Systematic Feature Simplification in AI-Based Revision of English Tense-Aspect Morphology

Yuka Fujii, Yuichi Ono
Degree Programs in Humanities and Social Sciences, Institute of Humanities and Social Sciences
University of Tsukuba
{s 2520049,ono.yuichi.ga}@u.tsukuba.ac.jp

Abstract

Recent studies on AI-assisted writing have primarily evaluated large language models (LLMs) in terms of grammatical correction accuracy. However, relatively little attention has been paid to the systematic characteristics of AI revision errors. This study investigates whether AI-generated revisions of English tense-aspect morphology exhibit random errors or consistent patterns of grammatical simplification. Using 150,641 verb phrases extracted from the EFCAMDAT learner corpus, teacher corrections were compared with AI-generated revisions across six tense-aspect categories: Simple Present, Simple Past, Present Progressive, Past Progressive, Present Perfect, and Past Perfect. Generalized linear mixed-effects models were first used to identify tense-aspect categories that required teacher correction and to estimate the probability that AI reproduced teacher revisions. A confusion matrix was then constructed to examine the directionality of AI substitutions, followed by a feature-simplification analysis based on an independently defined grammatical complexity hierarchy. The results showed that teacher corrections occurred most frequently in perfect constructions, whereas AI agreement with teacher revisions varied substantially across categories. Most importantly, AI errors were highly systematic rather than random: progressive and perfect constructions were consistently replaced with structurally simpler tense forms, and simplification rates increased as grammatical complexity increased. These findings suggest that AI revision errors provide a novel source of evidence for investigating grammatical complexity in English tense-aspect morphology.

Keywords: AI-assisted writing, tense-aspect morphology, grammatical error correction, feature simplification, EFCAMDAT

1. Introduction

Large language models (LLMs) have recently become widely used for grammatical error correction (GEC) in second language (L2) writing. Previous studies have generally evaluated AI-assisted revision in terms of correction accuracy and its agreement with human judgments, demonstrating that modern LLMs achieve high overall performance in correcting learner errors (Kasneci et al., 2023; Mizumoto & Eguchi, 2023). However, relatively little attention has been paid to the linguistic characteristics of AI revision errors.

From the perspective of second language acquisition (SLA), errors often provide important evidence about underlying grammatical representations. In particular, English tense-aspect morphology has long been recognized as one of the most difficult grammatical domains for L2 learners because it requires the integration of tense and aspectual features (Andersen & Shirai, 1996; Bardovi-Harlig, 2000). If AI revision reflects similar constraints, its errors may also exhibit systematic patterns rather than occurring randomly.

The present study investigates whether AI-generated revisions of English tense-aspect morphology show systematic feature simplification. Using 150,641 verb phrases extracted from the EFCAMDAT learner corpus (Geertzen et al., 2013), teacher corrections were compared with AI-generated revisions across six tense-aspect categories. Teacher correction rates, generalized linear mixed-effects models (GLMMs), confusion-matrix analysis, and feature-simplification analysis were conducted to examine the characteristics and directionality of AI revision errors. By focusing on AI failures rather than successful corrections, this study explores whether AI revision errors can provide new evidence for understanding grammatical complexity in English tense-aspect morphology.

2. Related Work

2.1 AI-assisted grammatical error correction

Recent advances in large language models (LLMs) have substantially improved AI-assisted grammatical error correction (GEC). Previous studies have demonstrated that LLMs outperform conventional GEC systems in correcting a wide range of learner errors and produce revisions that are generally comparable to those made by human instructors (Kasneci et al., 2023; Mizumoto & Eguchi, 2023). Consequently, most existing studies have evaluated AI systems primarily in terms of correction accuracy, fluency, and agreement with human judgments. However, relatively little research has examined the linguistic characteristics of AI revision errors or whether unsuccessful revisions exhibit systematic grammatical tendencies.

2.2 English tense-aspect morphology in second language acquisition

English tense-aspect morphology has long been regarded as one of the most challenging grammatical domains for L2 learners. Previous SLA research has shown that tense-aspect categories differ considerably in acquisition difficulty because they involve different form–meaning mappings (Andersen & Shirai, 1996; Bardovi-Harlig, 2000). In particular, progressive and perfect constructions are acquired later than simple tense forms and frequently remain problematic even for advanced learners. More recent theoretical accounts have argued that these difficulties arise because learners must acquire and reorganize multiple grammatical features rather than individual surface forms (Lardiere, 2009; Slabakova, 2009).

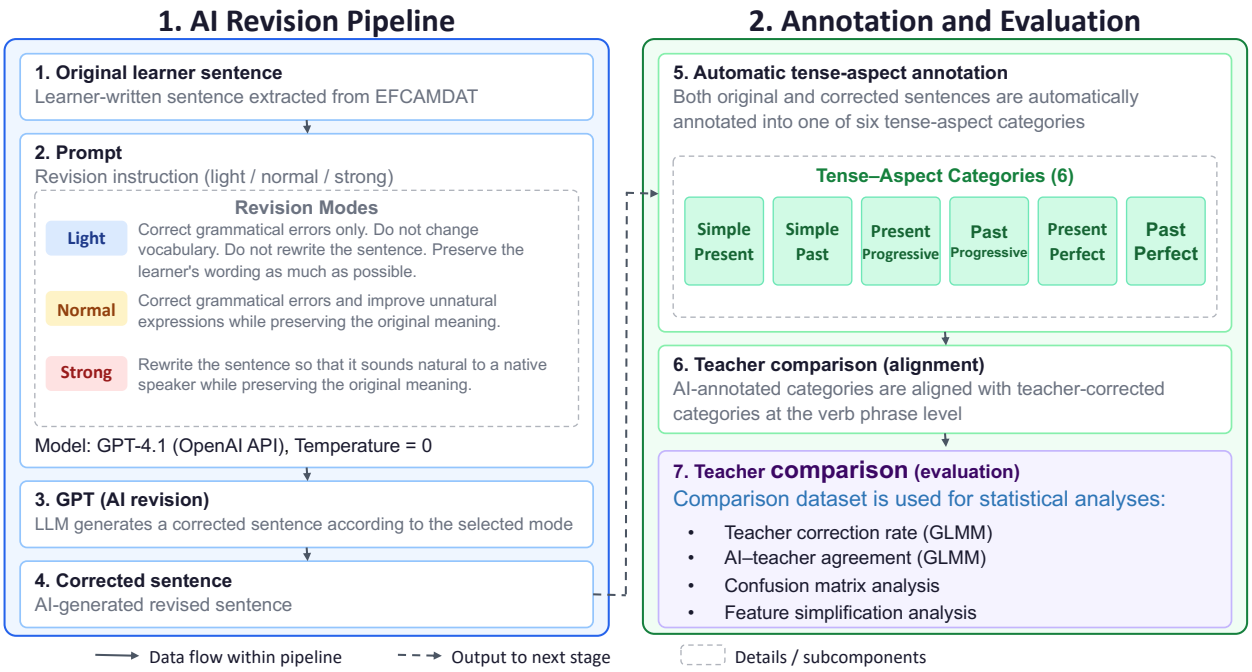

Figure 1 : AI revision and evaluation pipeline.

Building on these studies, the present research examines whether AI revision exhibits similar patterns of grammatical complexity. Rather than evaluating correction accuracy alone, this study investigates whether AI-generated revisions systematically simplify complex tense-aspect representations during grammatical revision.

2.3 From AI-Based Surface Normalization to Teacher-Oriented Revision

Our previous studies have investigated AI-assisted grammatical revision using the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 20xx), which contains writing produced under controlled writing tasks (Ono & Fujii, 2026; Ono et al., 2026). These studies examined the extent to which AI-based revision could normalize learner language toward native-speaker usage and identified grammatical features that consistently resisted automatic revision across different L1 backgrounds. Such persistent patterns were interpreted as L1 fingerprints, indicating grammatical representations that remain beyond the scope of surface-level normalization.

Although this line of research successfully identified grammatical features that AI failed to normalize, it did not evaluate whether AI revisions corresponded to the corrections that experienced teachers would actually produce. This limitation stems partly from the design of ICNALE, which provides learner production but does not include teacher-corrected reference texts.

The present study extends this research by employing the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2013), which contains professionally corrected learner writing. The availability of teacher corrections enables a direct comparison between AI-generated revisions and human expert judgments. Rather than asking which learner features remain after AI revision, the present study investigates whether AI reproduces teacher corrections consistently across different tense-aspect categories and whether systematic discrepancies emerge between AI revision and teacher correction.

3. Method

3.1 AI Revision Pipeline

Figure 1 illustrates the analytical pipeline adopted in the present study. The analysis began with learner-written sentences extracted from the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2013). Each sentence was submitted to a large language model (LLM) for grammatical revision using one of three predefined revision modes. The revised sentence was then automatically annotated for tense-aspect information and aligned with the corresponding teacher-corrected version at the verb-phrase level. The aligned dataset was subsequently used for statistical analyses of teacher corrections, AI-teacher agreement, confusion patterns, and feature simplification.

3.2 AI Revision Pipeline

Grammatical revision was performed using GPT-4.1 through the OpenAI API. Three prompting conditions were designed to control the degree of revision: light, normal, and strong. The prompts differed only in the extent to which the model was allowed to modify the original learner sentence.

The light mode instructed the model to correct grammatical errors only while preserving the learner's wording and sentence structure as much as possible. The normal mode additionally allowed improvements to unnatural expressions provided that the original meaning was maintained. The strong mode instructed the model to rewrite the sentence so that it sounded natural to a native speaker while preserving the intended meaning.

Since the objective of the present study was to compare AI revisions directly with teacher corrections at the

grammatical level, the analyses reported in this paper focus on the light revision condition. This setting minimizes lexical and stylistic rewriting while targeting grammatical errors, thereby enabling a more direct comparison between AI-generated revisions and teacher corrections.

3.3 AI Revision Pipeline

The dataset was constructed from the EF-Cambridge Open Language Database (EFCAMDAT), a large-scale learner corpus of English writing (Geertzen et al., 2013). A total of 150,641 verb phrases were extracted and automatically classified into six tense-aspect categories: Simple Present, Simple Past, Present Progressive, Past Progressive, Present Perfect, and Past Perfect. Teacher-corrected categories and AI-generated categories were aligned at the verb-phrase level to create the analysis dataset.

Four statistical analyses were conducted. First, teacher correction rates were examined using a generalized linear mixed-effects model (GLMM) with teacher correction as the dependent variable, tense-aspect category as the fixed effect, and writing ID as a random intercept. Second, AI-teacher agreement was analyzed using a second GLMM with revision version, tense-aspect category, and their interaction as fixed effects. Estimated marginal means were calculated to estimate the probability that AI reproduced teacher corrections across grammatical categories.

Third, a confusion-matrix analysis was conducted to investigate the directionality of AI substitutions between teacher-corrected and AI-generated tense-aspect categories. Finally, a feature-simplification analysis examined whether AI revisions systematically transformed teacher-corrected categories into grammatically simpler categories according to a predefined complexity hierarchy (**Simple Present < Simple Past < Present Progressive < Present Perfect < Past Progressive < Past Perfect**). A revision was classified as a simplification when the AI-generated category occupied a lower position than the corresponding teacher-corrected category.

All statistical analyses were conducted in R (version 4.3.3) using the lme4, car, emmeans, tidyverse, and ggplot2 packages. Generalized linear mixed-effects models were fitted using the glmer() function with a binomial distribution, and all figures and tables were generated within the R environment.

4. Results

4.1 Teacher correction rates across tense-aspect categories

We first examined how frequently each tense-aspect category was corrected by human teachers. Figure 2 presents the teacher correction rates for the six grammatical categories, while the corresponding GLMM results are summarized in Table 1.

Generally, teacher corrections were relatively infrequent across the corpus, reflecting the generally high quality of learner production. Nevertheless, substantial differences were observed among tense-aspect categories. Past Perfect showed the highest correction rate (1.96%), followed by Present Perfect (0.97%) and Simple Past (0.92%). Simple Present exhibited the lowest correction rate (0.31%), whereas Past Progressive (0.73%) and Present Progressive (0.61%) occupied intermediate positions.

The generalized linear mixed-effects model confirmed significant differences across grammatical categories. Compared with the reference category (Simple Present), teacher corrections were significantly more frequent for Simple Past ($p < .001$), Past Progressive ($p = .015$), Present Perfect ($p < .001$), and Past Perfect ($p < .001$), whereas Present Progressive did not differ significantly ($p = .257$). These findings indicate that perfect constructions, together with past-tense morphology, constitute the grammatical domains that most frequently require teacher intervention.

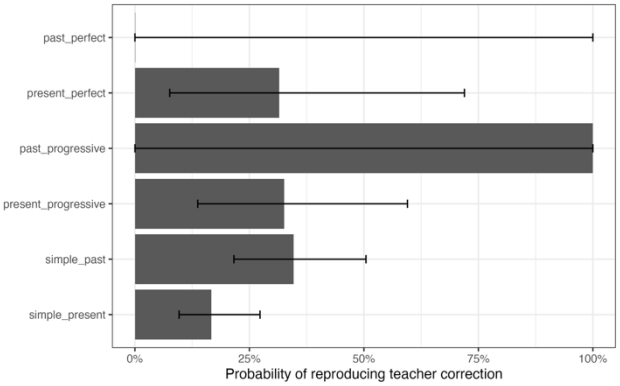

Figure 2: Teacher correction rates across six tense-aspect categories.

Predictor	Estimate	SE	*z*	*p*
Simple Past	1.011	0.146	6.904	< .001
Present Progressive	0.291	0.257	1.133	.257
Past Progressive	1.331	0.545	2.441	.015
Present Perfect	1.48	0.334	4.431	< .001
Past Perfect	1.484	0.584	2.539	.011

Table 1 : GLMM predicting teacher correction.

Notes: Simple Present was the reference category. The model included a random intercept for writing ID.

4.2 AI-teacher agreement

The second analysis examined the extent to which AI revisions reproduced teacher corrections. Figure 3 presents the estimated probabilities obtained from the generalized linear mixed-effects model.

Agreement between AI and teacher corrections varied considerably across grammatical categories. The highest agreement was observed for Past Progressive (95.5%), although this estimate was associated with a wide confidence interval because only a small number of teacher-corrected instances were available. For the remaining categories, agreement probabilities ranged from approximately 17% to 35%. Simple Present showed the lowest probability of agreement (17.2%), whereas Simple Past (35.1%), Present Progressive (33.1%), Present Perfect (33.3%), and Past Perfect (14.8%) exhibited relatively modest agreement.

These results suggest that, even when teachers identified grammatical errors, AI did not necessarily reproduce the same corrections. Descriptively, agreement probabilities

varied across tense-aspect categories. However, the generalized linear mixed-effects model did not detect statistically significant effects of tense-aspect category or revision version. These results should therefore be interpreted as descriptive rather than confirmatory.

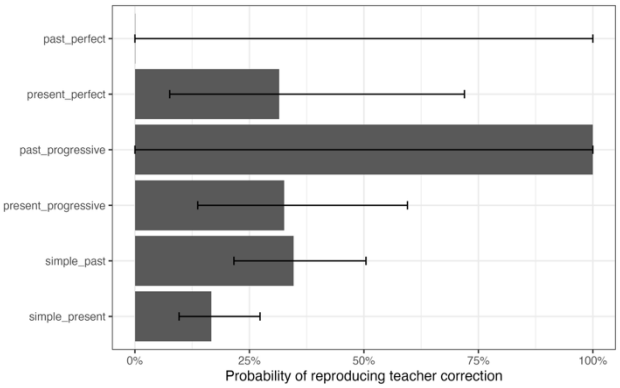

Figure 3 : Estimated probability that AI reproduces teacher corrections. Error bars represent 95% confidence intervals.

4.3 Directionality of AI substitutions

To investigate whether AI errors occurred randomly, a confusion-matrix analysis was conducted using the light revision condition (Figure 4).

The confusion matrix revealed highly systematic substitution patterns. Teacher-corrected Present Perfect forms were revised as Simple Present in 73% of cases and as Simple Past in the remaining 27%, with virtually no reproductions of the original category. Likewise, Past Perfect was most frequently transformed into Simple Present (43%) or Simple Past (29%). Present Progressive was revised as Simple Present in 50% of cases and as Simple Past in 11% of cases, while only 32% of revisions retained the progressive construction.

Simple tense forms exhibited comparatively stable behavior. Teacher-corrected Simple Present remained Simple Present in 57% of instances, whereas Simple Past remained unchanged in 34% of cases. Overall, AI substitutions were overwhelmingly directed toward simpler tense categories rather than distributed randomly across the six grammatical classes.

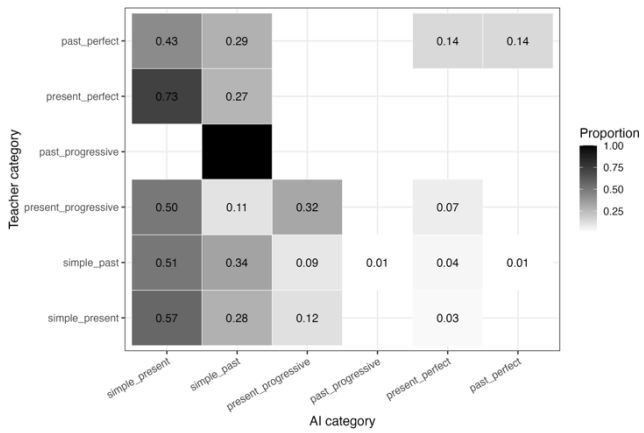

Figure 4 : Confusion matrix showing teacher-corrected and AI-generated tense-aspect categories.

4.4 Feature simplification

Finally, we examined whether AI revisions systematically reduced grammatical complexity. Figure 5 summarizes the proportion of revisions classified as feature simplification according to the predefined complexity hierarchy.

The results demonstrate a clear monotonic relationship between grammatical complexity and simplification. No simplification was observed for Simple Present (0%). In contrast, Simple Past showed a simplification rate of 50.5%, followed by Present Progressive (60.7%). Past Perfect exhibited a simplification rate of 85.7%, while both Present Perfect and Past Progressive reached 100%, indicating that every AI revision transformed these constructions into grammatically simpler categories.

In summary, these findings provide strong evidence that AI revision errors are not random. Instead, the model systematically favors grammatically simpler tense-aspect representations, particularly when revising progressive and perfect constructions.

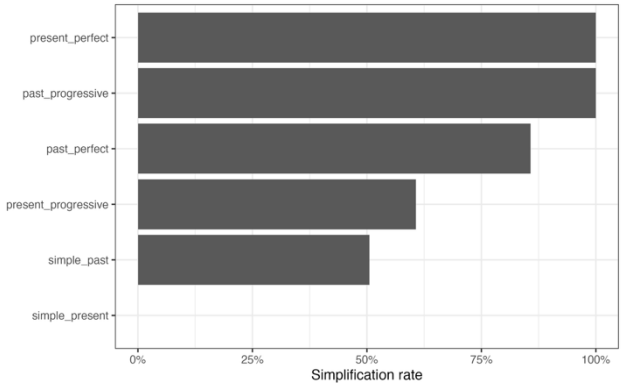

Figure 5. Feature simplification rates across tense-aspect categories.

5. Discussion

5.1 Teacher correction rates across tense-aspect categories

The first contribution of the present study extends previous research on AI-assisted grammatical error correction. As discussed in Section 2.1, existing studies have primarily evaluated AI revision in terms of grammatical accuracy, fluency, and overall writing quality (Kasneci et al., 2023; Mizumoto & Eguchi, 2023). Although these studies have demonstrated that large language models can generate linguistically acceptable revisions, they provide relatively little evidence regarding whether AI makes the same grammatical decisions as experienced teachers.

To address this gap, the present study introduced a teacher-oriented evaluation framework by directly comparing AI-generated revisions with teacher corrections in a large learner corpus. Rather than asking whether AI can produce grammatically correct output, we examined whether AI reproduced the revisions made by experienced teachers. Although the generalized linear mixed-effects model did not reveal statistically significant differences among tense-aspect categories, the descriptive analyses indicate that AI did not reproduce teacher corrections uniformly across

grammatical categories. These findings should therefore be interpreted as exploratory rather than confirmatory.

In summary, the present findings suggest that grammatical accuracy and teacher agreement represent different dimensions of AI performance. Teacher-corrected learner corpora therefore provide an important complementary benchmark for evaluating AI-assisted revision and extend the evaluation framework adopted in previous AI-assisted writing research.

5.2 AI Revision and the Complexity of English Tense-Aspect Morphology

T Section 2.2 reviewed SLA research demonstrating that English tense-aspect categories differ systematically in acquisition difficulty because they require different degrees of feature integration and form–meaning mapping (Andersen & Shirai, 1994; Bardovi-Harlig, 2000; Lardiere, 2009). The present study extends this line of research by examining whether grammatical complexity is also reflected in AI-assisted revision.

Although AI–teacher agreement did not differ significantly across grammatical categories, both the confusion-matrix analysis and the feature-simplification analysis revealed highly systematic revision behavior. Progressive and perfect constructions were frequently transformed into Simple Present or Simple Past forms, and simplification rates increased consistently as grammatical complexity increased. These findings indicate that AI revision errors are not randomly distributed but exhibit clear directional tendencies toward grammatically simpler tense-aspect representations.

Taken together, these findings suggest that AI revision is broadly consistent with the grammatical complexity hierarchy proposed in second language acquisition research. Rather than preserving more complex tense-aspect distinctions, AI tends to favor morphologically and semantically less complex alternatives when multiple revision options are available. Although the mechanisms underlying AI revision differ fundamentally from those of human language acquisition, the observed simplification patterns indicate that grammatical complexity provides a useful theoretical framework for understanding AI revision behavior.

5.3 From Surface Normalization to Teacher-Oriented Evaluation

The third contribution extends our previous research on AI-assisted surface normalization (Ono & Fujii, 2026; Ono et al., 2026). Using ICNALE, our earlier studies investigated AI revision in controlled writing tasks and examined how closely AI-generated revisions approximated native-speaker usage. Those studies identified grammatical features that consistently resisted automatic revision across different L1 backgrounds and interpreted these residual patterns as L1 fingerprints, representing aspects of learner language that remain beyond surface-level normalization.

However, because ICNALE does not provide teacher-corrected reference texts, previous studies could not determine whether AI revisions actually reflected expert correction practices. Consequently, the focus remained on identifying learner features that AI failed to normalize rather than evaluating the appropriateness of AI revisions from a pedagogical perspective.

By employing EFCAMDAT, which contains professionally corrected learner writing, the present study introduces a teacher-oriented perspective on AI revision. Instead of asking what AI fails to normalize, we ask whether AI reproduces teacher corrections. The results demonstrate that AI revision errors are systematic rather than random, exhibiting a strong tendency toward feature simplification. This finding extends the concept of surface normalization by showing that successful normalization does not necessarily imply agreement with expert human correction.

More broadly, these findings suggest that AI-assisted revision should be evaluated not only in terms of linguistic acceptability but also in relation to pedagogically appropriate grammatical decisions. Comparing AI revisions with teacher corrections provides a promising framework for understanding the strengths and limitations of generative AI in language education and offers a new direction for future research on AI-supported second language writing.

6. Conclusion

This study investigated AI-assisted grammatical revision from the perspective of teacher-oriented evaluation, focusing on English tense-aspect morphology. Using teacher-corrected learner writing from EFCAMDAT, we directly compared AI-generated revisions with human teacher corrections across six tense-aspect categories.

The results yielded three main findings. First, agreement between AI-generated revisions and teacher corrections varied substantially across grammatical categories, indicating that AI did not reproduce teacher corrections uniformly. Second, confusion-matrix and feature-simplification analyses demonstrated that AI revision errors were systematic rather than random. In particular, progressive and perfect constructions were frequently simplified into grammatically less complex tense categories. Third, by comparing AI revisions with teacher corrections, the present study extends previous research on AI-assisted surface normalization and introduces a new framework for evaluating AI revision in relation to expert pedagogical judgment.

These findings suggest that teacher-corrected learner corpora provide an important benchmark for assessing AI-assisted grammatical revision. While AI is capable of producing grammatically acceptable revisions, its revision strategy does not always align with the decisions made by experienced teachers, particularly for morphologically and semantically complex grammatical categories.

Future research should examine whether similar patterns of systematic simplification are observed in other grammatical domains, including articles, passive constructions, relative clauses, and modality. Expanding teacher-oriented evaluations across a wider range of grammatical phenomena will contribute to a deeper understanding of both the capabilities and the limitations of generative AI in second language writing instruction. Rather than viewing AI-assisted revision solely as an

automatic error-correction tool, future research should evaluate how closely AI revisions reflect the pedagogical decisions made by human teachers.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 24K00081.

References

Andersen, Roger W., and Yasuhiro Shirai. 1994. Discourse motivations for some cognitive acquisition principles. Studies in Second Language Acquisition, 16(2): 133–156.

Bardovi-Harlig, Kathleen. 2000. Tense and Aspect in Second Language Acquisition: Form, Meaning, and Use. Oxford: Blackwell.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1): 1–48.

Fox, John, and Sanford Weisberg. 2019. An R Companion to Applied Regression. Third Edition. Thousand Oaks, CA: Sage.

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In Proceedings of the 31st Second Language Research Forum (SLRF 2012), pp. 240‑254.

Ishikawa, Shin'ichiro. 2023. The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. London: Routledge.

Kasneci, Enkelejda, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103: 102274.

Lardiere, Donna. 2009. Some Thoughts on the Contrastive Analysis of Features in Second Language Acquisition. Second Language Research, 25(2): 173–227.

Lenth, Russell V. 2025. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 2.0.4.

Mizumoto, Atsushi, and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2): 100050.

Ono, Yuichi, and Yuka Fujii. 2026. Generative Constraints and Domain-General Learning: AI Editing as a Test Case for Tense-Aspect Feature Reassembly. Paper presented at the 18th Meeting of Generative Approaches to Second Language Acquisition (GASLA 18). Cambridge, UK, April 16–18.

Ono, Yuichi, Kasumi Takahashi, Reina Mogushi, Yuka Fujii, and Shota Kato. 2026. Surface Normalization Without Feature Reassembly: Overt Distribution, Underuse, and the Limits of Generative-AI Revision in L2 English. Paper presented at the 26th International Annual Conference of the Japan Second Language Association (J-SLA 2026). Mie University, Japan, July 4–5.

R Core Team. 2025. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Slabakova, Roumyana. 2016. Second Language Acquisition. Oxford: Oxford University Press.

Wickham, Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. Second Edition. Cham: Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43): 1686.