Effectiveness of ChatGPT in Korean Grammatical Error Correction

Junghwan Maeng

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
junghwan.maeng@polyu.edu.hk

Jinghang Gu

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
gujinghangnlp@gmail.com

Sun-A Kim

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
sun-a.kim@polyu.edu.hk

Abstract

As the first study in applying ChatGPT to L2 Korean grammar correction and learning, this study investigates the effectiveness of ChatGPT models as tools for Korean Grammatical Error Correction (GEC). The model comparison revealed that ChatGPT 4 outperforms ChatGPT 3.5 and the baseline model in all measures related to the precision of error detection and correction. Furthermore, a human evaluation shows that ChatGPT 4 outperforms its previous version in handling multiple grammatical errors and correcting semantic-level mistakes. The findings of the present study suggest that ChatGPT 4 as a self-learning tool would be more suitable for advanced to near-native level L2 learners of Korean to improve their semantic fluency of sentences with minimal errors than for less proficient L2 learners because the latest version of ChatGPT 4 still demonstrates a relatively lower accuracy rate in Korean GEC tasks.

1 Introduction

Recent years have witnessed remarkable progress in the field of large language models (LLM), particularly with the emergence of Generative Pre-trained Transformer (GPT) models. Among these, ChatGPT by OpenAI stands out as an effective tool for a variety of Natural Language Processing (NLP) tasks such as machine translation (Jiao et al., 2023; Hendy et al., 2023), question-answering (Bang et al., 2023), and text summarization (Yang et al., 2023).

Research has shown that ChatGPT can outshine existing models in grammatical error correction (GEC) (Fang et al., 2023), implying its potential as a valuable tool for second language (L2) learners striving to enhance their writing accuracy and fluency. ChatGPT has at least two advantages as an L2 self-learning tool. First, the feature of providing immediate feedback enables ChatGPT to foster a learning environment that is essential for self-directed L2 learning, particularly in improving grammatical accuracy. Second, ChatGPT exhibits an exceptional ability to understand contextual information (Brown et al., 2020), a critical aspect required for generating contextually appropriate GEC. While the role of ChatGPT as a GEC tool has garnered much interest, its impact on language learning awaits further exploration.

This study focused on the application of ChatGPT for Korean GEC tasks, aiming to assess whether the currently available versions of ChatGPT (ChatGPT 3.5-Turbo and ChatGPT+ 4) can serve as a reliable GEC instrument in Korean. It also considered the educational implications of ChatGPT to enhance L2 grammatical accuracy and writing fluency.

While previous studies have demonstrated the effectiveness of ChatGPT as a GEC system in English, German, and Chinese (e.g., Fang et al., 2023; Wu et al., 2023), its applicability to Korean, which is a highly agglutinative language, has little been explored to date. Recent Korean GEC studies have mostly proposed their own GEC models equipped with an enhanced performance relative to the existing Korean GEC service called Hanspell, which mainly focuses on spelling correction (Lee et al., 2021; Yoon et al., 2022). In order to test the validity of ChatGPT as a GEC tool, the present study compared two versions of ChatGPT (ChatGPT 3.5-Turbo and ChatGPT+ 4). Additionally, the current study conducted a statistical analysis on a human evaluation of the ChatGPT-based GEC to determine if ChatGPT+ 4 surpasses ChatGPT 3.5-Turbo in error correction. Considering that GEC is a task of correcting errors committed in spelling, grammar, and word choices (Ruder, 2022), grammatical errors were categorized in the human evaluation into four types: spelling, particles, conjugation, and expressions. *Spelling errors* involve simple typo mistakes, while *particle errors* refer to morphosyntactic errors associated with case marking. *Conjugation* errors refer to incorrect uses of verb inflection involving both morphosyntactic information (e.g., tense inflection) and semantics-syntax interface (e.g., clause connectives). *Expression errors* are mainly semantic-level errors where certain words need to be replaced by the ones that are more context-appropriate.

Even at advanced proficiency levels, L2 learners often experience persistent difficulties in good command of L2 vocabulary. This is partly due to their tendency to rely on L1 equivalents for selecting words in L2 (Jiang, 2000, 2002, 2004). This tendency can result in inappropriate word choices, making semantically unnatural sentences in their L2 production. Considering the persistent challenges of L2 acquisition in the lexical-semantic domain, the human evaluation-based analysis in the present study investigated the efficiency of current ChatGPT models in managing semantic-level grammatical errors in L2 written production. Based on the results of the human evaluation analysis, this study also considered a pedagogical implication of using ChatGPT as an instructional tool for writing in L2 Korean.

2 ChatGPT as a GEC Tool

Previous studies on using ChatGPT in GEC tasks have mainly concentrated on English as the target language and compared its performance to existing GEC models (e.g., Coyne et al., 2023; Fang et al, 2023; Wu et al., 2023). They consistently found that ChatGPT tends to over-correct but surpasses existing models in the domain of fluency. It achieves this by enhancing the readability of incorrect sentences, making them sound more natural.

Wu et al. (2023) evaluated ChatGPT (the version not mentioned in the paper) against Grammarly (as a commercially available GEC tool) and GECToR (as a state-of-the-art model) in English GEC tasks by comparing the performance of the three tools. Their findings indicated that ChatGPT was less efficient than Grammarly and GECToR in GEC tasks. Notably, ChatGPT's performance, measured by F0.5 metrics, declined significantly compared to Grammarly and GECToR, as sentence length increased. Overall, the study found that Grammarly consistently performed well in GEC tasks. GECToR, on the other hand, appeared to favor correcting only those errors it was confident with. Meanwhile, ChatGPT exhibited a tendency to make more corrections than necessary but still maintained grammatical accuracy. For instance, it would change "helpful for family potential disease" to "helpful in preventing potential family diseases." Nonetheless, the human evaluation conducted in this study indicated that the propensity of ChatGPT for over-correction underscores its potential as a valuable instrument for GEC tasks. The results demonstrated that compared to others, ChatGPT had the least instances of under-corrections and that its rate of mis-corrections was lower than the state-of-the-art model (GECToR).

Fang et al. (2023) also compared ChatGPT 3.5-Turbo to SOTA-based models and Grammarly and found that ChatGPT 3.5-Turbo corrected more than what is identified as grammatical errors. When tested on the M^2 scoring, ChatGPT 3.5-Turbo scored the highest in the recall value but the lowest in the precision and $F_{0.5}$ values, indicating its tendency for over-correction in GEC tasks. The study also displayed the importance of providing sufficient details in a prompt to elicit responses without generating superfluous comments. The application of the zero-shot CoT method (Kojima et al., 2022), which employs special tags to denote input sentences (e.g., <input>) and output sentences (e.g., <output>), significantly enhanced ChatGPT's performance in GEC tasks. This approach surpassed the performance of the plain zero-shot method, suggesting the value of using these specific tags to guide the model's response. The zero-shot method in utilizing ChatGPT has merit in improving text fluency, as its GEC performance was on par with the baseline SOTA model and outperformed the human evaluation. Fang and his colleagues also showcased the potential of ChatGPT as an effective tool for multilingual GEC tasks. They found that, under the zero-shot CoT method, ChatGPT surpassed Transformer-based models in Chinese and German GEC tasks. The findings from the multilingual GEC tasks suggest that ChatGPT can be utilized as a useful instrument in languages that are typologically different from Indo-European languages.

Within ChatGPT, a recent comparison between the capabilities of ChatGPT 3.5 and ChatGPT 4 (Coyne et al., 2023) revealed that ChatGPT 4 exceeded ChatGPT 3.5 in sentence revision tasks that emphasize editing fluency. Both versions were tested using the JFLEG dataset, designed for assessing fluency in GEC tasks (Napoles et al., 2017), where they outperformed the baseline ELECTRA-VERNet model. The GLEU score, a measure based on n-gram overlap between the corrected and reference sentences (Napoles et al, 2016), was higher for ChatGPT 4 than for ChatGPT 3.5. This suggests that ChatGPT 4 has better fluency-enhancing capabilities compared to its previous version. Conversely, when evaluated using the BEA-2019 dataset, which emphasizes minimal edits and prioritizes correction of detected errors without unnecessary modifications, both ChatGPT 3.5 and ChatGPT 4 fell short compared to the baseline models (GECToR+BIFI & ELECTRA-VERNet). This was reflected in their lower $F_{0.5}$ scores, a measure that is given more weight on the precision value in GEC. The findings from both the BEA-2019 and JFLEG datasets collectively substantiate that ChatGPT models tend to correct more than errors detected within sentences, which, for the most part, results in improved fluency in revised sentences.

In brief, prior studies consistently highlighted ChatGPT's capability to enhance the fluency or naturalness of sentences in GEC tasks. The strength of ChatGPT as a fluency revision tool implies that it can serve as an effective instrument for L2 learners in revising errors in the semantic domain. Considering that the lexical-semantic domain presents considerable challenges to L2 acquisition (Jiang, 2000, 2002, 2004), ChatGPT has great potential to be a self-directed learning instrument for L2 learners who aim to improve their written proficiency by revising word choices. Yet, literature has also shown that ChatGPT may not be as efficient as existing GEC models in terms of minimal grammatical edits, as it scored low in the precision and $F_{0.5}$ values. As the first study in applying ChatGPT to L2 Korean learning, the present study first aimed to establish whether ChatGPT can serve as a reliable instrument for Korean GEC tasks by using a KoBART model selected from Yoon et al. (2022) as the baseline. Furthermore, the present study compares ChatGPT 3.5 and ChatGPT 4 based on a human evaluation to confirm whether the latest version offers enhanced semantic-level error corrections.

3 Experiment Setup

3.1 Dataset

This study utilized the corpus dataset developed by Yoon et al. (2022), derived from the NIKL (National Institute of Korean Language) learner corpus data. The NIKL learner corpus consists of essays composed by Korean learners, featuring error correction and annotations provided by their instructors. The dataset used in Yoon et al. (2022) contains 28,427 sentence pairs, but due to the search limit of ChatGPT 4 (25 searches per 3 hours; 125 searches per day) at the moment of investigation, the present study randomly selected 400 sentences from the NIKL learner corpus to create a parallel corpus dataset for ChatGPT 3.5 and ChatGPT 4 respectively.

Although attempts were made to include all 400 sentences in the GEC evaluation based on the comparison of ChatGPT and the baseline model, 67 sentences were excluded because the KoBART model that the present study selected from Yoon et al. (2022) failed to generate corrected sentences for these sentences. As a result, 333 sentences were included in the analysis comparing the GEC capability of ChatGPT and the baseline KoBART model. On the other hand, all 400 sentences were analyzed in the human evaluation comparing ChatGPT 3.5 and ChatGPT 4.

3.2 Baseline GEC System

One of the goals of the present study was to examine whether ChatGPT 3.5 and ChatGPT 4 can serve as reliable tools for Korean GEC tasks. As the baseline, the present study selected a KoBART model from Yoon et al. (2022), which demonstrated enhanced GEC performance relative to the commercial GEC system called Hanspell. KoBART is a Korean GEC model based on BART, which has been effective in generational tasks (Katsumata & Komachi, 2020). Yoon et al. (2022) loaded the pre-trained weights from KoBART and finetuned them using their own parallel corpus dataset derived from the NIKL learner corpus.

3.3 ChatGPT System

The present study also compared the performance of ChatGPT 3.5-Turbo and ChatGPT+ 4 (hereafter, ChatGPT 3.5 and ChatGPT 4). OpenAI released ChatGPT 3.5 on October 21, 2022 and is available to the public for free with no search limit. Unlike ChatGPT 3.5, ChatGPT 4, released on March 14, 2023, is available only for paid subscribers and permits 25 searches per 3 hours. Due to the search limit of ChatGPT 4, the present study conducted analyses using 400 sentences randomly selected from the parallel corpus dataset. To perform the GEC task on ChaptGPT, the prompt (1) was used. Following the suggestions of Fang et al. (2023), a prompt was created using the CoT method with special tags indicating input and output sentences to avoid generating unnecessary information in responses.

(1) ChatGPT Prompt for the GEC Task

> You are a Korean grammatical error correction tool that can identify and correct grammatical errors in a Korean text.
>
> Please identify and correct any grammatical errors in the multiple Korean sentences below indicated by <input> ERROR </input> tag.
>
> You need to comprehend the sentence as a whole before identifying and correcting any errors step by step while keeping the original sentence structure unchanged as much as possible.
>
> Remember to format your corrected output results with the tag <output> Your Corrected Version </output>.

3.4 Evaluation Methods

The GEC performance was evaluated with the M^2 scores, which measure Precision, Recall, and $F_{0.5}$ scores (Dahlmeier & Ng, 2012). *Precision* assesses correction accuracy, while *Recall* measures error identification and correction. $F_{0.5}$ score is the weighted harmonic mean of precision and recall. The $F_{0.5}$ score gives more weight to precision, making it suitable when the focus is on reducing false positives (incorrect corrections) while maintaining reasonable recall. Table 1 presents the formulas for calculating precision, recall, and $F_{0.5}$ scores.

Measure	Formula
Precision	(Number of True Positives) / (Number of True Positives + Number of False Positives)
Recall	(Number of True Positives) / (Number of True Positives + Number of False Negatives)
$F_{0.5}$	((1 + 0.5^2) * Precision * Recall) / (0.5^2 * Precision + Recall)

*True Positives: Number of correctly corrected errors
*False Positives: Number of incorrectly corrected errors
*False Negatives: Number of errors that should have been corrected but were missed

Table 1. Formulas for Calculating M^2 Measures

The GEC performance was also evaluated with GLEU (Napoles et al., 2016). The GLEU score evaluates the GEC using n-gram overlap with a set of reference sentences. In addition, human evaluation was conducted by one of the authors who is a linguist and a native Korean speaker, on the sentences corrected by ChatGPT 3.5 and ChatGPT 4 to investigate (1) whether ChatGPT 4 has a better ability to deal with multiple grammar errors than the previous version and (2) whether ChatGPT 4 outperforms the previous version in all type of grammar error or only in (a) specific type(s). In the human evaluation, grammar errors were categorized as the four types (spelling, particle, conjugation, and expression). Examples of each grammar type are presented in Table 2. *Spelling* errors refer to orthographic-level mistakes, typically involving typos. *Particle* errors are syntactic-level errors, including incorrect usage of case markers. On the other hand, *conjugation* errors encompass morphosyntactic information (e.g., tense inflection) and semantics-syntax interface (e.g., clause connectives). *Expression* errors pertain to semantic-level errors including the inappropriate use of words, idioms, and fixed expressions. In the human evaluation, the number of each error type was counted manually and analyzed using the R software.

Type	Correct Sentence	Incorrect Sentence
Spelling	학교 Hakkyo ‘school’	*핵교 Haekkyo
Particle	그가 뛴다 keu.ka ttwin-ta he.nom run.decl ‘he runs.’	*그를 뛴다 keu-lul ttwin-ta He.acc run.decl
Conjugation	지을 때 chi-ul ttae build-RELtime ‘when building’	*짓을 때 chis-ul ttae
Exp ression	값에 추가 kaps-ey cwuka price-ACC add ‘(we) add to the price.’	*값에 증가 kaps-ey ceungka price-ACC increase

Table 2. Types of Grammatical Errors in Human Evaluation

4 Results

4.1 Analysis based on M_2 scores and GLEU

The performance comparison of Grammatical Error Correction (GEC) among ChatGPT 3.5, ChatGPT 4, and KoBART is displayed in Table 3.

	GLEU	M^2		
		Precision	Recall	$F_{0.5}$
Baseline	0.41	0.33	0.35	0.33
GPT 3.5	**0.45**	**0.40**	0.31	**0.38**
GPT 4	**0.50**	**0.44**	**0.42**	**0.43**

Table 3. Comparison of KoBART and ChatGPT

These results are based on M^2 and GLEU scores. In Precision, both ChatGPT versions outscored the baseline model, meaning they are more adept at accurately correcting the grammatical errors that they have identified. However, for the Recall value, only the most recent version (ChatGPT 4) showed better results than the baseline model. This suggests that the error detection capability of ChatGPT 3.5 is on par with the baseline model, rather than superior. The higher precision and comparatively lower recall of ChatGPT 3.5 imply that it avoids making superfluous corrections but may overlook some existing errors in sentences. The $F_{0.5}$ score, which weighs precision more heavily than recall, was higher for both versions of ChatGPT than the baseline model. This shows that both versions of ChatGPT are effective at preserving grammatical accuracy, avoiding needless changes, and making necessary corrections when required. Moreover, both ChatGPT models registered higher GLEU values. This indicates that their error corrections more closely matched the reference sentences than the baseline model in this study.

In breif, the analysis shows that ChatGPT 4 surpasses both its predecessor and the baseline model across all aspects of M^2 scoring and GLEU values. This suggests that ChatGPT 4 has superior capabilities in identifying and appropriately correcting grammatical errors compared to ChatGPT 3.5.

4.2 Comparisons between ChatGPT 3.5 and ChatGPT 4 in Human Evaluation

For grammatical error corrections made by ChatGPT 3.5 and ChatGPT 4, two comparisons were conducted in human evaluations. First, we examined the extent to which the number of grammar errors committed in each sentence modulates the accuracy of GEC in GPT 3.5 and GPT 4. The first comparison focused on confirming whether GPT 4 is a better measure of GEC than GPT 3.5, regardless of the number of grammar errors. The accuracy of GEC was the dependent variable and coded as “1” if all the grammatical errors in a sentence were corrected and “0” if one or more grammatical errors were uncorrected. The binary coding was implemented to assess the probability of GPT generating entirely grammatical sentences, which could serve as exemplary sentences for L2 learners aiming to enhance their grammatical accuracy. The GEC accuracy measured in percentage was 31.05% (SD = 0.46) in GPT 3.5 and 58.16% (SD = 0.49) in GPT 4. A binomial logistics regression analysis was conducted using the *glmer*() function in R for the comparison, given that the dependent variable are binary (1 or 0). A model included GPT Version (2 levels: Version 3.5 vs. Version 4) and the number of errors (continuous variable) as fixed effects, and an interaction between GPT Version and the number of errors, as well as a by-item random intercept. Table 4 displays the model output.

	Estimate	SE	z value
(Intercept)	-1.32	0.20	-6.75***
GPT4	1.83	0.24	7.56***
NoE	-1.14	0.22	-5.28***
GPT4:NoE	0.50	0.23	2.20*

Notes: * p < 0.05, *** p < 0.001; NoE: Number of errors in a sentence

Table 4. Model Output for Human Evaluation

The model yielded a main effect of GPT, indicating that GPT 4 is significantly more accurate in GEC than GPT 3.5 (Estimate = 1.83, SE = 0.24, p < 0.001). Also, the model revealed a main effect of the number of errors, which means that increased grammatical errors in each sentence leads to decreased GEC accuracy (Estimate = -1.14, SE = 0.22, p < 0.001). Furthermore, an interaction of GPT and the number of grammatical errors was found (Estimate = 0.50, SE = 0.23, p < 0.05). The higher the number of grammar errors in a sentence, the more accurate GPT-4 proves to be in GEC tasks, as opposed to GPT-3.5.

The second comparison examines the extent to which the GEC accuracy is modulated by types of grammatical errors in GPT 3.5 and GPT 4. For the second comparison analysis, sentences containing a single type of error were selected out of 800 sentences (400 sentences from each version), and a total of 340 sentences were included in the analysis as a result. The types of grammar errors analyzed are as follows: spelling (39 tokens), particle (92 tokens), conjugation (78 tokens), and expression (131 tokens). Table 5 displays the GEC accuracy for each type of grammar error in GPT 3.5 and 4 respectively.

	GPT 3.5 (SD)	GPT 4.0 (SD)
Spelling	85.00% (0.37)	84.21% (0.37)
Particle	64.44% (0.48)	89.36% (0.31)
Conjugation	55.56% (0.50)	64.29% (0.48)
Expression	20.31% (0.41)	52.24% (0.50)

Table 5. Correction Accuracy of GPT 3.5 and 4

A binomial logistics regression analysis was conducted using the *glmer*() function in R. The model included GPT Version (2 levels: GPT 3.5 vs. GPT 4) and Error Type (4 levels: spelling vs. particle vs. conjugation vs. expression) as fixed effects and an interaction between GPT Version and Error Type. The model also included a by-item random intercept.

Table 6 presents the model output. The model output of the second comparison analysis did not yield a main effect of GPT Version, indicating no difference in accuracy between GPT 3.5 and GPT 4 for sentences involving only one type of grammar mistakes. The model also revealed the main effects of Error Type in Conjugation, and Expression, demonstrating that spelling errors were more accurately corrected than conjugation errors (Estimate = -2.98, SE = 1.33, p < 0.05) and expression errors (Estimate = -.5,76, SE = 1.62, p < 0.001) across GPT 3.5 and GPT 4. The main effects of Error Type suggest that the orthographic level errors (spelling mistakes) are more likely to be detected by GPT than the semantic level errors (expression) and those involving an interface of syntax and semantics (conjugation). Moreover, the model also found an interaction of GPT 4 and Expression (Estimate = 2.77, SE = 1.37, p < 0.05). The interaction implies that GPT 4 outperforms GPT 3.5 in dealing with errors committed at the semantic level.

	Estimate	SE	z value
(Intercept)	3.25	1.20	2.70**
GPT4	-0.01	1.14	-0.01
Par	-1.91	1.19	-1.61
Conj	-2.98	1.33	-2.24*
Exp	-5.76	1.62	-3.56***
GPT4:Par	2.68	1.44	1.85
GPT4:Conj	0.89	1.33	0.67
GPT4:Exp	2.77	1.37	2.02*

Notes: * p < 0.05, *** p < 0.001; Par: particle, Conj, Conjugation, Exp: Expression

Table 6. Model Output for Human Evaluation

5 Discussion

The present study sought to examine whether ChatGPT can serve as a more reliable Korean GEC instrument by using a KoBART model as the baseline. The comparisons revealed that ChatGPT 4 outperformed both ChatGPT 3.5 and the baseline model in all measures of M^2 scoring (Precision, Recall and $F_{0.5}$) and the GLEU value.

This study found that in the Korean GEC task, GEC tasks ChatGPT 4 was successful at both detecting grammatical errors within sentences and providing corrections corresponding to those errors. This contrasts with the results of the previous studies on English GEC tasks, in which ChatGPT failed to surpass the baseline GEC models in the measures associated with precisions (e.g., Fang et al., 2023; Wu et al., 2023). Another notable finding from the current study was that ChatGPT 3.5 performed the least successfully in identifying grammatical errors in this study's Korean GEC task although ChatGPT 3.5 and 4 excelled the most in the recall value in the previous studies. This result indicates that compared to ChatGPT 4, ChatGPT 3.5 may not provide comprehensive corrections that learners need in their feedback as it may neglect some grammatical errors in sentences.

Overall, ChatGPT 4 showcased more reliable performance in both error detection and correction in the Korean GEC task, when compared to its performance in English GEC tasks evaluated in previous studies. Regarding the discrepancy in the results, two potential reasons can be put forward. The first is the typological difference between Korean and English. Fang et al. (2023) have demonstrated that ChatGPT may operate less successfully than the baseline models in English GEC tasks on grammatical errors involving long-distance dependencies such as subject-verb agreement and coreference. In Korean, the number of grammatical features corresponding to long-distance dependencies is limited, which could contribute to the relatively high precision that ChatGPT exhibited in the Korean GEC task. Second, the dissimilarity from the findings of the previous studies can also stem from the relatively unstable performance of the KoBART model selected for the present study as the baseline in the present study. The baseline KoBART model in the present study failed to generate responses for 67 sentences, with accounts for roughly 17% of the total selected for the analysis. In future research, it would be necessary to select a more reliable pre-trained model as the baseline if availabe.

A human evaluation comparing the error correction quality of ChatGPT 3.5 and ChatGPT 4 showed the superiority of the latest version in handling multiple grammatical errors and correcting semantic-level errors. This evaluation initially explored how the number of errors impacts the performance of ChatGPT in Korean GEC tasks. The findings imply a decline in overall correction accuracy as the number of errors increased in both versions of ChatGPT. Nonetheless, the interaction between the GPT version and the number of grammatical errors suggests that ChatGPT 4 can maintain accurate grammatical corrections more consistently, even with multiple errors in a sentence. This interaction emphasizes the tendency of ChatGPT 3.5 to overlook some grammatical errors, as shown by its relatively lower recall score in the model comparison analysis. Although ChatGPT 4 has an advantage over the previous version in handling multiple grammatical errors, it should be noted that its overall correction accuracy is still far from ideal as it hovers around 60%. This means that L2 learners may not entirely rely on ChatGPT for grammatical corrections but also need to consult human instructors for detailed feedback.

An additional analysis based on the human evaluation found that ChatGPT has better capabilities for dealing with spelling and particle errors than conjugation and expression errors. It shows that ChatGPT may have more difficulty handling semantic-level errors than orthographic and syntactic errors that can be computed with relative ease. Though grouped into a single category in the human evaluation, verb conjugation in Korean in fact denotes a variety of linguistic information in addition to morphosyntactic information (e.g., tense inflection). For instance, phrase connectives are often expressed with verb conjugations in Korean, and their correct usages involve a careful understanding of nuanced and subtle understanding of contexts, which cannot be computed in a straightforward manner as is the case in the subject-verb agreement. Since many types of verb conjugation in Korean embody the interface of syntax and semantics, it may present challenges for ChatGPT models to handle these errors properly as they need to consult the two separate domains (syntax and semantics) to correct a single error point. Future studies should consider implementing a detailed classification of verb conjugation to ascertain if ChaGPT can function as a reliable GEC tool for both purely morphosyntactic conjugation (e.g., tense inflection), and semantic-syntax interface conjugation (e.g., clause connectives).

Moreover, expression errors refer to the inappropriate use of words, idioms, or fixed expressions that undermines the overall naturalness of sentences. While relatively simple syntactic errors like particle errors can be fixed based on the local dependency by checking its validity based on the adjacent element, expression error correction requires a more global-level approach by taking the overall interpretation of a sentence into account. Furthermore, expression errors, in many cases, may not necessarily involve syntactic-level errors. Hence, they are likely to be overlooked in error corrections conducted by ChatGPT.

The additional analysis conducted on the human evaluation also revealed an interaction of the GPT version and the error type, showing that ChatGPT 4 is significantly better than ChatGPT 3.5 in correcting expression errors. The sentences in (2) illustrate an example of an expression error, which was corrected in ChatGPT 4, but not in ChatGPT 3.5.

(2) Expression error correction by GPT 4
a. * *I* *kes-un* ***haekyel-ha-l***
This thing-TOP resolve-do-FUT
pangpeb-uro *pol-su iss-da*
way-INS see-POT-DECL

b. *I* *kes-un* ***haekyelchaek**-uro*
This thing-TOP solution-INS
pol-su iss-da
see-POT-DECL
'This can be viewed a solution.'

In the incorrect sentence, the verb conjugation *haekyelha-l* solve-FUT 'to resolve' sounds incomplete, as it requires an object *mwunce-lul* problem-ACC 'problem' to create a more natural expression as in *mwunce-lul hakyelha-l pangpeb* problem-ACC resolve-FUT way 'a way to resolve a problem.' However, ChatGPT 4 takes one step further and replaces the ungrammatical part with a single lexical item that appropriately expresses the intended meaning (*haekyelchaek*) and even increases the formality.

Consistent with the findings from Coyne et al. (2023), ChatGPT in the present study was found to have a better capability for enhancing the fluency of sentence revision by providing a correction that is grammatical as well as semantically appropriate in contextual information. The advanced capability of ChatGPT-4 to correct semantic-level errors emphasizes its potential as a valuable resource for L2 learners in the future. It is expected to aid in the understanding of semantic nuances, an area often not successfully acquired by L2 learners even at advanced proficiency levels (Jiang, 2000, 2002, 2004). Given that the overall accuracy of error correction for ChatGPT 4 is approximately 60%—a rate that can decrease as the number of grammatical errors increases—it is anticipated that advanced to near-native level second language (L2) learners, who can construct sentences with minimal errors, will benefit the most from using the latest version of ChatGPT 4 for Korean GEC tasks by improving the naturalness of their written production via immediate feedback.

Another aspect to bear in mind when utilizing ChatGPT models for Korean GEC tasks is their tendency to bias toward formal language usage, often over-correcting casual grammatical expressions. Korean is a pragmatically rich language, where different types of connectives are used depending on the level of formality. For instance, the post-nominal connective *-wa/-kwa* 'with' can be used in both formal and informal contexts, whereas the use of *-rang* 'with' is limited to informal contexts. When ChatGPT conducts a GEC task in Korean, it replaces most informal usages of connectives with formal usages (e.g., *rang* 'with' → *-wa/-ka* 'with'). ChatGPT's inclination to formal language usage in Korean GEC tasks implies that it may be more suitable for L2 learners who aim to improve their written proficiency rather than oral proficiency.

6 Conclusion

The present study shows that ChatGPT 4 can function as a more reliable Korean GEC instrument than ChatGPT 3.5. Although the human evaluation revealed that ChatGPT 4 outperforms the previous version in managing the multiple grammatical errors within a sentence and correcting semantic-level mistakes, it is still unclear whether it can be applied to a wide range of L2 learners as a self-learning tool due to its relatively low accuracy rate. Based on the findings from the human evaluation, the present study suggests that ChatGPT 4 be utilized for advanced to near-native level L2 learners to enhance the semantic naturalness of sentences constructed with minimal errors. Furthermore, because currently available ChatGPT models favor formal over casual language usage, using ChatGPT in GEC tasks may be more beneficial for L2 learners who are invested in enhancing their L2 writing skills.

References

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do,

Q. V., Xu, Y., & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. http://arxiv.org/abs/2302.04023

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. https://commoncrawl.org/the-data/

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). *Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction*. http://arxiv.org/abs/2303.14342

Dahlmeier, D., & Ng, H. T. (2012). *Better Evaluation for Grammatical Error Correction*. http://groups.google.com/group/hoo-nlp/

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). *Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation*. http://arxiv.org/abs/2304.01746.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. http://arxiv.org/abs/2302.09210

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21. 10.1093/applin/21.1.47

Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. Studies in Second Language Acquisition. 24. 617 - 637. 10.1017/S0272263102004047.

Jiang, N. (2004). *6. Semantic transfer and development in adult L2 vocabulary acquisition* (pp. 101–126). https://doi.org/10.1075/lllt.10.09jia

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine*. http://arxiv.org/abs/2301.08745

Katsumata, S., & Komachi, M. (2020). *Stronger Baselines for Grammatical Error Correction Using a Pretrained Encoder-Decoder Model*. https://github.com/pytorch/fairseq

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners*. http://arxiv.org/abs/2205.11916

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2016). *GLEU Without Tuning*. http://arxiv.org/abs/1605.02592

Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *the Association for Computational Linguistics* (Vol. 2).

Ruder, S. (2022). NLP-Progress.

Wu, H., Wang, W., Lyu, M. R., Wan, Y., & Jiao, W. (2023). *ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark Keyphrase generation for social media posts View project Question Generation View project ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark*. https://chat.openai.com/chat

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). *Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization*. http://arxiv.org/abs/2302.08081

Yoon, S., Park, S., Kim, G., Cho, J., Park, K., Kim, G., Seo, M., & Oh, A. (2022). *Towards standardizing Korean Grammatical Error Correction: Datasets and Annotation*. http://arxiv.org/abs/2210.14389