ChatGPT's Self-Correction between Error Marking and Fine-Grained Taxonomy: Is Taxonomy Worth the Effort?

Adel Hezam [1,*]
Khalil A Nagi [2]

Received: 23 July 2025
Revised: 05 August 2025
Accepted: 06 August 2025

[1] bachelor's in English education and Master in Diaspora and Migration Studies, Department of Language Translation Center, University of Saba Region, Ma'rib, Yemen
[2] Assistant Professor of Linguistics, Department of English, University of Saba Region, Ma'rib, Yemen
* Corresponding author. E-mail. adeltaiz55@gmail.com , khalil.naji@usr.ac

https://doi.org/10.20428/jss.v31i7.3182

ChatGPT's Self-Correction between Error Marking and Fine-Grained Taxonomy: Is Taxonomy Worth the Effort?

Abstract:

This study aims to investigate external feedback strategies and their effects on ChatGPT's self-correction for the purpose of improving its translation performance. The study evaluates the effectiveness of the proposed strategies and points out which strategy is more effective. To achieve these objectives, the researchers built a test suite that is composed of 200 English sentences. These sentences were translated into Arabic by ChatGPT using a default translation prompt. The translated sentences were manually annotated, and an error marking and an error taxonomy were provided. The researchers, then, selected the 60 sentences with the most errors to be retranslated using self-correction strategies. A professional manual evaluation of the initial translation and the retranslation was performed to evaluate the effectiveness of the used strategies. The study investigated whether the effort spent on error classification is needed to get a better translation when using ChatGPT's self-correction strategy. The study is very important for translators, post-editors, researchers, and developers of MT, and it provided recommendations for future work.

Keywords: *self-correction, feedback, error marking, error taxonomy, ChatGPT.*

التصحيح الذاتي لـ ChatGPT بين تحديد الأخطاء والتصنيف الدقيق لها: هل يستحق تصنيف الأخطاء كل هذا الجهد

عادل محمد قائد حزام [1،*]
خليل عبدالسلام خالد ناجي [1]

الملخص:

تسعى هذه الدراسة إلى اختبار استراتيجيات التغذية الراجعة الخارجية وتأثيراتها على التصحيح الذاتي لـ ChatGPT، وذلك بهدف تحسين أداء الترجمة. تقيّم الدراسة فاعلية الاستراتيجيات المقترحة وتحدد الاستراتيجية الأكثر فاعلية. لتحقيق هذه الأهداف، طوّر الباحثان مجموعة اختبارات مكونة من 200 جملة إنجليزية، تم ترجمتها بواسطة ChatGPT إلى اللغة العربية باستخدام موجه ترجمة افتراضي. تم بعد ذلك استكشاف وتصنيف الأخطاء في الجمل المترجمة من قبل مترجمين خبراء. اختار الباحثون 60 جملة والتي تحتوي على أكبر عدد من الأخطاء لإعادة ترجمتها باستخدام استراتيجيات التصحيح الذاتي. تم تقييم الترجمة الأولية والترجمة المعادة من قبل فريق الخبراء وذلك لغرض تقييم فاعلية الاستراتيجيات المستخدمة. كما تناولت الدراسة ما إذا كان الجهد المبذول في تصنيف الأخطاء ضرورياً للحصول على ترجمة أفضل عند استخدام استراتيجية التصحيح الذاتي لـ ChatGPT. تعد هذه الدراسة مهمة جداً للمترجمين، ومدققي الترجمة، والباحثين، ومطوري الترجمة الآلية. كما قدمت الدراسة عدد من التوصيات.

الكلمات المفتاحية: ***التصحيح الذاتي، التغذية الراجعة، تحديد الأخطاء، تصنيف الأخطاء، ChatGPT.***

[1] **بكالوريوس انجليزي تربية وماجستير الهجرة والشتات، قسم اللغة والترجمة جامعة إقليم سبأ، مارب، اليمن**
[2] **أستاذ اللغويات المساعد، قسم اللغة الإنجليزية، جامعة إقليم سبأ، مارب، اليمن**
[*] **عنوان المراسلة:** adeltaiz55@gmail.com , khalil.naji@usr.ac

https://doi.org/10.20428/jss.v31i7.3182

Introduction

Machine translation (MT) is considered to be alternative to human translation that is both inexpensive and time-saving. Nevertheless, translation outputs should meet the quality requirements. It is crucial that the quality of machine translation meet certain standards, and accordingly the field of translation quality and its evaluation becomes a very interesting field of research. The performance of machine translation systems should be thoroughly investigated; strategies to improve the outputs of machine translation should be proposed. In short, investigating and finding ways to improve machine translation outputs must be the focus of the recent research of machine translation.

Heated discussions regarding the quality of machine translation still form the focus of many research works. The proposals in the literature fall between proposals that see machine translation outputs as ones that are on par with professional human translation outputs (Hassan et al., 2018; Barrault et al., 2019) and proposals that affirm that such parity has not been achieved and professional human translation outputs are still superior to the machine translation outputs (Läubli et al., 2018; Toral et al., 2018; Freitag et al., 2021).

A fact that cannot be denied is that machine translation is advancing quickly, and the translation outputs provided by machine translation are of high quality. There is still a gap between machine translation and professional human translation outputs, and error analysis studies have been performed recently that have come out with a comparatively long list of translation errors in MT outputs (Popović, 2021; Kocmi, 2022; among others).

With the advent of new large language models (LLMs) recently, there is a new wave of research that investigates the quality of translation provided by the new models such as ChatGPT and Gemini. Recent studies have shown that LLMs' translation outputs are still lacking in performance and are still riddled with many errors (Zhu et al., 2023; Nagi et al., 2024). Accordingly, post-editing efforts are required. Recent studies have also shown that LLMs have the capability of post-editing when provided with external feedback (Chen et al., 2023; Raunak et al., 2023; Feng et al., 2024; and Nagi et al., 2024, among others).

ChatGPT is one of the most prominent LLMs. However, when it comes to translation, the model still suffers from the recurrence of translation errors, especially when translating from English to Arabic due to the wide divergence in morphological and syntactic structures. It is also because of the fact that there are not many annotated Arabic corpora. Due to that, post-editing effort is required, which can be minimized by inducing high-quality translation using self-correction strategies. Effective strategies need to be proposed. Therefore, using external feedback, the study will investigate ChatGPT's capability to self-correct or post-edit and improve its translation performance.

The study aims, therefore, to evaluate the effectiveness of the proposed strategies and point out whether error marking or fine-grained taxonomy is more effective as external feedback in the self-correction process. This study is the first that investigates the capacity of self-correction of ChatGPT in the case of translating English sentences to Arabic. Translation errors are marked and classified, and both error marking and error classification are used as feedback to initiate ChatGPT's self-correction process. This is crucial to the development of automatic translation. Therefore, the study will be a great addition to the literature of automatic translation and translation in general. To the best of the researchers' knowledge, it is the first study related to Arabic translation that investigates various feedback strategies of ChatGPT to perform self-correction.

The study is, however, limited to English/Arabic translation using ChatGPT. The test suite is also restricted to a limited number of sentences. Nonetheless, it should be noted here that the researchers have ensured that the test suite includes various types of sentences with different structures. While the proposed self-correcting strategies may apply to all languages, the nature of detected errors may differ when it comes to other languages. This study examines two strategies of external feedback, and self-provided feedback will not be examined here.

Literature Review

ChatGPT

ChatGPT is considered to be the most well-known recent artificial intelligence (AI) application. It gained its popularity even before its official launch. Various companies and news agencies like BBC, CNN, and People's Daily announced the forthcoming AI revolution, which made ChatGPT quickly rise in popularity. ChatGPT also became popular because it is capable of performing various tasks in an effective way. These tasks include translating languages, generating text, classifying text, answering questions, and writing code (Siu, 2023).

Jiao et al. (2023) and Hendy et al. (2023) indicated that the translation performance of ChatGPT comes on par with the performance of the previous commercial translation systems such as Google Translate when translating a high-resource European language. It, however, lags behind when translating a low-resource language. These studies also showed that ChatGPT struggles with translating scientific texts, such as biomedical abstracts, but it performs well when translating spoken language. Jiao et al. (2023) also stated that GPT-3.5 does not perform well in specific domains, but its performance in the translation of spoken languages appears to be of good quality. It also stated that when it comes to low-resource languages, GPT-4 does not achieve the desired level of performance despite the fact that its multilingual translation capabilities are improving (Zhu et al., 2023). ChatGPT struggles when it translates specialized texts, such as scientific texts, legal texts, and literary texts, but it performs well with simple content when Arabic is

involved (Khoshafah, 2023). Nagi et al. (2024) also stated that ChatGPT translation outputs exhibit a high-level error frequency when it translates complex English sentences into Arabic. Alzain et al. (2024) also compared the translation performance of both ChatGPT and Google Translate, pointing out that Google Translate outperforms ChatGPT when translating scientific texts.

Prompts and Self-Correction

A prompt is identified as "a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities" (White et al., 2023). Using prompts has also been proven to be effective in improving the LLMs' output quality, and there is a direct relationship between the prompt quality and the output quality. Using specific or informed prompts can improve the output effectively (White et al., 2023; Giray, 2023; Liu et al., 2023, among others).

In the field of machine translation, there has been limited research on how effective the prompt engineering is on the translation outputs. A few prompting strategies in literature have been proven to be effective when used to improve translation output (Jiao et al., 2023; Gao et al., 2023; Siu, 2023; Nagi et al., 2024).

In the literature, it has been stated clearly that special prompting strategies contribute greatly to the quality of the translation produced by LLMs. Jiao et al. (2023) have suggested a prompting strategy that they dubbed the pivot strategy. According to this strategy, it is required that ChatGPT translate a source text into a high-source pivot language before translating it into the target language. They have stated that this strategy improves the translation quality significantly.

Gao et al. (2023) have also pointed out that ChatGPT surpasses commercial translation systems when properly designed prompts are used. In their study, they have proposed prompts that contain translation task information (both the target language and the source language are identified), context domain information (the domain of the text is identified, such as news, legal, etc.), or part-of-speech tags. It has been indicated that the results of the study have shown that the proposed prompts significantly enhance the performance of ChatGPT in translation. It has also been indicated that prompts with contextual information enable ChatGPT to produce improved translation output (Siu, 2023).

An interesting study in this aspect has been conducted by Gu (2023), in which linguistically informed prompts have been introduced and used in the translation of Japanese attributive clauses into Chinese. It has been stated that such prompts improve the translation accuracy by more than 35%.

There is also a very interesting recent approach to informed prompting in LLM translation, which is the use of a self-correction strategy where the model is asked to modify the original translation using a proposed strategy. Chen et al. (2023), Raunak et al. (2023), and Feng et al. (2024) have used adopted strategies that depend on prompting the model to self-correct its previous translation. Nagi et al. (2024) have also used self-correction strategies to improve ChatGPT translation performance. This approach has been proven to be effective, and better translation outputs have been produced by the investigated LLMs using such an approach.

Since the use of the self-correction approach to improving the translation output seems promising, further investigation is required. New research will help propose more effective and stable strategies. This study, therefore, is an endeavor in examining strategies that improve LLMs translation performance within the framework of the self-correction approach.

To the best of the researchers' knowledge, this will be the first study that examines the capacity of ChatGPT's self-correction using different feedback strategies that improve the Arabic translation of English texts.

Human Evaluation

Despite the fact that several automatic evaluation techniques are presented in the literature, human evaluation remains the main and the most accurate technique. Automatic evaluation is undoubtedly important. However, this study will focus not only on the evaluation of machine translation but also on how to use annotating error as feedback to LLMs and get them to produce an error-free translation. The researcher, therefore, will use the feedback provided by two error annotation strategies to exploit the LLMs' translation performance to the fullest. These annotation strategies are the Error Span Annotation (ESA) and the Multidimensional Quality Metrics (MQM).

Error Marking

Error marking is a human annotation method where the annotators mark the problematic parts in MT outputs without assigning error labels. According to Kreutzer et al. (2020), this method substantially reduces the annotation effort. Popović (2020) has also argued that this method is more advantageous than error classification since it is both informative and less demanding. Such an annotation method can be used further to provide different types of analyses (Popović, 2020).

Kocmi et al. (2024) also discussed what is called error span annotation (ESA), where translation errors are marked and either a major or a minor severity level is assigned to the annotated errors. According to Kocmi et al. (2024), errors with a major severity level may change the meaning, make the text difficult to read, or

decrease its usability. Errors with a minor severity level, on the other hand, are related to style, grammar, or lexical choice. Missing errors are also indicated if something is missing from the translated text.

Multidimensional Quality Metrics

The study performs error taxonomy. The taxonomy of the annotated errors in the study is guided by the one provided by Multidimensional Quality Metrics (MQM) introduced in Lommel et al. (2014). The typology of errors provided by MQM classified translation errors into eight dimensions: terminology, accuracy (adequacy), linguistic conventions (fluency), style, locale conventions, audience appropriateness, design and markup, and custom. Such dimensions are defined and classified further. For instance, the main subdimensions of accuracy are mistranslation, addition, and omission, whereas the main subdimensions related to linguistic conventions are grammar, punctuation, and spelling. These subdimensions are divided further. For a complete classification of errors under MQM, check out https://themqm.org/the-mqm-full-typology/. Therefore, it can be stated that MQM provides a fine-grained analysis of translation errors. In addition, the MQM framework is flexible, and it is well-established in the literature.

Methodology and Results

The methodology and results are presented in this section of the study. It covered the datasets utilized, the methods for error analysis, and the outcomes of error annotation. Additionally, it included the prompting strategy, the assessment of the initial translation, and the evaluation of the retranslation after using the suggested prompting strategy.

The Main Dataset

The researcher creates a dataset consisting of 200 sentences with diverse structural patterns. The sentences are selected from recent news essays to make sure they are not included in the ChatGPT training data. The selected sentences are translated by ChatGPT-4 using the default prompt, Translate these sentences into Arabic.

Error Taxonomy

The sentences of the main dataset are annotated by 3 professional annotators. Errors are marked and then classified according to Multidimensional Quality Metrics (MQM), which was first described in Lommel et al. (2014). The annotated errors fall under four MQM dimensions. These dimensions are the terminology, accuracy (adequacy), linguistic conventions (fluency), and style. They are further classified into different sub-categories as represented in Table 1 below.

Table 1: Types and Number of Annotated Errors in ChatGPT Original Translation

Types of Errors		**Number of Errors**
Terminology	Wrong Term	82
	Total Terminology Errors	82
Accuracy	Omission	53
	Ambiguous target content	4
	Ambiguous source content	1
	Overly literal	8
	Untranslated parts	4
	Addition	9
	Total Accuracy Errors	79
Fluency	Incorrect word form	57
	Incorrect FW	27
	Missing FW	95
	Extraneous FW	30
	Cohesion	30
	Incorrect Argument	4
	Word Order	51
	Punctuation	64
	Total Fluency Errors	358
Style		17
Total Number of Errors		536

Error Distribution

The error taxonomy, as shown in Table 1, indicates that ChatGPT translation outputs are still full of various types. Since the test suite is composed of 200 sentences, the error frequency, then, amounts to 2.68 errors per sentence, which is very high. It should be noted that, according to MQM main dimensions, the errors are distributed as follows.

- Fluency errors come in front with 358 errors (66.79% of the annotated errors).
- Terminology errors follow with 82 errors (15.3% of the annotated errors).
- Accuracy errors are next with 79 errors (14.74% of the annotated errors).
- Last come the style errors with 17 errors (3.17% of the annotated errors).

The distribution of errors according to MQM main dimensions is represented in Figure 1 below.

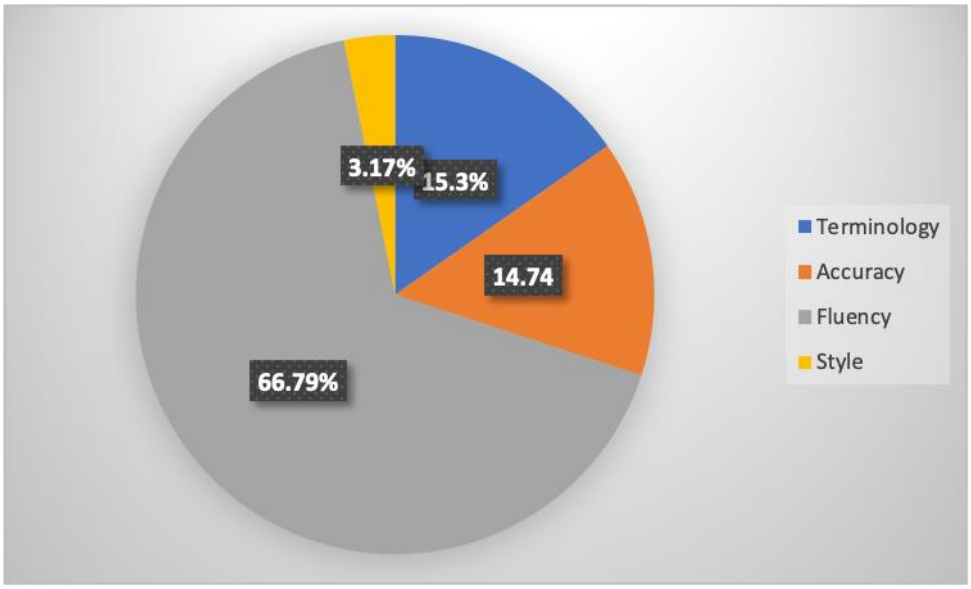

Figure 1: The Distribution of Errors Following MQM Main Dimensions

Self-Correction Test Suite and Process

Following the error annotation process, the researchers selected 60 sentences with the most errors. These sentences are extracted from the main test suites to be used in the various self-correction processes.

The extracted sentences are divided into two equivalent test suites with 30 sentences each to be used in both the error marking or error taxonomy feedback and processes are used as the external feedback of ChatGPT's self-correction process. This data is used along with the source sentences and original translations in the process. Based on that, ChatGPT is prompted to provide a better translation. The new translation is evaluated and compared with the initial translation. The Self-correction process is achieved through the following processes: self-correction with error marking and self-correction with error taxonomy.

Self-Correction with Error Marking

The first extracted 30 sentences of the 60 sentences that contained the most translation errors are in this process. Each sentence in this suite was processed by providing the source text, the initial Arabic translation, and a list of translation errors without categorization. The following format illustrates this process.

Source text: If it fails to do so, it will risk becoming thoroughly irrelevant and condemning Mozambique to being reduced to a failed state in the not so distant future.

Initial Translation: إذا فشلت في ذلك، فإنها تخاطر بأن تصبح غير ذات صلة تمامًا، مما قد يدفع موزمبيق إلى أن تصبح دولة فاشلة في المستقبل القريب.

Initial Translation Errors: "إذا", "تخاطر", "صلة ", "تمامًا ", "مما ", "قد يدفع "

Based on reviewing the initial translation errors, please provide the final Arabic translation.

Self-Correction with Error Taxonomy

This process focuses on the second test suite, which also includes 30 sentences translated using ChatGPT's Self-Correction prompt. In contrast to the prior suite, the sentences in this set were all modified in accordance with a more organized error analysis that includes an error taxonomy. The kinds of errors made in the original Arabic translation are categorized in this taxonomy (e.g., word order, cohesion, function words, terminology). The process once more includes the source text, the initial translation, the classified errors, and the revised final translation. The example below demonstrates this approach.

Source text: On the day she was killed, she was with several colleagues in a safe area away from clashes and crossfire, although there was an Israeli army convoy about 200 metres (660 feet) away.

Initial Translation: في اليوم الذي قُتلت فيه، كانت مع عدة زملاء في منطقة آمنة بعيدًا عن الاشتباكات وإطلاق النار، رغم أن هناك قافلة للجيش الإسرائيلي كانت على بُعد حوالي 200 متر (660 قدمًا).

Initial Translation Errors: "في اليوم الذي قُتلت فيه، كانت / word order", "كانت / cohesion", "رغم / two missing function words", "أن / incorrect function word", "هناك / wrong term", "كانت / extraneous function word",

Based on reviewing the initial translation errors, please provide the final Arabic translation.

Evaluation

To evaluate the effectiveness of the used feedback strategies, manual evaluation of the initial translation and the retranslation will be performed by the three professional annotators using a secular quality metric (Freitag et al., 2021). This metric uses a 0-6 Likert-like scale. Its ranks are as follows:

- **6: Perfect Meaning and Grammar:** The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.
- **4: Most Meaning Preserved and Few Grammar Mistakes:** The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.

- **2: Some Meaning Preserved:** The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- **0: Nonsense/No meaning preserved:** Nearly all information is lost between the translation and source. Grammar is irrelevant.

The capacity of self-correction was evaluated based on the improvement rate of the retranslation. Accordingly, the most effective feedback is decided. The result indicates both feedback strategies showed a marked improvement in the translation. The following table shows the evaluation of the original translation as well as the retranslation outputs and the improvement rate.

Table 2: Evaluation of the Original Translation and Retranslation

Feedback Strategy	**Original Translation**	**Retranslation**	**Improvement**
Error Marking	61.67	75.56	13.89
Error Taxonomy	66.39	78.33	11.94

The table above displays the results of an automatic evaluation to measure the quality of original translations and re-translations under two feedback strategies, error marking and error taxonomy. It also shows the improvement rate of initial translation that results from both strategies.

The result shows that the quality of translation significantly improved as a result of using the error-marking feedback strategy. The original translation score (61.67) indicates moderate accuracy. After providing feedback, the score increased to 75.56. The 13.89-point improvement shows that the error-marking strategy is effective in recognizing and correcting errors. However, despite this significant improvement, the final score indicates that the result was good but not perfect.

Similarly, the original score for error taxonomy feedback was higher, maybe because of a better initial comprehension. The retranslation score of 78.33 indicated high accuracy. The 11.94-point improvement confirms the effectiveness of this method. Although the extent of improvement is slightly less than the error-marking strategy, the final score is stronger and approaches a very good level.

The improvement of using both feedback strategies shows that error marking has achieved good improvement and the final quality is moderate. On the other hand, error taxonomy has also achieved a good improvement, but the final quality of translation is higher. The following diagram shows such improvement.

Figure 2: The Improvement of Retranslation Using Both Feedback Strategies

Discussion

This study explored how two external feedback strategies, namely, the error marking feedback strategy and the error taxonomy feedback strategy, affect ChatGPT's ability to self-correct in translation tasks. According to the findings, both feedback strategies resulted in significant improvements in ChatGPT's translation performance. It is shown that ChatGPT can make significant self-corrections and that it responds well to external feedback. Error marking led to a faster rate of improvement since it requires less effort and less time, which makes it ideal for immediate error correction. Error taxonomy feedback, on the other hand, resulted in a greater final quality. This proves its value in error correction. However, it should be mentioned that error taxonomy feedback requires more effort and time.

Despite these improvements, the ultimate translation scores are still below 80, indicating that there are shortcomings in attaining expert post-editing level. The results also indicated that there is little difference in the evaluation of retranslations where the error taxonomy feedback has achieved a higher translation quality. Despite the fact that the difference is not very high, it still indicates that the error taxonomy feedback strategy has achieved better results, which can be improved further by investigating more prompting strategies.

In general, the research shows that such external feedback strategies improve the ChatGPT translation quality, but not to the point where it can take the place of professional human translators.

Conclusion

This study investigated the efficiency of external feedback strategies in improving ChatGPT's translational accuracy. The findings indicate that external feedback can significantly improve ChatGPT's performance. The model, though, is not yet able to create translations at the professional level on its own due to its shortcomings in contextual understanding, semantic depth, and stylistic elegance. Although both Error Taxonomy and Error Marking improved translation quality, neither method produced flawless translations, demonstrating the ongoing need for advancement in AI language modeling and feedback integration. Nevertheless, ChatGPT's responsiveness to feedback points to its potential as an instructional resource or a helpful translator, particularly in collaborative or guided environments. Future studies should concentrate on evaluating cross-linguistic effectiveness, creating adaptive feedback systems, and enhancing ChatGPT's contextual understanding in order to broaden the use of AI-assisted translation in both academic and professional contexts.

The study recommends that more research be done in this aspect to come up with new prompting and feedback techniques. This can lead to achieving more accurate translation outputs.

Acknowledgment

This research received grant no. (523/2024) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

Alzain, E., Nagi, K. A., & Algobaei, F. (2024). The quality of Google Translate and ChatGPT English to Arabic translation: The case of scientific text translation. *Forum for Linguistic Studies, 6*(4), 837–849. https://doi.org/10.30564/fls.v6i3.6799

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... & Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 1–61). http://www.statmt.org/wmt19/pdf/53/WMT01.pdf

Chen, P., Guo, Z., Haddow, B., & Heafield, K. (2023). Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*. https://doi.org/10.48550/arXiv.2306.03856

Feng, Z., Zhang, Y., Li, H., Liu, W., Lang, J., Feng, Y., Wu, J., & Liu, Z. (2024). Improving LLM-based machine translation with systematic self-correction. *arXiv preprint arXiv:2402.16379*. https://doi.org/10.48550/arXiv.2402.16379

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics, 9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437

Gao, Y., Wang, R., & Hou, F. (2023). How to design translation prompts for ChatGPT: An empirical study. *arXiv preprint arXiv:2304.02182*. https://doi.org/10.48550/arXiv.2304.02182

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, 1–5. https://doi.org/10.1007/s10439-023-03272-4

Gu, W. (2023). Linguistically informed ChatGPT prompts to enhance Japanese Chinese machine translation: A case study on attributive clauses. *arXiv preprint arXiv:2303.15587*. https://doi.org/10.48550/arXiv.2303.15587

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*. https://doi.org/10.48550/arXiv.1803.05567

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*. https://doi.org/10.48550/arXiv.2302.09210

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*, *1*(10). https://doi.org/10.48550/arXiv.2301.08745

Khoshafah, F. (2023). ChatGPT for Arabic-English translation: Evaluating the accuracy. *ResearchSquare*. https://doi.org/10.21203/rs.3.rs-2814154/v1

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... & Popović, M. (2022). Findings of the 2022 Conference on Machine Translation

(WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 1–45). https://aclanthology.org/2022.wmt-1.1

Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., ... & Shmatova, M. (2024). Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*. https://doi.org/10.48550/arXiv.2406.11580

Kreutzer, J., Berger, N., & Riezler, S. (2020, November). Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 135–144). https://aclanthology.org/2020.eamt-1.15

Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). https://doi.org/10.18653/v1/D18-1512

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, (12)*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Nagi, K. A., Alzain, E., & Naji, E. (2024). Informed prompts and improving ChatGPT English to Arabic translation. *Al-Andalus Journal for Humanities & Social Sciences, 98*(11). https://doi.org/10.35781/1637-000-098-007

Popović, M. (2020). Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5059–5069). https://doi.org/10.18653/v1/2020.coling-main.444

Popović, M. (2021). On nature and causes of observed MT errors. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)* (pp. 163–175). https://aclanthology.org/2021.mtsummit-research.14

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12009–12024). https://aclanthology.org/2023.findings-emnlp.804

Siu, S. C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. *SSRN*. http://dx.doi.org/10.2139/ssrn.4448091

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 113–123). https://aclanthology.org/W18-6312

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., ... & Huang, S. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *arXiv e-prints, arXiv:2304.04675*. https://doi.org/10.48550/arXiv.2304.04675