier to revise than HT due to its higher frequency of accuracy errors. Moreover, this impact was especially significant in literary texts, as the number of accuracy errors in NMT-translated literary text was twice that in human-translated one.

For revision changes to error-free translation segments, the results revealed that participants were more likely to over-edit segments without errors, especially in NMT-translated literary text and human-translated news text. A possible reason is that the participants did not fully comply with the guidelines for post-editing and revision. These two guidelines required participants to retain as much of the original NMT and HT as possible; however, the results indicated that some participants did not adhere to the guidelines closely. It aligns with earlier observations (Mellinger and Shreve, 2016; Nitzke and Gros, 2020), which suggest that when the original translation conflicts with translators' personal preferences, they may struggle to follow the guidelines and tend to favour their own version, leading to the over-editing of error-free segments. Although most changes to error-free segments do not severely affect the final translation quality, they do increase the cognitive effort and should be avoided (Mossop, 2020).

5.4 Revision quality

As for revision quality, post-editing NMT resulted in higher quality score compared to revising HT, irrespective of text type. This difference in revision quality was significant in the case of literary texts, where participants performed significantly better in post-editing NMT than in revising HT. Given that revision quality is measured by 'necessary changes', 'overrevisions', and 'total translation errors', factors influencing revision changes, such as the types of errors and the translator's adherence to guidelines, may also impact the final revision quality.

An interesting result emerged when comparing revision quality in relation to whether participants' perceived translation sources were consistent with the actual ones. For NMT of news text, the revision quality in the inconsistent situation (revising NMT) was higher than that in the consistent situation (post-editing NMT). In contrast, for HT of news text, the revision quality in the consistent situation (revising HT) was higher than that in the inconsistent situation (post-editing HT). For NMT of literary text, a higher revision quality could be found in the consistent situation (post-editing NMT), compared to the inconsistent situation (revising NMT). The findings are partly in line with

the study by Daems and Macken (2020), where the revision quality in two inconsistent situations (revising NMT and post-editing HT) was higher than that in consistent situations (revising HT and post-editing NMT) respectively.

It suggests that the relationship between perception and quality might be mediated by text type. Participants were more likely to achieve higher revision quality not when their perceived translation sources diverged from the actual ones, but when they believed the news text was human-translated and the literary text NMT-generated. It might be attributed to two possible reasons. First, participants might exhibit less tolerance toward HT of news text and NMT of literary text. Their attitude toward HT of news texts is largely influenced by the complex nature of news translation (Yang et al., 2023) and their concerns that the translator's subjectivity might compromise the quality of news translation (Chen, 2011). Their distrust to NMT-translated literary texts may arise from the intricate nature of literary texts (Hu and Li, 2023) and the limitations of current NMT technology (Yu, 2022). However, their sceptical attitudes may foster critical thinking, motivating them to identify and correct errors and ultimately improving revision quality.

Second, this phenomenon can be interpreted through the lens of the Pygmalion Effect (Rosenthal and Jacobson, 1968), which posits that individuals' expectations about the outcome of a task shape their behavioural engagement in it. In HT of news text and NMT of literary text, participants likely held clearer expectations about the desired outcomes and felt more confident in addressing the task. This confidence could motivate them to allocate greater cognitive resources to editing the original translations, driven by their commitment to achieving the anticipated translation quality.

In addition, it is worth noting that the revision quality achieved by participants in this study did not exceed 0.30, lower than that in the study by Daems and Macken (2020). This discrepancy may stem from differences in language pair or translation experience. While Daems and Macken (2020) involved professional translators with rich experience, this study focused on student translators with limited translation experience. Professional translators are typically more adept at employing reading strategies to identify and correct translation errors effectively (Schaeffer et al., 2019), a skill that student translators may lack.

6 Conclusion

This study was conducted to explore student translators' ability to differentiate between NMT and HT across news text and literary text, how they justify their choices, and how they improve the translations. The findings suggest that participants were more adept at identifying NMT-translated news text and human-translated literary text. When determining the translation sources, participants relied heavily on translation strategies and, to a lesser extent, on translation errors. For revision changes, participants detected and corrected more errors in NMT than in HT, regardless of text types, and this difference in error detection and correction between NMT and HT was significant in the case of literary texts. Similarly, participants achieved a higher revision quality in post-editing NMT compared to revising HT, irrespective of text type. This difference in revision quality was significant in the case of literary texts. Furthermore, it was found that participants were more likely to achieve higher revision quality not when their perceptions diverged from the actual translation source, but when they believed the news text was human-translated and perceived the literary text as NMT-generated.

It is also important to note that some limitations should be taken into account when interpreting the findings. To begin with, the small sample size and inadequate sample representation limit the generalization of the findings. The findings of our cross-sectional study reflect students' ability to differentiate between NMT and HT across two text types at a specific point in time. Future studies could, therefore, adopt a longitudinal approach with larger and more diverse samples to track changes in students' ability over time and explore whether translation revision and post-editing practice can effectively enhance their revision and post-editing competence. Second, the text type might have a influence on participants' translation performance, which is worth further investigation.

In conclusion, although this study is exploratory, its findings aim to inspire further research into the potential of NMT for creative texts and the difference between translation revision and post-editing. Such investigations could significantly advance human-computer interaction research in both translation education and the industry.

References

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International conference on spoken language translation*, pages 280–290.

Michael Carl and M Cristina Toledo Báez. 2019. Machine translation errors and the translation process: a study across different languages. *The Journal of Specialised Translation*, 31:107–132.

Ya-Mei Chen. 2011. The translator's subjectivity and its constraints in news transediting: A perspective of reception aesthetics. *Meta*, 56(1):119–144.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences.

Gloria Corpas Pastor and Laura Noriega-Santiáñez. 2024. Human versus neural machine translation creativity: A study on manipulated mwes in literature. *Information*, 15(9):530.

Ella Creamer. 2024. Dutch publisher to use AI to translate 'limited number of books' into English. *The 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*, pages 53–78. Taylor & Francis.

Joke Daems and Lieve Macken. 2020. Post-editing human translations and revising machine translations. *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*, pages 50–70.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, 62(2):245–270.

Guangrong Dai and Siqi Liu. 2024. Towards predicting post-editing effort with source text readability: An investigation for english-chinese machine translation. *The Journal of Specialised Translation*, (41):206–229.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39.

Orphée De Clercq, Gert De Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. 2021. Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated french. *Translation Quarterly*, (101):21–45.

Félix Do Carmo and Joss Moorkens. 2020. Differentiating editing, post-editing and revision. In *Translation revision and post-editing*, pages 35–49. Routledge.

Sabrina Girletti. 2022. Working with pre-translated texts: Preliminary findings from a survey on post-editing and revision practices in swiss corporate in-house language services. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 271–280.

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

Ana Guerberof-Arenas, Susana Valdez, and Aletta G Dorst. 2024. Does training in post-editing affect creativity? *The Journal of Specialised Translation*, (41):74–97.

James Luke Hadley, Kristiina Taivalkoski-Shilov, Carlos SC Teixeira, and Antonio Toral. 2022. *Using technologies for creative-text translation*. Taylor & Francis.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Yifan He, Yanjun Ma, Johann Roturier, Andy Way, and Josef van Genabith. 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Kaibao Hu and Xiaoqian Li. 2023. The creativity and limitations of ai neural machine translation: A corpus-based study of deepl's english-to-chinese translation of shakespeare's plays. *Babel*, 69(4):546–563.

Kristian Tangsgaard Hvelplund. 2011. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Frederiksberg: Copenhagen Business School (CBS).

ISO17100. 2015. Translation services — requirements for translation services.

Arnt Lykke Jakobsen. 2018. Moving translation, revision, and post-editing boundaries. In *Moving boundaries in translation studies*, pages 64–80. Routledge.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine translation*, 23(4):241–263.

Kalle Konttinen, Leena Salmi, and Maarit Koponen. 2020. Revision and post-editing competences in translator education. In *Translation Revision and Post-Editing*, pages 187–202. Routledge.

Maarit Koponen, Brian Mossop, Isabelle S Robert, and Giovanna Scocchera. 2021. *Translation Revision and Post-Editing*. London: Routledge.

Bernhard Korte and Jens Vygen. 2018. *Combinatorial optimization: theory and algorithms*. Springer.

Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Ralph Krüger. 2022. Some translation studies informed suggestions for further balancing methodologies for machine translation quality evaluation. *Translation Spaces*, 11(2):213–233.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research*, 67:653–672.

Xiaoye Li, Xiangling Wang, and Wentian Lai. 2025. The usability of neural machine translation in creative-text post-editing: Evidence from users' performance and perception. *International Journal of Human-Computer Interaction*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Lieve Macken, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: Machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110. European Association for Machine Translation.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation*, pages 10–19.

Christopher D Mellinger and Gregory M Shreve. 2016. Match evaluation and over-editing in a translation memory environment. In *Reembedding translation process research*, pages 131–148. John Benjamins Publishing Company.

Ahmed Mohammed Moneus and Yousef Sahari. 2024. Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, 10(6).

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Brian Mossop. 2020. *Revising and Editing for Translators*. routledge.

Jean Nitzke and Anne-Kathrin Gros. 2020. Preferential changes in revision and post-editing. In *Translation revision and post-editing*, pages 21–34. Routledge.

Jean Nitzke and Silvia Hansen-Schirra. 2021. *A short guide to post-editing (Volume 16)*. Language Science Press.

Jean Nitzke, Silvia Hansen-Schirra, and Carmen Canfora. 2019. Risk management and post-editing competence. *The Journal of Specialised Translation*, 31(1):239–259.

Daniel Ortiz-Martínez, Jesús González-Rubio, Vicent Alabau, Germán Sanchis-Trilles, and Francisco Casacuberta. 2016. Integrating online and active learning in a computer-assisted translation workbench. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 57–76.

Marta Petrak, Mia Uremović, and Bogdanka Pavelin Lešić. 2022. Fine-grained human evaluation of nmt applied to literary text: case study of a french-to-croatian translation. In *Language Technologies and Digital Humanities Conference*, pages 141–146.

Thierry Poibeau. 2022. On" human parity" and" super human performance" in machine translation evaluation. In *Language Resource and Evaluation Conference*.

Anthony Pym. 2013. Translation skill-sets in a machine-translation age. *Meta*, 58(3):487–503.

Katharina Reiss and Eroll F Rhodes. 2014. *Translation criticism-potentials and limitations: Categories and criteria for translation quality assessment*. Routledge.

Celia Rico Pérez and Enrique Torrejón. 2012. Skills and profile of the new role of the translator as mt post-editor. *Tradumàtica*, (10):0166–178.

Isabelle S Robert, Aline Remael, and Jim JJ Ureel. 2017. Towards a model of translation revision competence. *The Interpreter and Translator Trainer*, 11(1):1–19.

Isabelle S Robert, Iris Schrijver, and Jim J Ureel. 2024. Measuring translation revision competence and post-editing competence in translation trainees: methodological issues. *Perspectives*, 32(2):177–191.

Isabelle S Robert, Iris Schrijver, and Jim JJ Ureel. 2023. Comparing l2 translation, translation revision, and post-editing competences in translation trainees: An exploratory study into dutch–french translation. *Babel*, 69(1):99–128.

Isabelle S Robert, Jim JJ Ureel, Aline Remael, and Ayla Rigouts Terryn. 2018. Conceptualizing translation revision competence: a pilot study on the 'fairness and tolerance'attitudinal component. *Perspectives*, 26(1):2–23.

Robert Rosenthal and Lenore Jacobson. 1968. Pygmalion in the classroom. *The urban review*, 3(1):16–20.

Andrew Rothwell, Andy Way, and Roy Youdale. 2023. *Computer-Assisted Literary Translation*. Taylor & Francis.

Mehmet Şahin and Sabri Gürses. 2019. Would mt kill creativity in literary retranslation? In *Proceedings of the qualities of literary machine translation*, pages 26–34.

Leena Salmi, Aletta G Dorst, Maarit Koponen, and Katinka Zeven. 2023. Do humans translate like machines? students' conceptualisations of human and machine translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 295–304.

Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2):26.

Moritz Schaeffer, Jean Nitzke, Anke Tardel, Katharina Oster, Silke Gutermuth, and Silvia Hansen-Schirra. 2019. Eye-tracking revision processes of translation students and professional translators. *Perspectives*, 27(4):589–603.

Lane Schwartz. 2014. Monolingual post-editing by a domain expert is highly effective for translation triage. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 34–44.

Giovanna Scocchera. 2020. The competent reviser: A short-term empirical study on revision teaching and revision competence acquisition. *The Interpreter and Translator Trainer*, 14(1):19–37.

Fedor Sizov, Cristina España-Bonet, Josef van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of llms, nmts, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199.

Arda Tezcan, Joke Daems, and Lieve Macken. 2019. When a 'sport'is a person and other issues for nmt of novels. In *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49.

Henry C. Thode. 2002. *Testing for Normality*. Marcel Dekker.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? *Translation quality assessment: From principles to practice*, pages 263–287.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, pages 2203–2213. Association for Computational Linguistics (ACL).

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232.

Muriel Vasconcellos and Dale A Bostad. 2002. Machine translation in a high-volume translation environment. In *Computers in Translation*, pages 78–97. Routledge.

Lyu Wang and Xiangling Wang. 2021. Building virtual communities of practice in post-editing training: A mixed-method quasi-experimental study. *The Journal of Specialised Translation*, 36:193–219.

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. In *Informatics*, volume 7, page 32. MDPI.

Yanxia Yang, Runze Liu, Xingmin Qian, and Jiayue Ni. 2023. Performance and perception: machine translation post-editing in chinese-english news translation by novice translators. *Humanities and Social Sciences Communications*, 10(1):1–8.

Yuxiu Yu. 2022. [retracted] english characteristic semantic block processing based on english-chinese machine translation. *Advances in Multimedia*, 2022(1):1458394.

Binghan Zheng, Sandra Báez, Li Su, Xia Xiang, Susanne Weis, Agustín Ibáñez, and Adolfo M García. 2020. Semantic and attentional networks in bilingual processing: fmri connectivity signatures of translation directionality. *Brain and cognition*, 143:105584.

Xiaoqian Zhu, Yanpeng Chang, and Jianping Li. 2024. A cross-institutional database of operational risk external loss events in chinese banking sector 1986–2023. *Scientific Data*, 11(1):939.

A Appendix

		Error types (based on MQM)			Number of errors	Error severity weight		Severity weight of errors
Text type	Translation source	Accuracy	Fluency	Style		Critical	Minor	
News text	HT	1	2	5	8	6	2	14
News text	NMT	2	1	5	8	6	2	14
Literary text	HT	2	0	6	8	6	2	14
Literary text	NMT	4	1	3	8	6	2	14

Figure 6: Error types for each text and translation source