IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Do Typological Similarities Matter? Cross-Lingual GEC from a Low-Resource Korean Model

Jiseop Kim

Lynbrook High School, 1280 Johnson Ave, San Jose, CA, 95129, USA; uniquestella1122@gmail.com
Mentor: Sangyoon Bae

ABSTRACT: As the global demand for multilingual education grows, supporting English Language Learners (ELLs) through tools like grammatical error correction (GEC) has become increasingly important. However, most neural GEC models rely on large, annotated datasets that are only available for high-resource languages, leaving underrepresented linguistic communities at a disadvantage. This study investigates whether a GEC model trained on Korean—a morphologically rich, low-resource language—can generalize to other languages through cross-lingual transfer. We fine-tuned a pretrained sequence-to-sequence model on Korean GEC data. We evaluated its performance on grammatically erroneous sentences in four typologically diverse target languages: Japanese, Russian, German, and Thai. Evaluation metrics included the BLEU score and semantic similarity to assess both surface-level accuracy and meaning preservation. The model performed best on Japanese, a typologically similar language with shared subject-object-verb (SOV) word order and agglutinative morphology, supporting the Typological Similarity Hypothesis. Surprisingly, Russian ranked second, despite structural differences, aligning with the Morphological Richness Hypothesis and suggesting that deeper morphosyntactic complexity and flexible syntax may enable transfer. German and Thai showed more limited transfer, particularly at the surface level. These findings demonstrate that morphologically informed models trained on low-resource languages like Korean can serve as effective cross-lingual scaffolds. This approach offers a scalable and inclusive strategy for developing GEC tools in under-resourced contexts, with potential applications in education, translation, and language learning.

KEYWORDS: Behavioral and Social Sciences, Grammatical Error Correction, Computational Linguistics, English Language Learners.

Introduction

As technological development accelerates globally, the importance of multilingual education has grown rapidly. In the United States, English Language Learner (ELL) students—those whose native language is not English—have become the fastest-growing student population. The number of ELL students in grades 7–12 increased by approximately 70% between 1992 and 2002. By 2021, ELLs made up 10.4% of all public-school students nationwide.[7] In California alone, the number reached over 1 million, accounting for 18.6% of public-school enrollment. However, data from the National Center for Education Statistics (NCES) shows that the four-year high school graduation rate for ELLs in 2020–21 was only 71%, compared to 86% for non-ELLs—a 15 percentage point gap.5 California's own records indicate similar disparities, with ELL graduation rates at 72.4%, compared to 88.7% for non-ELLs and 89.2% for reclassified students.[6] These statistics underscore significant educational inequities and highlight the urgency of supporting students learning English as a second language.

One of the key challenges faced by ELL students is difficulty with English grammar, which is directly linked to lower academic achievement.[1,2] Providing corrective feedback on grammar errors can significantly improve writing skills for English as a Second Language (ESL) learners. Therefore, grammatical error correction (GEC)—the task of detecting and correcting grammatical mistakes in writing—has become a vital component in supporting ELLs' academic progress.

Neural network–based GEC systems, in particular, require large-scale annotated corpora for effective training. However, most existing datasets are available only for high-resource languages like English, Chinese, or German. To address class imbalance or data sparsity, prior research has often downsampled data by error type or learner group, which can distort real-world error distributions and limit model robustness. Training GEC systems in low-resource language settings is particularly difficult due to the lack of sufficient labeled data.[3] This has spurred interest in multilingual approaches using pretrained models that can transfer learning across languages.

Moreover, developing separate GEC models for each language is costly and inefficient, requiring expert annotation and extensive computational resources. Such language-specific systems do not scale well.[3] In contrast, multitask models capable of handling multiple languages simultaneously are more scalable. Similarly, pretraining GEC models on data-rich languages and transferring them to low-resource languages can significantly reduce development cost and time, proving the practical value of cross-lingual transfer.[4]

Therefore, if a GEC model trained on one language can generalize well to others with similar grammatical structures, it could reduce dependence on language-specific datasets and

DOI: 10.36838/IJHSR812.53

improve access to educational tools in underrepresented linguistic communities.

This study explores whether a GEC model trained on Korean—a morphologically rich, low-resource language—can effectively generalize to typologically similar languages. Specifically, we evaluate its performance on Japanese, which shares subject-object-verb (SOV) word order and agglutinative morphology, as well as on other agglutinative languages such as Turkish, Uzbek, Kazakh, Finnish, and Hungarian. For comparison, we also include typologically dissimilar fusional or inflectional languages such as Spanish, French, German, and Russian. If the Korean-trained GEC model performs well on structurally similar languages, it would reduce the need for large annotated datasets in each target language and support more inclusive, scalable language education technologies.

Our main research question is:

Can a GEC model fine-tuned on Korean grammatical error correction effectively transfer to typologically similar languages?

We also examine its performance on unrelated languages.

To guide our inquiry, we propose two hypotheses: A GEC model trained on Korean will perform better on target languages that either (a) are typologically similar to Korean (e.g., share agglutinative morphology and/or SOV word order), or (b) exhibit high morphological richness, even if their surface syntax differs.

This study contributes to cross-lingual natural language processing (NLP) by demonstrating how a single GEC model trained on a low-resource but structurally rich language can be extended to other languages. It offers a cost-efficient, inclusive, and scalable approach to multilingual education and error correction tools.

Methods

We fine-tuned a pretrained sequence-to-sequence model on a Korean GEC dataset, then evaluated it on parallel test sets in other languages with controlled grammatical errors.

To evaluate cross-lingual generalization, we tested the Korean GEC model on grammatically erroneous sentences in four target languages: Japanese, Russian, German, and Thai.

These four languages were selected to represent a spectrum of typological similarity to Korean. Japanese was chosen as the most typologically similar language, as it shares both SOV word order and agglutinative morphology with Korean. Russian was selected for its complex morphological system and partial word order flexibility, offering moderate similarity despite being a fusional language. German, also a fusional language, was included due to its verb-second (V2) word order and rich inflection, making it structurally less similar to Korean. Finally, Thai was chosen as the most distant language typologically—it has subject-verb-object (SVO) word order and an isolating morphological structure, with no inflection or affixation. Because we had only 3108 sentences of the German dataset, we used 3108 sentences to train every four languages. To metricize our evaluation, we used semantic similarity between the corrected output and reference sentences.

For our experiments, we compared two training approaches: from-scratch and fine-tuning. In the from-scratch setup, we initialized a sequence-to-sequence model using AutoModelForSeq2SeqLM.from_config with the configuration of facebook/bart-base, which creates a model with random weights and no prior linguistic knowledge. In contrast, in the fine-tuning setup, we loaded a pretrained model checkpoint specifically trained for Korean grammatical error correction using AutoModelForSeq2SeqLM.from_pretrained. This allowed the model to start with extensive knowledge of Korean grammar, morphology, and syntax. We used a sequence-to-sequence (Seq2Seq) architecture based on the BART model, which has shown strong performance across a wide range of text generation and correction tasks (Lewis *et al.*). Seq2Seq models are well-suited for grammatical error correction because they directly map an erroneous input sentence to a corrected output sequence. BART, in particular, is pretrained using denoising objectives that resemble the GEC task, making it an effective backbone for correction models. We selected this architecture due to its proven effectiveness in prior GEC research and its ability to generalize across languages through shared representations.[10]

Results and Discussion

To evaluate correction quality, we used several automatic metrics. The BLEU score (Bilingual Evaluation Understudy) measures n-gram overlap between the model output and a reference correction, and is widely used in machine translation and GEC evaluation.[11] Higher BLEU scores indicate closer surface-level similarity to the reference.

We also measured semantic similarity using sentence-embedding–based metrics, which estimate how well the corrected sentence preserves the meaning of the original reference. This metric captures meaning-level alignment even when surface forms differ.

Table 1: Summary of cross-lingual GEC performance across five automatic evaluation metrics.

Language	BLEU	METEOR	CHRF++	GLEU	Semantic
Japanese	0.4781	0.7097	0.7501	0.1728	0.8876
Russian	0.5151	0.7362	0.7764	0.1903	0.8985
German	0.5549	0.7470	0.7805	0.1545	0.9017
Thai	0.4156	0.6629	0.7183	0.1712	0.8559

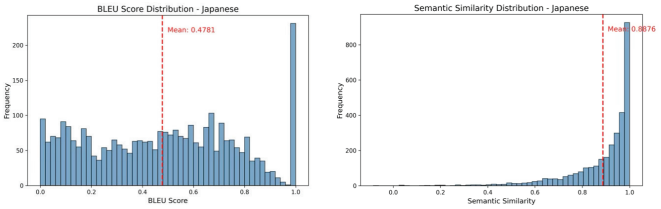

Figure 1: Distributions of BLEU score (top) and semantic similarity (bottom) for the Japanese-finetuned model. The histograms illustrate the frequency of different BLEU and semantic similarity scores, with the red dashed line and text indicating the mean values. The average BLEU score is approximately 0.4727, while the average semantic similarity is around 0.8996.

Japanese (depicted in Figure 1) achieved strong cross-lingual performance, with a mean BLEU score of 0.4727 and a semantic similarity score of 0.8996. Beyond BLEU, additional metrics such as METEOR and CHRF++ also showed

consistently high values (0.7097 and 0.7501, respectively), suggesting that both surface-level token overlap and character-level precision were well preserved. This multi-metric consistency strengthens the interpretation that the model's transfer was not limited to n-gram memorization but reflected structurally meaningful generalization.

Japanese shares subject-object-verb (SOV) word order and agglutinative morphology with Korean, meaning that grammatical roles are expressed through explicit morpheme stacking rather than rigid word position. This structural similarity likely facilitated the transfer of learned correction patterns, particularly those involving case marking, tense suffixes, and clause-final predicates. The high concentration of semantic similarity scores above 0.9 further indicates that sentence-level meaning was reliably preserved, supporting the Typological Similarity Hypothesis.

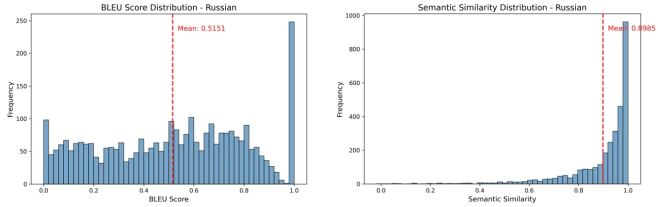

Figure 2: Distributions of BLEU score (top) and semantic similarity (bottom) for the Russian-finetuned model. The histograms illustrate the frequency of different BLEU and semantic similarity scores, with the red dashed line and text indicating the mean values. The average BLEU score is approximately 0.4712, while the average semantic similarity is around 0.8926.

Russian (depicted in Figure 2), despite being typologically distinct from Korean in both canonical word order (subject-verb-object, SVO) and morphological type (fusional rather than agglutinative), demonstrated unexpectedly strong transfer performance. The model achieved a BLEU score comparable to Japanese and high values across METEOR and CHRF++, indicating robust lexical and subword-level alignment.

The BLEU distribution exhibited greater variance than in Japanese, suggesting less stable surface-level correction. However, semantic similarity scores remained high, with most outputs clustering above 0.85. This pattern indicates that while exact morphological realization was sometimes imperfect, the model preserved core grammatical relations and propositional meaning.

This finding lends support to the Morphological Richness Hypothesis. Russians' extensive case marking and verbal inflection may provide dense grammatical cues that activate structural attention patterns learned from Korean. Moreover, Russian's relatively flexible word order may reduce interference from the Korean SOV bias. Together, these factors suggest that deep morphosyntactic complexity—rather than superficial typological similarity alone—can enable effective cross-lingual generalization.

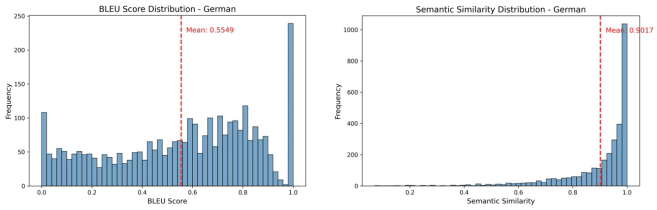

Figure 3: Distributions of BLEU score (top) and semantic similarity (bottom) for the German-finetuned model. The histograms illustrate the frequency of different BLEU and semantic similarity scores, with the red dashed line and text indicating the mean values. The average BLEU score is approximately 0.4451, while the average semantic similarity is around 0.8687.

German (depicted in Figure 3) showed moderate cross-lingual transfer, with intermediate scores across BLEU, METEOR, and CHRF++. While lexical and character-level overlap remained relatively stable, performance dropped compared to Japanese and Russian, indicating reduced structural alignment. The dispersion observed in the BLEU distribution suggests inconsistent realization of surface-level corrections, particularly in morphosyntactic agreement and clause structure.

A key source of difficulty likely lies in German's strict verb-second (V2) word order and its fusional inflectional system, which encodes multiple grammatical features—such as case, gender, and number—within single morphemes. Unlike Korean's agglutinative morphology, where grammatical markers are segmentable and sequential, German's internal morphological alternations provide less transparent cues. Furthermore, deviations from Korean's subject-object-verb (SOV) order introduce structural mismatches at the clause level.

Despite these divergences, semantic similarity scores remained comparatively high, suggesting that the model retained sentence-level meaning even when form-level accuracy decreased. This pattern indicates that semantic abstraction may transfer more readily than precise morphosyntactic realization when typological and syntactic structures differ.

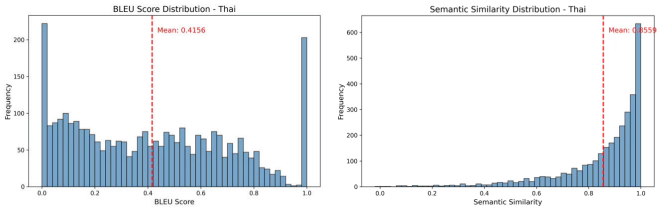

Figure 4: Distributions of BLEU score (top) and semantic similarity (bottom) for the Thai-finetuned model. The histograms illustrate the frequency of different BLEU and semantic similarity scores, with the red dashed line and text indicating the mean values. The average BLEU score is approximately 0.4028, while the average semantic similarity is around 0.8781.

Thai (depicted in Figure 4) exhibited the lowest BLEU score among the tested languages, while maintaining moderate semantic similarity. Across additional metrics such as METEOR and CHRF++, performance remained consistently below that of the morphologically rich languages, indicating weaker token- and character-level alignment. This pattern reflects limited structural transfer.

Typologically, Thai differs substantially from Korean. It follows subject-verb-object (SVO) word order and is an isolating

DOI: 10.36838/IJHSR812.53

language with minimal inflectional morphology. Grammatical relationships are expressed primarily through word order and function words rather than affixation. As a result, many of the morphological cues that the Korean-trained model relies upon—such as explicit tense or case markers—are absent. This structural gap likely constrained the model's ability to perform accurate form-level grammatical correction.

Nonetheless, semantic similarity scores remained relatively high compared to BLEU, suggesting that propositional meaning was often preserved even when surface grammatical corrections were imperfect. This divergence between form-level and meaning-level metrics reinforces the distinction between structural transfer and semantic abstraction. The Thai results, therefore, indicate that while morphological richness may facilitate transfer, meaning-oriented representations can still generalize across typologically distant languages to a limited extent.

Our main findings convey that cross-lingual grammatical error correction (GEC) is feasible, especially for languages that share structural characteristics—such as word order or morphological features—with the source language used for fine-tuning. The successful case of Japanese and the unexpectedly strong performance of Russian indicate that the model may have developed language-agnostic correction strategies, capable of extending beyond surface-level typological similarity.

Russian performance, in particular, supports this idea. Although typologically distant from Korean, Russian shares certain deeper features: morphological richness (e.g., extensive case and verb inflections), relatively flexible word order, and access to large multilingual corpora. These factors may have allowed the model to generalize abstract grammatical roles (subject, object, tense) learned from Korean to Russian, even if surface structures differ. This finding opens a valuable future research direction: to test whether models fine-tuned on other morphologically rich languages (e.g., Finnish, Hungarian) can likewise transfer effectively to other grammatically complex languages, regardless of typological distance.

The potential real-world applications of this approach are significant. Suppose a GEC model trained on Korean can generalize to languages like Russian, Turkish, or Uzbek. In that case, it suggests a cost-effective path toward building writing support tools for under-resourced and morphologically rich languages—without requiring large, language-specific error-annotated corpora. This could benefit: (1) Digital education platforms by enabling multilingual grammar checkers for heritage and second-language learners. (2) Governmental and institutional communication through improved grammar correction in machine-translated content. (3) Language learning tools, especially in mobile apps or AI tutors, offering personalized correction in a user's native or heritage language.

Despite these promising results, this study has several limitations that must be addressed.

First, the evaluation relied primarily on automatic metrics, including BLEU, METEOR, CHRF++, GLEU, and semantic similarity. While these metrics provide useful quantitative comparisons, they do not fully capture human judgments of grammaticality, fluency, or naturalness. For example, a corrected sentence may achieve a high BLEU score due to lexical overlap with the reference but still sound awkward or unnatural to a native speaker. Due to resource constraints and the multilingual scope of the experiment, human evaluation was not conducted in this study. Future work should incorporate human judgments using structured evaluation protocols, such as ratings for grammaticality, fluency, and meaning preservation, along with inter-rater agreement measures to ensure reliability.

Second, potential dataset bias across languages may have influenced cross-lingual comparisons. Although we controlled for dataset size by using the same number of sentences for each target language, other properties of the datasets were not fully aligned. The distribution of grammatical error types, lexical diversity, and syntactic complexity likely varied across languages. For instance, one dataset may contain predominantly tense or agreement errors, while another may emphasize word-order or particle usage errors. Such differences could affect model performance independently of typological similarity, making direct cross-language comparisons less precise. As a result, the observed ranking of languages may partially reflect dataset composition rather than purely structural transfer effects.

Third, the analysis did not examine performance by specific grammatical error categories. The current evaluation focused on aggregate scores, which makes it difficult to determine which linguistic features transferred most effectively across languages. For example, the model may have performed well on case-marking or tense-related errors while struggling with word-order or agreement errors in certain languages. Without a fine-grained error-type analysis, it is difficult to directly test the mechanisms behind the Typological Similarity and Morphological Richness hypotheses. Future research should include error-type-wise evaluation, using annotation schemes such as ERRANT or similar frameworks, to identify which grammatical categories transfer successfully and which remain challenging.

To address these limitations, future research should: Incorporate human evaluation using acceptability judgments, error severity scoring, and task-specific fluency ratings, Include additional automated metrics such as METEOR (for synonym and paraphrastic matching), CHRF++ (for character-level precision/recall suited to morphologically rich languages), and GLEU or ERRANT for fine-grained edit analysis, Analyze error-type-wise transferability, especially to identify which grammatical categories (e.g., case, aspect, modality) are most or least transferrable across language pairs.[8]

Finally, the significance of this study lies in demonstrating that a GEC model fine-tuned on a morphologically complex, low-resource language like Korean can generalize meaningfully to typologically diverse languages. This not only lowers the data barrier for building educational tools in less-resourced contexts, but also opens new directions in cost-efficient, multilingual NLP. Suppose the hypothesis about morphological richness as a transferable bias holds across other training languages. In that case, we may eventually be able to build foundational GEC models that serve as cross-lingual scaffolds for fine-tuning in dozens of languages with minimal effort.

Conclusion

This study demonstrates that a grammatical error correction (GEC) model fine-tuned on Korean—a morphologically rich, low-resource language—can successfully generalize to other languages. The model achieved its highest performance on Japanese, which shares the same SOV word order and agglutinative characteristics as Korean. This result supports the Typological Similarity Hypothesis, suggesting that greater structural similarity enhances cross-lingual transfer.

Notably, the model also performed exceptionally well on Russian, a language that is typologically distant but shares a high degree of morphological complexity with Korean. This finding strongly supports the Morphological Richness Hypothesis, indicating that deep grammatical complexity can facilitate transfer even when surface-level structures differ. In contrast, German and Thai showed more limited performance due to significant structural divergences.

These findings present a cost-effective and scalable approach for developing GEC tools in low-resource language settings. By leveraging the rich morphological features of a single source language, this method can reduce the reliance on large, language-specific datasets and promote the development of more inclusive multilingual education technologies.

Acknowledgments

I would like to express my sincere gratitude to my mentor, Sangyoon Bae from Seoul National University, for her invaluable guidance and insightful feedback throughout this research project. Her unwavering support and encouragement were instrumental in bringing this work to completion.

References

1. Ferris, D. R. Does Error Feedback Help Student Writers? New Evidence on the Short- and Long-Term Effects of Written Error Correction. In *Feedback in Second Language Writing: Contexts and Issues*; Hyland, K., Hyland, F., Eds.; Cambridge University Press, 2006; pp 81–104.
2. Sheen, Y. The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles. *TESOL Q.* 2007, *41* (2), 255–283.
3. Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. GECToR – Grammatical Error Correction: Tag, Not Rewrite. *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, 2020, pp 163–170.
4. Yamashita, I., Katsumata, S., Kaneko, M., Imankulova, A., and Komachi, M. Cross-Lingual Transfer Learning for Grammatical Error Correction. *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, 2020, pp 4704–4715.
5. National Center for Education Statistics. *Public High School Graduation Rates*. U.S. Department of Education, 2022. https://nces.ed.gov/programs/coe/indicator/coi (accessed Aug 26, 2025).
6. California Department of Education. *Cohort Outcome Data for the Class of 2021–22*. DataQuest, 2022. https://dq.cde.ca.gov/dataquest/ (accessed Aug 26, 2025).
7. National Council of Teachers of English. *English Language Learners: A Policy Research Brief*. NCTE. https://web.archive.org/web/20100821114640/http://www.ncte.org/library/NCTEFiles/Resources/PolicyResearch/ELLResearchBrief.pdf (accessed Aug 26, 2025).
8. Grundkiewicz, Roman, Christopher Bryant, and Mariano Felice. *A Crash Course in Automatic Grammatical Error Correction*. Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts. International Committee for Computational Linguistics, 2020. https://aclanthology.org/2020.coling-tutorials.6/
9. Keita, Sekou, Marie-Francine Moens, and Chris Emmery. Grammatical Error Correction for Low-Resource Languages: The Case of Zarma. *arXiv preprint*, 2024. https://arxiv.org/abs/2410.15539v2
10. Lewis, M., Liu, Y., Goyal, N., *et al.* BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ACL 2020*.

Author

Jiseop Kim is a high school student passionate about AI, linguistics, and cross-cultural education. Their research explores the intersection of natural language processing and second language learning. Jiseop plans to major in Cognitive Science or Computer Science, focusing on applications of AI in multilingual and heritage language learning.

DOI: 10.36838/IJHSR812.53