Unified Automated Essay Scoring and Grammatical Error Correction

SeungWoo Song* **Junghun Yuk*†** **ChangSu Choi** **HanGyeol Yoo** **HyeonSeok Lim**
KyungTae Lim§ **Jungyeul Park‡**
Seoul National University of Science and Technology, South Korea
§Korea Advanced Institute of Science and Technology, South Korea
†Hanbat National University, South Korea ‡The University of British Columbia, Canada
{sswoo, choics2623, 21102372, gustjrantk}@seoultech.ac.kr
†20191780@hanbat.ac.kr §ktlim@kaist.ac.kr ‡jungyeul@mail.ubc.ca

Abstract

This study explores the integration of automated writing evaluation (AWE) and grammatical error correction (GEC) through multitask learning, demonstrating how combining these distinct tasks can enhance performance in both areas. By leveraging a shared learning framework, we show that models trained jointly on AWE and GEC outperform those trained on each task individually. To support this effort, we introduce a dataset specifically designed for multitask learning using AWE and GEC. Our experiments reveal significant synergies between tasks, leading to improvements in both writing assessment accuracy and error correction precision. This research represents a novel approach for optimizing language learning tools by unifying writing evaluation and correction tasks, offering insights into the potential of multitask learning in educational applications.

1 Introduction

Learner corpus applications such as automated writing evaluation (AWE) and grammatical error correction (GEC) systems offer promising solutions for reducing the workload of language instructors and simplifying the feedback process for learners. These systems can serve as indispensable tools by providing prompt scoring alongside corrective feedback, thus addressing the growing need for efficient and effective assessment practices for both instructors and students. AWE predicts continuous values for a holistic score or a set of trait scores, offering a comprehensive assessment of the writing quality. GEC automates the detection and correction of grammatical errors in writing. As the number of both first- and second-language learners continues to increase and the demand for timely feedback intensifies, AWE and GEC have gained increasing popularity in recent years (Jiang et al., 2023; Bryant et al., 2023, *inter alia*). At present, AWE is commonly approached through linear regression models, while GEC is often treated as a translation task, in which incorrect sentences are "translated" into their correct forms. AWE and GEC have traditionally been treated as separate tasks.

In this study, we aim to introduce a system that integrates AWE and GEC through prompting, designed to enhance the language learning process and promote better learning outcomes. Although various approaches have leveraged AWE or GEC systems to support language learners, our goal is to develop a more robust and cohesive method that transforms traditional language education using state-of-the-art techniques. Throughout this paper, we use the term *prompt* in two distinct contexts: (1) In the context of essay writing, a *prompt* refers to a specific question, topic, or statement serving as the starting point for an essay. (2) In the context of generative pre-trained transformers, a *prompt* refers to the input provided to the model to generate a response.

The integration of technology into language learning and teaching has become ubiquitous, with automatic essay scoring (AES) emerging as a widely adopted and effective method for assessing writing proficiency and providing instant, individualized feedback to learners. AES systems automate the process of assigning numerical grades or scores to essays that are typically written in the first language in educational settings. For example, English AES systems commonly rely on the ASAP dataset, which was initially introduced in the 2012 Kaggle competition.[1] This dataset has been extensively used in various AES systems. On the other hand, GEC focuses on automatically identifying and correcting grammatical errors in sentences. English GEC systems have leveraged the Cambridge English Write & Improve (W&I) corpus, which is

*Equally contributed authors.
Corresponding authors: KyungTae Lim and Jungyeul Park.

[1] https://www.kaggle.com/c/asap-aes

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4412–4426

manually annotated with CEFR proficiency levels (Yannakoudakis et al., 2018). This corpus was introduced at the Building Educational Applications (BEA) 2019 Shared Task: Grammatical Error Correction (BEA2019) (Bryant et al., 2019) and has since been widely used in GEC research.

Only a few studies have attempted to integrate pre-neural AES and GEC systems, with most focusing on the perceived feedback from language learners (Ranalli, 2018; O'Neill and Russell, 2019; Zhang, 2020; Ariyanto et al., 2021; Reynolds et al., 2021). While empirical studies have explored student perceptions, beliefs, and preferences regarding feedback, its direct influence on writing performance remains unclear (Truscott, 1999; Ferris, 1999; Chandler, 2003; Ferris, 2004, 2014). Therefore, this study focuses on integrating AES and GEC systems through prompting. For the first time, we combine neural AES and GEC systems, achieving results that approach state-of-the-art performance.

2 Enriching the corpus

Korean AWE dataset We used the Korean AWE dataset (Lim et al., 2023), which includes proficiency levels ranging from Levels 1 to 6 (equivalent to CEFR levels A1 to C2). The dataset also contains additional information such as students' native languages by nationality, gender, teacher-assigned overall scores, writing prompts, and complete writing samples, which have been manually segmented into sentences and essays. This dataset has been used in several studies on the applications of the Korean learner corpus (Sung and Shin, 2023a,b, 2024). While some writing prompts are specific to learners at certain proficiency levels (e.g., the prompt *My weekend* for Level 1), others are shared across multiple proficiency levels (e.g., *The day I remember the most* for Levels 3 and 5). The dataset includes more than 100 writing prompts, with 21 prompts assigned to multiple proficiency levels, representing 42.96%.

Enriching the AWE dataset with GEC AWE and GEC face challenges owing to their different basic processing units. AWE evaluates entire essays and assesses all aspects of a student's writing, whereas GEC processes the text on a sentence-by-sentence basis. To address this, a native speaker manually segmented the essays used in AWE into smaller sentence units as the corpus contains grammatical errors, making automatic sentence boundary detection unreliable. This yielded 48,937 sentences. We then conducted GEC annotation by correcting these segmented sentences. To ensure GEC accuracy, two native Korean speakers with backgrounds in applied linguistics were employed to verify the results. In line with the essay writing evaluation data guidelines, we ensured that the GEC annotations minimized changes to the author's intended content. Following manual corrections, we achieved an 82.6% agreement between the linguists. In cases of annotation conflicts, a post-processing step was implemented in which disagreements were resolved through discussion. Details regarding the annotation guidelines, annotators, and final dataset are provided in Appendix A. We excluded a small subset of the test dataset from the 4,011 processed essays, which included prompts that were not part of the training dataset. Because a fair evaluation requires an even distribution of proficiency levels, the prompt *The most memorable day of my life*, evenly distributed between Levels 3 (B1) and 5 (C1), was selected for the test dataset. The remaining dataset, consisting of 3,804 essay samples and 44,865 sentences, was used for training. The test dataset included 207 essays and 4,072 sentences. The detailed statistics of the training data and the distribution of grammatical errors can be found in Appendix B, E.

3 Experiments and Analysis

We employed the Bllossom model and fine-tuned it for our experiments, which we describe in Appendix C[2]. Additionally, we refer to our enriched dataset as K-UEED.

3.1 Experiments environment

For our experiments, we introduced two evaluation methods to assess the validity of the training dataset and proposed model: (1) an internal evaluation using only the K-UEED dataset, and (2) an external evaluation incorporating additional GEC datasets. To ensure reproducibility and enable a fair comparison with previous studies, we followed the evaluation methods proposed in previous studies. The evaluation metrics employed are presented in Table 1, as follows: (1) AWE: Lim et al. (2023) proposed methods for predicting proficiency levels using accuracy along with scoring based on the quadratic weighted kappa (QWK). (2) GEC: Yoon

[2] https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B

Model	AWE	QWK	GEC (GLEU)
Llama3-8B-MUL	47.34	36.46	49.06
Llama3-8B (5-shot)	16.42	4.86	37.01
GPT-4o (5-shot)	41.55	39.25	56.19
BLLOSSOM-AWE	36.72	46.18	-
BLLOSSOM-GEC	-	-	**50.76**
BLLOSSOM-MUL	46.38	46.86	49.92
BLLOSSOM-GA	**60.39**	**64.16**	50.13
BLLOSSOM-AG	41.06	58.60	50.01

Table 1: Experimental results using K-UEED data

et al. (2023) suggested the use of F0.5 scores and GLEU (Napoles et al., 2015) as evaluation metrics. The experiments were conducted using identical prompt formats to ensure that all models were evaluated under the same conditions. The instruction tuning for all LLMs was run over 10 epochs, with the tailored training sequences and hyperparameter settings provided in Appendix D.

3.2 Evaluation on K-UEED

Table 1 presents the experimental results obtained using the K-UEED dataset. Specifically, BLLOSSOM-AWE refers to a version of the Bllossom model based on Llama3-8B Dubey et al. (2024) without vocabulary expansion, trained exclusively on the AWE dataset from K-UEED to specialize in AWE. In contrast, BLLOSSOM-GEC was trained solely on the GEC dataset. Finally, the BLLOSSOM-(MUL,AG,GA) model is a multitask model trained using a combination of AWE and GEC. A key advantage of K-UEED is its capability to support diverse combinations of multitask training. During the training process, all essays were divided into sentence- and essay-level GEC tasks. The AG and GA methods involved merging the GEC and AWE datasets at the essay level into a single sample with 3,804 samples. In the AG method, the models were first trained to predict the AWE for an essay, followed by the GEC. The GA method followed the reverse order. The MUL method involved training by combining sentence-level GEC with essay-level data from the GA method. However, while these methods used essays as inputs during training (which did not affect AWE evaluation), the GEC evaluation required inputs at the sentence level; therefore, the dataset was split accordingly.

Overall comparison It can be observed from the results in Table 1 that GPT-4o achieved the highest GEC scores, followed by BLLOSSOM-GEC, which was 5.43 points lower. Notably, the Llama-8B (5-shot) model, which was not fine-tuned and was provided with only five examples, occasionally generated text outside the proficiency level, resulting in lower performance.

Impact of multitask training on GEC and AWE When comparing the results of BLLOSSOM-GEC with the various multitask-trained models (BLLOSSOM-*) in Table 1, the GEC models trained in a single-task format exhibited the highest performance. Incorporating the AWE dataset into GEC training did not yield positive effects. Conversely, in AWE, the models trained in a multitask format scored significantly higher, with increases ranging from 0.68 to 17.98 points on the QWK scale. Notably, the BLLOSSOM-GA model, which was trained to generate GEC results for a paragraph before predicting the AWE score, showed a substantial improvement in AWE performance. This suggests that the model's ability to reference its own GEC results first enhances the accuracy of the subsequent AWE score calculations. In contrast, the BLLOSSOM-AG model, which predicts AWE first and then generates GEC, did not show significant performance gains in the GEC task.

3.3 Evaluation on existing GEC dataset

What potential benefits can be achieved by combining the proposed K-UEED dataset with existing GEC data? Table 2 shows that the KAGAS GEC dataset consists of separate training and testing sets for L1 and L2 learners (Yoon et al., 2023). We evaluated the performance on the KAGAS test dataset by integrating it with the K-UEED dataset in various configurations.

Impact of dataset combination on GEC BLLOSSOM+KAGAS and BLLOSSOM-*+KAGAS models in Table 2 differ in their use of KAGAS data. The observations indicate that simultaneous training with L2 datasets from K-UEED and KAGAS L2 led to significant performance improvements across all cases. Notably, the BLLOSSOM-GEC+KAGAS model showed increases of 1.01 points in GLEU and 1.77 points in the F-score, suggesting that combining similar L2 datasets can yield positive results. In contrast, training with the L2 dataset from K-UEED combined with KAGAS L1 did not result in significant performance improvements.

	L2 (learner)				L1 (native)				Union (L1+L2)				Gen. time
	GLEU	M^2			GLEU	M^2			GLEU	M^2			
		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$	
Hanspell (Yoon et al., 2023)	30.36	29.45	5.33	15.46	57.08	81.93	47.36	71.50	-	-	-	-	189.69
KoBART (Yoon et al., 2023)	45.06	43.35	24.54	37.58	67.24	75.34	55.95	70.45	-	-	-	-	38.25
LLAMA3-8B-INST (5-SHOT)	33.23	26.33	39.01	24.45	39.73	32.65	29.54	31.98	35.97	28.85	22.59	33.91	63.24
LLAMA3-8B-INST+KAGAS	52.35	55.64	37.22	50.63	82.18	90.70	78.66	88.01	62.73	68.67	50.24	63.98	53.07
BLLOSSOM+KAGAS	53.64	58.93	38.49	53.22	82.29	92.13	78.37	89.01	63.42	72.08	51.05	66.59	59.7
BLLOSSOM-GEC+KAGAS	54.65	**60.33**	40.66	**55.01**	82.03	92.05	78.05	88.87	64.18	71.58	51.65	66.45	57.14
BLLOSSOM-MUL+KAGAS	54.21	59.30	39.78	54.00	81.56	90.92	77.40	87.85	63.40	70.94	50.90	65.76	51.06
BLLOSSOM-GA+KAGAS	**54.96**	60.12	**40.76**	54.90	82.35	92.21	78.64	89.13	**64.28**	**72.39**	**51.90**	**67.09**	55.92
BLLOSSOM-AG+KAGAS	54.44	59.72	40.01	54.37	**82.74**	**92.36**	**78.77**	**89.28**	63.96	71.88	51.29	66.54	55.12
BLLOSSOMV-GEC+KAGAS	49.00	53.46	34.22	48.05	77.81	84.81	74.41	82.51	59.41	67.97	46.56	62.24	52.41

Table 2: Experimental results using the Korean KAGAS GEC data proposed by Yoon et al.

Impact of AWE dataset on KAGAS-GEC The BLLOSSOM-GA+KAGAS model in Table 2 used the K-UEED dataset for simultaneous training on AWE and GEC, followed by fine-tuning with the KAGAS training data. This model exhibited the best performance in the Union training, which involved simultaneous training on L1 and L2, scoring 1.5 points higher than the BLLOSSOM+KAGAS model. These results suggest that multitask training of AWE and GEC can positively influence the GEC performance, depending on the training method.

Impact of pre-training on the GEC task The Bllossom model is fundamentally based on Llama3 and was subjected to additional pre-training with approximately 100 GB of Korean-English data. In our evaluation, the BLLOSSOM+KAGAS model outperformed the LLAMA3-8B-INST+KAGAS model in several metrics. Notably, the two models showed a significant score difference of 2.61, suggesting that the additional Korean pre-training had a highly positive impact on performance.

Impact of vocabulary extension on GEC The BLLOSSOMV model in Table 2, which is based on the Llama3-8B model, incorporated an additional 23,000 Korean words before undergoing pre-training with 100 GB of Korean-English data. When comparing the BLLOSSOM-KAGAS model, which was pre-trained without vocabulary expansion, to the BLLOSSOMV-KAGAS model, the inclusion of word extension showed a negative impact on the performance.

3.4 Analysis

The top three lines in Figure 1 display the actual GEC error distributions for K-UEED and KAGAS levels L1 and L2, respectively. In addition, the bottom four lines indicate the error prediction distributions of the proposed Bllossom-based model. Our analysis yielded two key observations. First, excluding the "noop" error type unique to K-UEED, the GEC error distributions for K-UEED and L2 were strikingly similar. Second, a strong alignment between the dataset and model error distributions correlated with improved GEC performance, which was particularly notable in the BLLOSSOM-GA and K-UEED dataset, as listed in Table 2. The additional findings are presented in Appendix E.

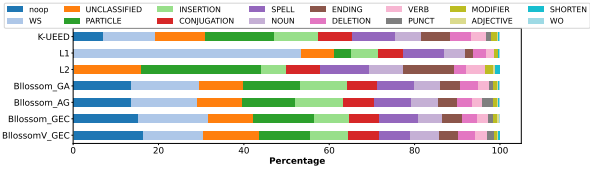

Figure 1: Distributions of GEC error types in K-UEED and KAGAS, and error predictions by the Bllossom model

Table 3 shows the distribution of grammatical error types (GEC) in relation to AWE scores within the K-UEED dataset. It illustrates how the occurrence of certain errors varies depending on the score. Interestingly, a distinction was observed between lower and higher AWE scores. As AWE scores increase, the proportion of the [noop, WS] type, where no corrections are made to the sentences, also increases. In contrast, the frequency of 'INSERTION' errors decreases as scores improve, showing a strong correlation between specific error types and the overall writing quality.

4 Conclusion

In conclusion, our work with K-UEED demonstrates the potential for integrating AWE and GEC into

Type	AWE Score										
	0	10	20	30	40	50	60	70	80	90	100
ADJECTIVE	1 (0.60%)	2 (0.76%)	4 (0.41%)	12 (0.55%)	24 (0.71%)	67 (0.67%)	103 (0.67%)	121 (0.61%)	117 (0.63%)	63 (0.45%)	3 (0.22%)
CONJUGATION	12 (7.19%)	21 (7.95%)	84 (8.62%)	196 (8.96%)	276 (8.17%)	819 (8.17%)	1226 (8.02%)	1369 (6.90%)	1197 (6.46%)	700 (5.04%)	56 (4.05%)
DELETION	7 (4.19%)	10 (3.79%)	36 (3.70%)	88 (4.02%)	122 (3.61%)	442 (4.41%)	692 (4.53%)	918 (4.63%)	781 (4.21%)	545 (3.93%)	43 (3.11%)
ENDING	13 (7.78%)	11 (4.17%)	66 (6.78%)	102 (4.66%)	244 (7.22%)	648 (6.46%)	1010 (6.61%)	1240 (6.25%)	1153 (6.22%)	803 (5.79%)	61 (4.41%)
INSERTION	35 (20.96%)	22 (8.33%)	201 (20.64%)	389 (17.79%)	415 (12.28%)	1217 (12.14%)	1931 (12.63%)	2167 (10.93%)	1610 (8.68%)	1052 (7.58%)	97 (7.01%)
NOUN	16 (9.58%)	18 (6.82%)	46 (4.72%)	160 (7.32%)	229 (6.78%)	714 (7.12%)	1050 (6.87%)	1376 (6.94%)	1231 (6.64%)	796 (5.74%)	80 (5.78%)
PARTICLE	16 (9.58%)	35 (13.26%)	126 (12.94%)	277 (12.67%)	551 (16.30%)	1561 (15.57%)	2284 (14.94%)	2745 (13.84%)	2615 (14.10%)	1618 (11.66%)	144 (10.40%)
PUNCT	4 (2.40%)	0 (0.00%)	12 (1.23%)	31 (1.42%)	41 (1.21%)	93 (0.93%)	96 (0.63%)	189 (0.95%)	133 (0.72%)	192 (1.38%)	24 (1.73%)
SPELL	19 (11.38%)	30 (11.36%)	82 (8.42%)	219 (10.01%)	363 (10.74%)	1077 (10.74%)	1562 (10.22%)	2091 (10.54%)	1885 (10.17%)	1238 (8.92%)	91 (6.58%)
UNCLASSIFIED	28 (16.77%)	55 (20.83%)	149 (15.30%)	340 (15.55%)	485 (14.35%)	1298 (12.95%)	1938 (12.68%)	2408 (12.14%)	1964 (10.59%)	1368 (9.86%)	135 (9.75%)
VERB	4 (2.40%)	6 (2.27%)	30 (3.08%)	59 (2.70%)	103 (3.05%)	297 (2.96%)	448 (2.93%)	608 (3.07%)	630 (3.40%)	396 (2.85%)	31 (2.24%)
WS	6 (3.59%)	25 (9.47%)	77 (7.91%)	206 (9.42%)	309 (9.14%)	1087 (10.84%)	1722 (11.26%)	2420 (12.20%)	2410 (13.00%)	2077 (14.97%)	202 (14.60%)
noop	6 (3.59%)	27 (10.23%)	53 (5.44%)	89 (4.07%)	188 (5.56%)	597 (5.96%)	1057 (6.91%)	1923 (9.70%)	2564 (13.83%)	2867 (20.66%)	399 (28.83%)
MODIFIER	0 (0.00%)	1 (0.38%)	6 (0.62%)	17 (0.78%)	18 (0.53%)	85 (0.85%)	128 (0.84%)	184 (0.93%)	158 (0.85%)	94 (0.68%)	9 (0.65%)
SHORTEN	0 (0.00%)	1 (0.38%)	2 (0.21%)	0 (0.00%)	6 (0.18%)	16 (0.16%)	30 (0.20%)	55 (0.28%)	76 (0.41%)	62 (0.45%)	9 (0.65%)
WO	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (0.09%)	6 (0.18%)	6 (0.06%)	11 (0.07%)	16 (0.08%)	17 (0.09%)	5 (0.04%)	0 (0.00%)
TOTAL	166	262	970	2175	3356	9957	15185	19709	18424	13813	1381

Table 3: Types of GEC errors in relation to the AWE scores from the K-UEED dataset.

a unified framework, offering significant advancements in language assessment and correction tasks. By leveraging multitask learning with generative language models, we observed improvements in both writing evaluation and error correction, highlighting the synergy between these seemingly distinct tasks. The results underscore the importance of developing comprehensive datasets such as the K-UEED, which enable more efficient and accurate language processing models. As language learning continues to evolve, the integration of AWE and GEC within a single system opens new avenues for enhancing automated educational tools, contributing to more personalized and effective feedback for learners.

Limitations

While this study provides a comprehensive analysis and presents promising results, the integration of AWE and GEC through multi-task learning, although effective in this case, may present challenges when applied to other languages or less-structured learner corpora. Future research could explore how these models generalize to different linguistic datasets and educational environments, further expanding the applicability of the findings.

Acknowledgments

We would like to thank the reviewer for their insightful feedback throughout the study, particularly regarding the error types in GEC from a syntactic perspective. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (No.RS-2024-00456709, A Development of Self-Evolving Deepfake Detection Technology to Prevent the Socially Malicious Use of Generative AI) and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City awarded to KyungTae Lim.

References

Meilisa Sindy Astika Ariyanto, Nur Mukminatien, and Sintha Tresnadewi. 2021. College Students' Perceptions of an Automated Writing Evaluation as a Supplementary Feedback Tool in a Writing Class. *Jurnal Ilmu Pendidikan*, 27(1):41–51.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

Jean Chandler. 2003. The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3):267–296.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dana R. Ferris. 1999. The case for grammar correction in L2 writing classes: A response to truscott (1996). *Journal of Second Language Writing*, 8(1):1–11.

Dana R. Ferris. 2004. The "Grammar Correction" Debate in L2 Writing: Where are we, and where do we go from here? (and what do we do in the meantime ...?). *Journal of Second Language Writing*, 13(1):49–62.

Dana R. Ferris. 2014. Responding to student writing: Teachers' philosophies and practices. *Assessing Writing*, 19(1):6–23.

Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of The Tenth International Conference on Learning Representations*, pages 1–13, Virtual. ICLR.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving Domain Generalization for Prompt-Aware Essay Scoring via Disentangled Representation Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

KyungTae Lim, Jayoung Song, and Jungyeul Park. 2023. Neural automated writing evaluation for Korean L2 writing. *Natural Language Engineering*, 29(5):1341–1363.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground Truth for Grammatical Error Correction Metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Ruth O'Neill and Alex M. T. Russell. 2019. Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology*, 35(1):42–56.

Jim Ranalli. 2018. Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning*, 31(7):653–674.

Barry Lee Reynolds, Chian-Wen Kao, and Yun-yin Huang. 2021. Investigating the Effects of Perceived Feedback Source on Second Language Writing Performance: A Quasi-Experimental Study. *The Asia-Pacific Education Researcher*, 30(6):585–595.

Hakyung Sung and Gyu-Ho Shin. 2023a. Diversifying language models for lesser-studied languages and language-usage contexts: A case of second language Korean. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11461–11473, Singapore. Association for Computational Linguistics.

Hakyung Sung and Gyu-Ho Shin. 2023b. Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 72–82, Toronto, Canada. Association for Computational Linguistics.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a Dependency Treebank for Second Language Learners of Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758, Torino, Italia. ELRA and ICCL.

John Truscott. 1999. The case for "The Case Against Grammar Correction in L2 Writing Classes": A response to Ferris. *Journal of Second Language Writing*, 8(2):111–122.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean Grammatical Error Correction: Datasets and Annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.

Zhe (Victor) Zhang. 2020. Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43(January):100439.

Supplementary Materials

A Data Production Process

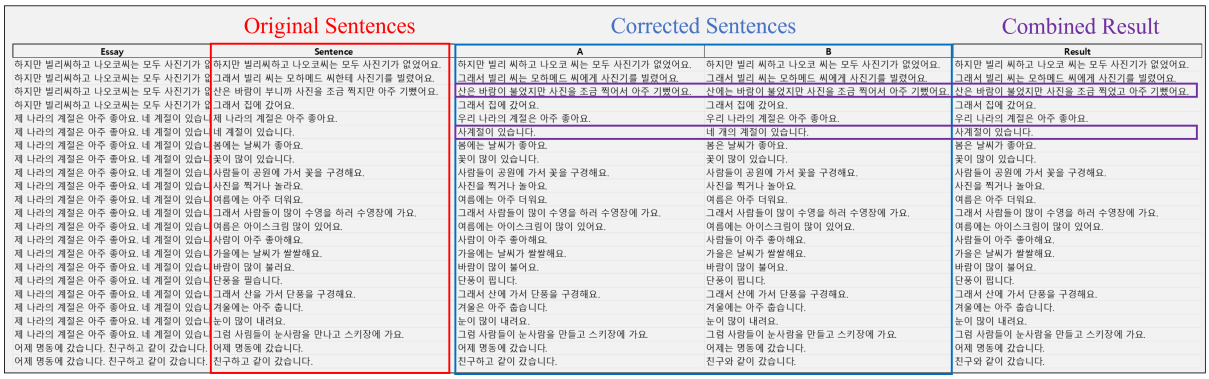

Figure 2: A part of the K-UEED data. 'Original Sentences' refers to sentences containing grammatical errors, 'Corrected Sentences' are sentences in which grammatical errors have been corrected by two annotators, and 'Combined Result' represents the final outcome determined through discussion between the two annotators.

To produce high-quality K-UEED data, we employed two students currently studying linguistics who are native Korean speakers. In order to exclude various Korean dialects, we hired students who reside in Seoul, an area known for using the standard language. The employment details for the two annotators are as follows.

- Major: Linguistics
- Time required: 50 hours over two weeks
- Wage: 9,860 KRW per hour, confirming South Korea's minimum hourly wage
- Age: 23 years old

To generate data optimized for model training, we provided two annotators with clear data creation guidelines. The annotators were instructed to make corrections while minimizing changes to the meaning of the sentences and preserving the author's intent. Additionally, efforts were made to minimize the difference in sentence length before and after corrections to retain the characteristics of the original sentences as much as possible. In cases where there were discrepancies in the corrections made by the annotators, a discussion was held to adjust the text towards a more natural direction. The guidelines for building the K-UEED dataset were informed by the GEC data construction procedures from Ai Hub and the National Institute of the Korean Language. The final format of the K-UEED dataset is shown in Figure 2.

B Details of the K-UEED Dataset

Level	Title	Number of titles
1	Weekend stories, Birthday party, Plans for Next Week, ...	16
2	Seasons and weather in my country, My future plans, ...	29
3	The most memorable day of my life, My hobbies, ...	28
4	Environmental pollution, Movie review ...	27
5	The most memorable day of my life, ...	19
6	Death penalty, Education, ... system	11

Table 4: The titles and levels of the K-UEED dataset.

Table 4 lists the titles and their associated proficiency levels within the K-UEED dataset. The titles span various topics and levels, ranging from simple everyday scenarios (e.g., "Weekend stories" at Level 1)

to more complex, abstract topics (e.g., “Death penalty” at Level 6), providing a diverse set of writing prompts for learners across different proficiency levels.

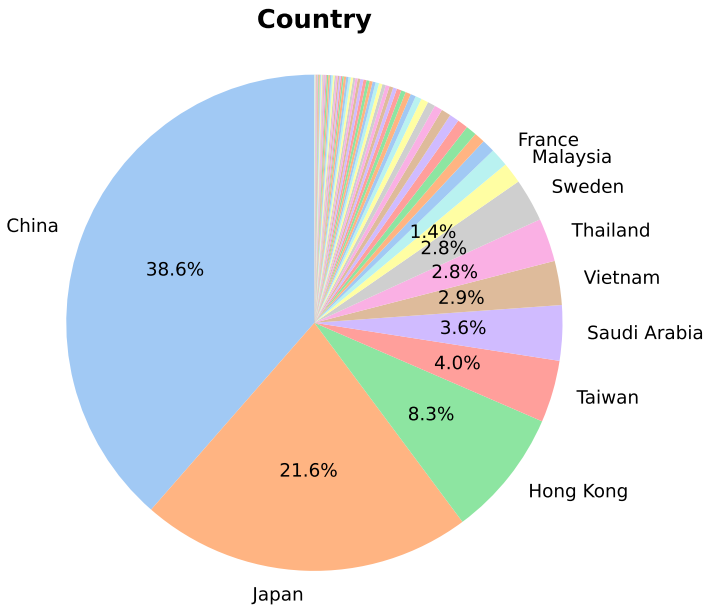

Figure 3: Distribution of learners’ native languages within the K-UEED dataset.

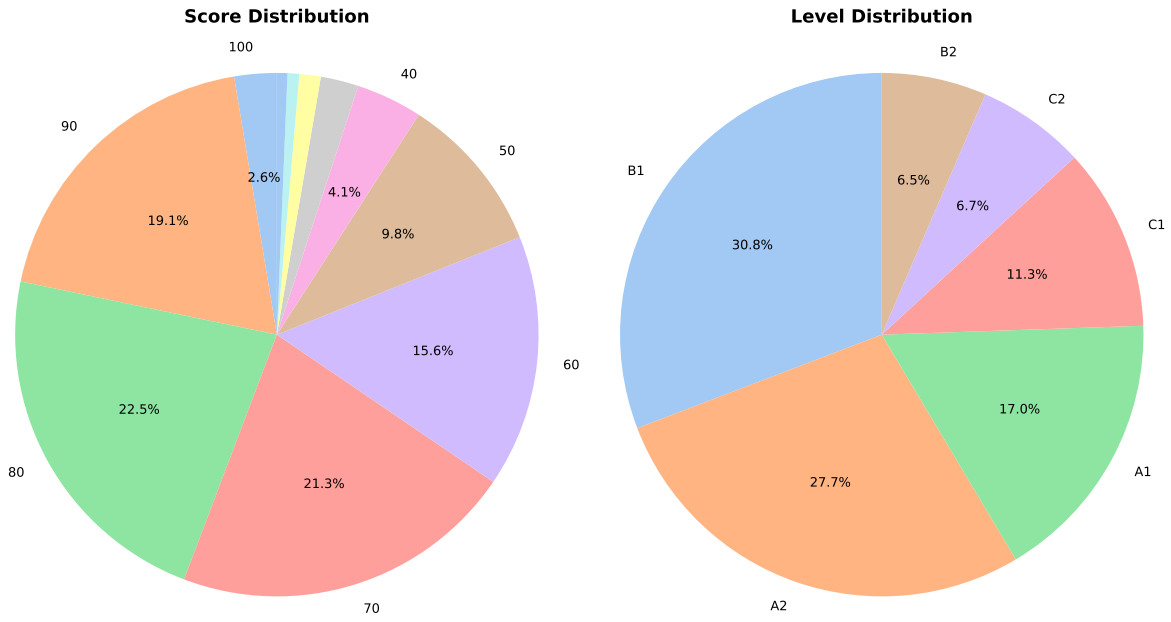

Figure 4: Distribution of scores and levels for 4,011 contexts within the complete K-UEED dataset.

Figure 3 shows the distribution of learners’ native languages. Of the 71 countries represented in the dataset, more than 75% of the participants come from Asian countries, offering a detailed breakdown of the learners’ linguistic backgrounds.

Figure 4 presents the score and level distributions for 4,011 contexts in the K-UEED dataset. Scores range from 0 to 100, with an average score of 69.63. Figure 5 illustrates the word distribution in paragraphs and sentences, indicating an average word count of 79 and 147 for paragraphs in the train and test sets, respectively, and 6.7 and 7.5 words for sentences.

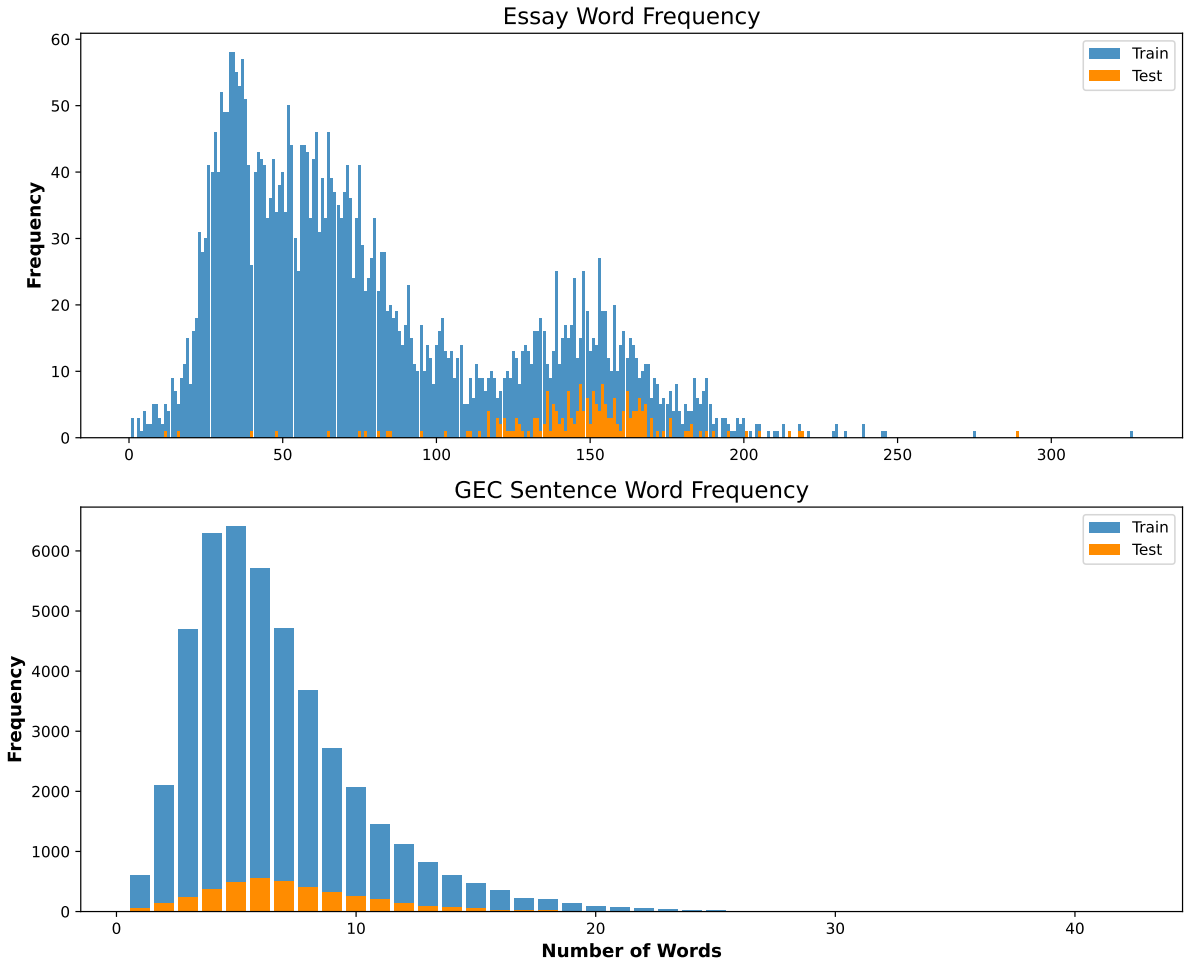

Figure 5: Word distribution in paragraphs and sentences of the K-UEED dataset.

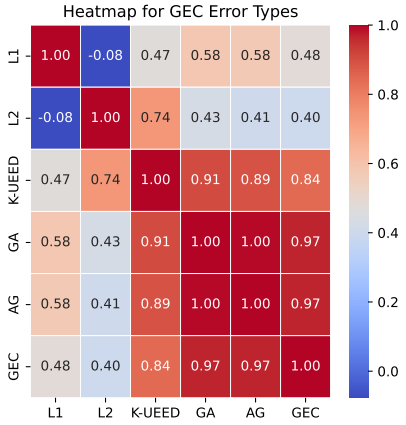

Figure 6: Heatmap of the correlation coefficients between the actual error types in the L1, L2, and K-UEED datasets and those generated by the BLLOSSOM-(GEC, GA, AG) models during the GEC task.

Figure 6 shows a heatmap of error types across datasets L1, L2, and K-UEED, as well as between the error types generated by the model and those present in the K-UEED dataset during the GEC task. Notably, the K-UEED and L2 datasets exhibit highly similar error distributions, while the correlation between L1 and L2 error types is the lowest. Interestingly, the BLLOSSOM-(GEC, GA, AG) models that achieved

outstanding performance on the K-UEED dataset also excelled on the L2 dataset, highlighting their robust capabilities across similar error distributions.

C Details of Bllossom

In this section, we introduce the Bllossom model[3] utilized in this paper. Tasks such as AWE and GEC are significantly influenced by linguistic grammatical and lexical features. Therefore, the configuration of the vocabulary in the language model used, along with the type of data employed for its pre-training, can alter the model's understanding of the task. Unfortunately, recent Large Language Models (LLMs) primarily utilize publicly available multilingual LLMs, which are predominantly focused on English or other Latin-based languages. Consequently, in models for languages such as Korean, Chinese, and Japanese, there is a relative deficiency in understanding vocabulary and grammar. For instance, while Llama3[4] is a Multilingual LLM (MLLM) trained in multiple languages, an internal review revealed that Latin-based tokens represent 113,966 (88.86%) of the total, compared to a considerably smaller proportion of 2,281 (1.78%) for tokens exclusive to Korean. This imbalance frequently results in responses to Korean queries being in English or produced in a code-switching format, as shown in Table 5.

Prompt
Tokenized Input: '맛', '있는', '식', '사를', ' 하', '셨', '습', '니까', '?'
Code-switching Output: 저는 안 먹었어요. Have you had a good meal? ...

Table 5: Example of the code-switching

Vocabulary Enhancement for LLMs To enhance the Korean capabilities of the multilingual language model, BllossomV underwent additional pre-training after expanding its vocabulary. The dataset, denoted as $\mathcal{D}$, comprised 144,782 tokens, which includes 16,782 Korean vocabulary items that do not overlap with the 128,256 tokens of the existing Llama3. The newly added words were randomly initialized. Consequently, to facilitate the learning of representations for the newly added words and Korean grammatical knowledge, additional pre-training was conducted using Causal Language Modeling (CLM). The fundamental concept of CLM, as described in Equation (1), involves receiving a sequence of input tokens $x_{<i} = (x_0, x_1, \ldots, x_{i-1})$ and predicting the next token x_i that will appear. The model obtains a loss value by taking the negative log-likelihood of the probability of the predicted token and aims to minimize this loss.

$$L_{pt}(\theta) = \mathbb{E}_{x \sim \mathscr{D}_{PT}} \left\{ -\sum_i log P(x_i | x_{<i}; \theta, \mathcal{D}) \right\} \tag{1}$$

In Equation (1), $L_{pt}(\theta)$ represents the loss function of the language model for the pre-training dataset $\mathscr{D}_{PT}$. Here, $\mathscr{D}_{PT}$ denotes the pre-training dataset, which utilizes data from the Korean Wikipedia. θ refers to the parameters of the model, and (x_i) represents the token to be predicted. This loss function considers the loss associated with how well each token in the pre-training dataset is predicted.

D Training Details and Hyperparameters

Hyperparameters Similar to Bllossom, we employed Low-Rank Adaptation (LORA) (Hu et al., 2022) for instruction tuning. During the training, the hyperparameters of LORA were set in accordance with the recommendations proposed by Bllossom, as detailed in Table 7.

Training details Table 6 presents the instruction tuning training time for all models on an A100 GPU. The GEC consists of 40K samples, while the AWE consists of 4K. Therefore, AWE requires significantly less time. For GA and AG, it takes more time compared to AWE, as they need to output not only the scores for 4k paragraphs but also the results of the GEC. However, this process is still considerably faster

[3] MLP-KTLim/llama-3-Korean-Bllossom-8B

[4] https://ai.meta.com/blog/meta-llama-3

than that of GEC. Nevertheless, the fact that GA performs better indicates that the K-UEED dataset, which allows for the simultaneous training of AWE and GEC, is highly effective.

GPU	Training Type	Duration
A100x1	AWE	1.1h
A100x1	GEC	11.5h
A100x1	GA and AG	2.4h
A100x1	MUL	13.1h
A100x1	L1	3.5h
A100x1	L2	5.5h

Table 6: Instruction tuning training time for all models on an A100 GPU.

	value
Dropout	0.05
Learning rate	5e-5
Optimizer	AdamW_bnb_8bit
Epoch for IT	10
Batch size	2
Low-rank size	64
lora_alpha	128
lora_trainable	q,v,k,o,gate,down,up_proj
LoRA layer,	q, k, v
Random Seed	42

Table 7: Applied hyperparameter settings for instruction tuning.

E Detail of Grammatical Error Types.

Type	K-UEED(GEC)	Llama3_8b_MUL	Bllossom_GEC	Bllossom_MUL	Bllossom_GA	Bllossom_AG
ADJECTIVE	38 (0.45%)	34 (0.45%)	30 (0.42%)	30 (0.43%)	33 (0.46%)	26 (0.35%)
CONJUGATION	679 (7.98%)	499 (6.57%)	491 (6.92%)	448 (6.40%)	496 (6.84%)	525 (7.06%)
DELETION	403 (4.73%)	287 (3.78%)	259 (3.65%)	250 (3.57%)	310 (4.27%)	272 (3.66%)
ENDING	584 (6.86%)	327 (4.31%)	320 (4.51%)	306 (4.37%)	331 (4.56%)	330 (4.44%)
INSERTION	861 (10.11%)	982 (12.93%)	594 (8.37%)	581 (8.30%)	814 (11.22%)	843 (11.34%)
MODIFIER	102 (1.20%)	56 (0.74%)	58 (0.82%)	64 (0.91%)	51 (0.70%)	53 (0.71%)
NOUN	522 (6.13%)	461 (6.07%)	411 (5.79%)	386 (5.52%)	445 (6.14%)	459 (6.17%)
PARTICLE	1382 (16.23%)	1024 (13.48%)	1000 (14.09%)	968 (13.84%)	963 (13.28%)	922 (12.40%)
PUNCT	110 (1.29%)	118 (1.55%)	90 (1.27%)	71 (1.01%)	72 (0.99%)	192 (2.58%)
SHORTEN	23 (0.27%)	14 (0.18%)	16 (0.23%)	16 (0.23%)	18 (0.25%)	22 (0.30%)
SPELL	864 (10.15%)	614 (8.09%)	652 (9.19%)	630 (9.01%)	634 (8.74%)	656 (8.82%)
noop	606 (7.12%)	1060 (13.96%)	1077 (15.17%)	1193 (17.05%)	986 (13.60%)	1014 (13.64%)
UNCLASSIFIED	1002 (11.77%)	817 (10.76%)	764 (10.76%)	748 (10.69%)	754 (10.40%)	792 (10.65%)
VERB	296 (3.48%)	189 (2.49%)	175 (2.47%)	173 (2.47%)	193 (2.66%)	181 (2.43%)
WO	7 (0.08%)	1 (0.01%)	2 (0.03%)	0 (0.00%)	1 (0.01%)	4 (0.05%)
WS	1034 (12.15%)	1111 (14.63%)	1159 (16.33%)	1132 (16.18%)	1151 (15.87%)	1143 (15.38%)
TOTAL	8513	7594	7098	6996	7252	7434

Table 8: This table represents the distribution of GEC error types generated by the model compared to the K-UEED (GEC) dataset.

Table 8 provides a comparative analysis of error type distributions between the K-UEED dataset and various models, such as the Llama3-8b-based models and the Bllossom models. Notably, significant differences are observed in error types like 'noop' and 'WS,' where the BLLOSSOM-MUL model generated a higher proportion of these errors compared to the original K-UEED GEC dataset.

Table 9 presents the distribution of GEC error types in the L1 dataset and compares it with various models, including Llama3-8b, BLLOSSOM-GEC, and multi-task models. Significant differences are observed in error types like 'noop' and 'WS,' where the models generate a higher number of these errors compared to the L1 dataset.

Table 10 shows the distribution of GEC error types for the L2 dataset and compares it with multiple models, including Llama3-8b and BLLOSSOM-GEC. It highlights notable differences in categories like

Type	L1	Llama3_8b	Bllossom_GEC	Bllossom_MUL	Bllossom_GA	Bllossom_AG	BllossomV_MUL
ADJECTIVE	8 (0.18%)	6 (0.15%)	5 (0.13%)	3 (0.07%)	5 (0.13%)	6 (0.15%)	7 (0.16%)
CONJUGATION	263 (5.89%)	196 (4.88%)	204 (5.13%)	220 (5.48%)	210 (5.25%)	211 (5.29%)	262 (6.04%)
DELETION	138 (3.09%)	89 (2.22%)	79 (1.99%)	98 (2.44%)	88 (2.20%)	88 (2.20%)	163 (3.76%)
ENDING	88 (1.97%)	37 (0.92%)	41 (1.03%)	40 (1.00%)	37 (0.93%)	36 (0.90%)	46 (1.06%)
INSERTION	291 (6.52%)	217 (5.40%)	221 (5.56%)	218 (5.43%)	219 (5.48%)	216 (5.41%)	279 (6.43%)
MODIFIER	40 (0.90%)	30 (0.75%)	27 (0.68%)	25 (0.62%)	28 (0.70%)	28 (0.70%)	29 (0.67%)
NOUN	220 (4.93%)	88 (2.19%)	77 (1.94%)	92 (2.29%)	81 (2.03%)	80 (2.00%)	189 (4.36%)
PARTICLE	170 (3.81%)	91 (2.27%)	96 (2.41%)	98 (2.44%)	90 (2.25%)	84 (2.10%)	132 (3.04%)
PUNCT	0 (0.00%)	4 (0.10%)	4 (0.10%)	4 (0.10%)	3 (0.08%)	3 (0.08%)	4 (0.09%)
SHORTEN	9 (0.20%)	5 (0.12%)	3 (0.08%)	6 (0.15%)	4 (0.10%)	4 (0.10%)	6 (0.14%)
SPELL	420 (9.41%)	352 (8.77%)	365 (9.18%)	356 (8.87%)	360 (9.00%)	360 (9.02%)	340 (7.84%)
noop	0 (0.00%)	171 (4.26%)	214 (5.38%)	220 (5.48%)	216 (5.40%)	218 (5.46%)	203 (4.68%)
UNCLASSIFIED	346 (7.75%)	270 (6.72%)	255 (6.41%)	263 (6.55%)	261 (6.53%)	255 (6.39%)	394 (9.08%)
VERB	76 (1.70%)	45 (1.12%)	42 (1.06%)	43 (1.07%)	45 (1.13%)	42 (1.05%)	61 (1.41%)
WO	6 (0.13%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
WS	2388 (53.51%)	2414 (60.12%)	2344 (58.94%)	2327 (57.99%)	2352 (58.81%)	2361 (59.14%)	2222 (51.23%)
TOTAL	4463	4015	3977	4013	3999	3992	4337

Table 9: Distribution of GEC error types generated by the model compared to the L1 dataset.

Type	L2	Llama3_8b	Bllossom_GEC	Bllossom_MUL	Bllossom_GA	Bllossom_AG	BllossomV_MUL
ADJECTIVE	57 (0.64%)	30 (0.42%)	32 (0.45%)	24 (0.34%)	29 (0.41%)	27 (0.38%)	30 (0.42%)
CONJUGATION	706 (7.97%)	542 (7.65%)	540 (7.64%)	511 (7.19%)	501 (7.10%)	502 (7.06%)	549 (7.71%)
DELETION	255 (2.88%)	249 (3.51%)	232 (3.28%)	192 (2.70%)	219 (3.10%)	214 (3.01%)	281 (3.95%)
ENDING	1059 (11.95%)	800 (11.29%)	760 (10.75%)	720 (10.14%)	755 (10.70%)	733 (10.31%)	659 (9.26%)
INSERTION	518 (5.85%)	291 (4.11%)	263 (3.72%)	296 (4.17%)	298 (4.22%)	292 (4.11%)	304 (4.27%)
MODIFIER	157 (1.77%)	77 (1.09%)	71 (1.00%)	72 (1.01%)	72 (1.02%)	78 (1.10%)	78 (1.10%)
NOUN	706 (7.97%)	462 (6.52%)	440 (6.22%)	449 (6.32%)	424 (6.01%)	450 (6.33%)	555 (7.79%)
PARTICLE	2499 (28.20%)	2095 (29.57%)	2029 (28.70%)	2068 (29.11%)	2100 (29.76%)	2103 (29.58%)	1844 (25.90%)
PUNCT	0 (0.00%)	2 (0.03%)	7 (0.10%)	3 (0.04%)	5 (0.07%)	4 (0.06%)	7 (0.10%)
SHORTEN	89 (1.00%)	106 (1.50%)	102 (1.44%)	88 (1.24%)	105 (1.49%)	104 (1.46%)	83 (1.17%)
SPELL	1014 (11.44%)	725 (10.23%)	834 (11.80%)	831 (11.70%)	809 (11.46%)	807 (11.35%)	701 (9.85%)
noop	0 (0.00%)	315 (4.45%)	380 (5.37%)	435 (6.12%)	364 (5.16%)	416 (5.85%)	479 (6.73%)
UNCLASSIFIED	1397 (15.76%)	1114 (15.73%)	1073 (15.18%)	1113 (15.67%)	1102 (15.62%)	1103 (15.51%)	1261 (17.71%)
VERB	394 (4.45%)	232 (3.27%)	260 (3.68%)	247 (3.48%)	237 (3.36%)	234 (3.29%)	237 (3.33%)
WO	0 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	1 (0.01%)	0 (0.00%)	2 (0.03%)
WS	11 (0.12%)	44 (0.62%)	46 (0.65%)	55 (0.77%)	36 (0.51%)	43 (0.60%)	50 (0.70%)
TOTAL	8862	7084	7070	7104	7057	7110	7120

Table 10: Distribution of GEC error types generated by the model compared to the L2 dataset.

'DELETION,' 'INSERTION,' and 'noop,' where the models exhibit varied behavior in error detection compared to the L2 dataset.

		ADJ	CONJ	DEL	END	INS	MOD	NOUN	PART	PUNCT	SHORT	SPELL	noop	UNK	VERB	WO	WS
Llama3-8b	Pre	0.14	6.24	3.43	0.92	6.64	1.10	2.41	2.62	0.0	0.11	11.81	0.0	7.74	1.28	0.0	64.10
	Rec	50.00	69.20	72.46	29.55	69.07	77.50	31.36	43.53	100.0	33.33	81.90	100.0	67.34	47.37	0.0	92.63
	$F_{0.5}$	0.18	7.62	4.24	1.15	8.11	1.37	2.96	3.22	0.0	0.13	14.24	0.0	9.40	1.59	0.0	68.31
Bllossom_GEC	Pre	0.15	6.21	3.25	0.94	6.78	1.01	2.35	2.94	0.0	0.11	12.43	0.0	7.94	1.37	0.0	64.36
	Rec	50.00	67.68	67.39	29.55	69.42	70.00	30.00	48.24	100.0	33.33	85.00	100.0	68.21	50.00	0.0	91.04
	$F_{0.5}$	0.18	7.59	4.02	1.16	8.27	1.26	2.88	3.62	0.0	0.14	14.99	0.0	9.64	1.71	0.0	68.36
Bllossom_MUL	Pre	0.11	6.38	3.55	0.90	6.66	0.97	2.45	2.76	0.0	0.18	12.12	0.0	7.63	1.23	0.0	63.82
	Rec	37.50	69.58	73.91	28.41	68.04	67.50	31.36	45.29	100.0	55.56	82.62	100.0	65.32	44.74	0.0	90.62
	$F_{0.5}$	0.14	7.80	4.39	1.12	8.12	1.21	3.01	3.40	0.0	0.23	14.62	0.0	9.27	1.53	0.0	67.83
Bllossom_GA	Pre	0.15	6.49	3.64	0.98	6.90	1.05	2.54	2.71	0.0	0.15	12.22	0.0	7.94	1.34	0.0	64.70
	Rec	50.00	70.72	75.36	30.68	70.45	72.50	32.27	44.12	100.0	44.44	83.10	100.0	67.92	48.68	0.0	91.71
	$F_{0.5}$	0.18	7.94	4.49	1.22	8.42	1.31	3.11	3.33	0.0	0.18	14.73	0.0	9.65	1.67	0.0	68.75
Bllossom_AG	Pre	0.18	6.45	3.66	0.98	6.82	1.12	2.46	2.80	0.0	0.11	12.27	0.0	7.97	1.27	0.0	64.79
	Rec	62.50	70.34	76.09	30.68	69.76	77.50	31.36	45.88	100.0	33.33	83.57	100.0	68.21	46.05	0.0	91.79
	$F_{0.5}$	0.23	7.89	4.52	1.21	8.33	1.39	3.02	3.45	0.0	0.14	14.80	0.0	9.68	1.58	0.0	68.84
BllossomV_MUL	Pre	0.14	5.64	3.44	0.87	6.28	0.97	2.50	2.83	0.0	0.11	11.32	0.0	7.39	1.19	0.0	59.44
	Rec	50.00	63.50	73.91	28.41	66.32	70.00	33.18	48.24	100.0	33.33	80.00	100.0	65.32	44.74	0.0	86.89
	$F_{0.5}$	0.17	6.90	4.25	1.08	7.67	1.21	3.07	3.49	0.0	0.13	13.67	0.0	8.98	1.47	0.0	63.45

Table 11: M2 Scores for each grammatical error type in the L1 training data

Table 11 and Table 12 represent the M2 Scores for each grammatical error type in the L1 and L2 datasets, respectively, across different models. These tables provide insight into how effectively each model performs in identifying specific error types such as 'ADJECTIVE,' 'CONJUGATION,' and 'SPELL.' The results show a consistent improvement in model performance, particularly in categories like 'noop' and

		ADJ	CONJ	DEL	END	INS	MOD	NOUN	PART	PUNCT	SHORT	SPELL	noop	UNK	VERB	WO	WS
Llama3-8b	Pre	0.34	3.17	1.18	6.27	1.55	0.89	4.38	23.72	0.0	0.96	12.39	0.0	9.65	1.87	0.0	0.02
	Rec	28.07	21.42	21.96	28.61	14.29	26.75	29.60	50.02	100.0	50.56	60.36	100.0	33.86	22.34	100.0	9.09
	$F_{0.5}$	0.43	3.82	1.46	7.43	1.89	1.11	5.28	26.51	0.0	1.20	14.73	0.0	11.26	2.28	0.0	0.03
Bllossom_GEC	Pre	0.30	3.97	1.33	7.14	1.71	0.97	4.76	24.83	0.0	1.02	14.26	0.0	10.79	2.15	0.0	0.04
	Rec	24.56	26.52	24.31	32.29	15.44	28.66	31.73	51.82	100.0	52.81	68.93	100.0	37.51	25.38	100.0	18.18
	$F_{0.5}$	0.38	4.79	1.64	8.46	2.08	1.21	5.73	27.72	0.0	1.27	16.95	0.0	12.58	2.63	0.0	0.05
Bllossom_MUL	Pre	0.22	3.76	1.12	6.88	1.54	1.02	4.85	24.65	0.0	0.96	13.84	0.0	10.94	2.18	0.0	0.02
	Rec	17.54	24.96	20.39	30.88	13.90	29.94	32.15	51.14	100.0	49.44	66.77	100.0	38.01	25.63	100.0	9.09
	$F_{0.5}$	0.27	4.53	1.39	8.15	1.88	1.27	5.84	27.50	0.0	1.19	16.45	0.0	12.76	2.67	0.0	0.03
Bllossom_GA	Pre	0.34	3.80	1.36	7.05	1.78	0.92	4.72	24.98	0.0	1.06	13.88	0.0	11.01	2.13	0.0	0.04
	Rec	28.07	25.53	25.10	32.11	16.22	27.39	31.59	52.38	100.0	55.06	67.75	100.0	38.58	25.38	100.0	18.18
	$F_{0.5}$	0.43	4.58	1.68	8.36	2.16	1.15	5.69	27.89	0.0	1.31	16.51	0.0	12.84	2.61	0.0	0.05
Bllossom_AG	Pre	0.35	3.54	1.25	6.92	1.73	1.02	4.53	25.24	0.0	0.96	13.88	0.0	10.63	2.20	0.0	0.04
	Rec	28.07	23.55	22.75	31.16	15.64	29.94	30.03	52.62	100.0	49.44	66.86	100.0	36.86	25.89	100.0	18.18
	$F_{0.5}$	0.43	4.26	1.54	8.20	2.10	1.26	5.46	28.18	0.0	1.19	16.50	0.0	12.39	2.69	0.0	0.05
BllossomV_MUL	Pre	0.31	3.20	0.92	5.68	1.40	1.06	3.90	21.95	0.0	0.91	12.24	0.0	9.56	1.59	0.0	0.04
	Rec	24.56	20.85	16.47	24.93	12.36	30.57	25.35	44.06	100.0	46.07	57.50	100.0	32.36	18.27	100.0	18.18
	$F_{0.5}$	0.39	3.85	1.14	6.71	1.70	1.31	4.69	24.39	0.0	1.13	14.53	0.0	11.13	1.94	0.0	0.06

Table 12: M2 Scores for each grammatical error type in the L2 training data

Model	AWE	QWK	GEC (GLEU)
BLLOSSOM3.1-AWE	45.89	53.14	-
BLLOSSOM3.1-GEC	-	-	59.09
BLLOSSOM3.1-MUL	**86.47**	54.93	59.26
BLLOSSOM3.1-GA	71.01	**60.69**	**59.71**
BLLOSSOM3.1-AG	46.38	59.22	59.09

Table 13: Experimental results on the K-UEED dataset using Bllossom3.1.

'WS,' which significantly contribute to overall error correction.

F K-UEED in Bllossom3.1

The Bllossom 3.1 model, pre-trained in Korean based on Llama 3.1[5], also showed positive synergy from the interaction between K-UEED's GEC and AWE tasks when trained simultaneously, similar to the Bllossom model. In Table 13, it was observed that the performance of the AWE task using the MUL method improved compared to Table 1, and in Table14, the GEC, MUL, GA, and AG methods showed similar performance to Table 2.

G Detail of Instruction Format

Figure 7 displays the six different paraphrased versions of instructions used during training. For each data sample, one of these versions was randomly selected to ensure diversity in the instructions. However, during evaluation, a single instruction was used consistently.

Figure 8 provides an example of the data structure used for training each model. The instructions for each model were randomly selected from the list in Figure 7, and the dataset for the MUL model was created by combining datasets from both the AWE and GA models.

[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

	L2 (learner)				L1 (native)				Union (L1+L2)				Gen. time
	GLEU	M^2			GLEU	M^2			GLEU	M^2			
		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$	
BLLOSSOM3.1-GEC+KAGAS	52.02	60.26	37.71	53.82	81.89	92.57	78.04	89.25	62.76	73.31	50.42	67.21	50.2
BLLOSSOM3.1-MUL+KAGAS	52.79	**61.48**	38.51	**54.93**	81.52	91.91	77.91	88.72	62.23	73.27	49.65	66.91	50.47
BLLOSSOM3.1-GA+KAGAS	53.02	60.68	38.62	54.46	**82.86**	92.26	78.97	89.25	63.06	**73.35**	50.8	**67.37**	56.02
BLLOSSOM3.1-AG+KAGAS	**53.29**	60.62	**39.26**	54.67	82.69	**93.18**	**79.13**	**89.98**	**63.19**	73.25	**50.48**	67.19	54.29

Table 14: Experimental results of Bllossom3.1 on the Korean KAGAS GEC data proposed by Yoon et al.

Instruction list

GEC

1. "아래의 한국어 글에 대해 문법을 교정한 문장만 출력해줘."
(For the Korean text below, only output grammar-corrected sentences.)

2. "다음 한국어 글쓰기에 대해 문법을 교정한 문장만 출력해줘."
(For the following Korean text, please print only grammatically corrected sentences.)

3. "아래의 한국어 글에 대해 맞춤법을 검사해서 고친 문장만 출력해줘."
(For the Korean text below, please only output sentences that have been spell-checked and corrected.)

4. "다음 글에 대해 맞춤법을 검사해서 고친 문장만 출력해줘."
(Only spell-checked and corrected sentences for the following text.)

5. "한국어 글에 대해 맞춤법을 검사해서 고친 문장만 출력해줘."
(Please print only spell-checked and corrected sentences for the Korean text.)

6. "아래 글의 맞춤법을 교정해서 교정한 문장만 출력해줘."
(Spell-check the following article and print only the corrected sentences.)

AWE

1. "아래의 {topic}에 대한 글의 종합적인 수준과 점수를 매겨줘."
(Please rate the overall quality and score of the article below on {Topic}.)

2. "{topic}에 대해 작성된 아래 글의 전체적인 수준과 점수를 평가해줘."
(Please rate the overall quality and score of the article below on {Topic}.)

3. "아래의 {topic}에 관한 글을 평가하고, 전반적인 내용의 질에 따른 수준과 점수를 매겨줘."
(Please rate the article below on {Topic} and give it a level and score based on the overall quality of the content.)

4. "{topic}에 대한 아래 글을 보고, 글 전체에 대한 수준과 점수를 평가해줘."
(Please look at the article below on {Topic} and rate the overall quality and score of the article.)

5. "이 {topic}에 대한 글을 검토하고, 종합적인 수준과 점수를 통해 평가해줘."
(Please review this article on {Topic} and rate it with an overall level and score.)

6. "다음에 제시된 {topic}에 대한 글을 읽고, 종합적인 품질에 대한 수준과 점수를 매겨줘."
(Read the following article on {Topic} and rate it on a scale and score for overall quality.)

GA / AG

1. "아래의 {Topic}에 대한 글의 종합적인 수준과 점수를 매기고 문법을 교정해줘."
(Rate and score the overall quality of the writing below on {Topic} and correct its grammar.)

2. "{Topic}에 대해 작성된 아래 글의 전체적인 수준과 점수를 평가하고 글의 문법을 교정해줘."
(Rate the overall quality and score of the following text on {Topic} and correct its grammar.)

3. "아래의 {Topic}에 관한 글을 평가하고, 전반적인 내용의 질에 따른 수준과 점수를 매기고 글의 맞춤법을 교정해줘."
(Please evaluate the article below on {Topic}, give it a level and score based on the overall quality of the content, and correct the spelling.)

4. "{Topic}에 대한 아래 글을 보고, 글 전체에 대한 수준과 점수를 평가하고 글의 문법을 교정해줘."
(Please look at the article below on {Topic}, rate and score the overall quality of the article, and correct the grammar.)

5. "이 {Topic}에 대한 글을 검토하고, 종합적인 수준과 점수를 통해 평가하고 맞춤법을 교정해줘."
(Please review this article on {Topic}, rate it with an overall level and score, and correct spelling.)

6. "다음에 제시된 {Topic}에 대한 글을 읽고, 종합적인 품질에 대한 수준과 점수를 매기고 글의 문법을 교정해줘."
(Read the following article on {Topic}, rate and score it for overall quality, and correct my grammar.)

Figure 7: Instruction selection candidates for each model used in instruction tuning.

Data Sample

[Topic]
우리 가족
(My Family)

[Original Essay]
'제 어머니는 **의사읍니다**. **의사기** 때문에 **매일 매일** 너무 **바빱니다**. 하지만 지금 어머니 세계여행를 **시작습니다**. 저는 세계 **여행었으면** 좋겠습니다. 제 어머니는 영어를 **줄** 압니다. 제 어머니는 얼굴이 **의할머니처럼습니다**.'

[Corrected Essay]
'제 어머니는 **의사입니다**. **의사이기** 때문에 **매일매일** 너무 **바쁩니다**. 하지만 지금 어머니**는** 세계여행을 **시작하셨습니다**. 저는 세계 **여행을 했으면** 좋겠습니다. 제 어머니는 영어를 **잘** 압니다. 제 어머니는 얼굴이 **외할머니처럼 생겼습니다**.'

[English Ver.]
'My mother **is a doctor**. **Because she is a doctor**, she is very **busy** every day. But now my mother **has started** traveling the world. And I **want to travel** the world. My mother knows English **very well**. My motheAr has a face **like my maternal grandmother**.'

[Level, Score]
2급, 20

Data Structure

System Prompt

[Prompt]
당신은 문법 오류를 수정하고 자동 글쓰기 평가를 제공하는 유용한 AI 어시스턴트입니다. 모든 답변은 정확하고 명료해야 합니다.
You are a helpful AI assistant tasked with correcting grammar errors and providing automated writing assessments. All responses must be accurate and clear.

AWE

[Instruction]
아래의 우리 가족에 대한 글의 종합적인 수준과 점수를 매겨줘.
(Please rate the overall quality and score of the article below on my family.)

[Input]
제 어머니는 의사읍니다. 의사기 때문에 매일 매일 너무 바빱니다. 하지만…

[Output]
Level: 2
Score: 20

AG

[Instruction]
아래의 우리 가족에 대한 글의 종합적인 수준과 점수를 매기고 문법을 교정해줘.
(Rate and score the overall quality of the writing below on my family and correct its grammar.)

[Input]
제 어머니는 **의사읍니다**. **의사기** 때문에 **매일 매일** 너무 **바빱니다**. 하지만…

[Output]
Level: 2
Score: 20
제 어머니는 **의사입니다**. **의사이기** 때문에 매일매일 너무 바쁩니다. 하지만 …
(My mother **is a doctor**. Because **she is a doctor**, she is very **busy every day**. But …)

MUL

GEC

[Instruction]
아래의 한국어 글에 대해 문법을 교정한 문장만 출력해줘.
(For the Korean text below, only output grammar-corrected sentences.)

[Input]
제 어머니는 **의사읍니다**.

[Output]
제 어머니는 **의사입니다**.
(My mother **is a doctor**.)

GA

[Instruction]
아래의 우리 가족에 대한 글의 종합적인 수준과 점수를 매기고 문법을 교정해줘.
(Rate and score the overall quality of the writing below on my family and correct its grammar.)

[Input]
제 어머니는 **의사읍니다**. 의사기 때문에 매일 매일 너무 바빱니다. 하지만…

[Output]
제 어머니는 **의사입니다. 의사이기** 때문에 **매일매일** 너무 **바쁩니다**. 하지만 …
(My mother **is a doctor.** Because **she is a doctor**, she is very **busy every day**. But …)
Level: 2
Score: 20

Figure 8: Training data samples used for instruction tuning and the data structure input to each model.