IJMR | International Journal of Multidisciplinary Research
Международный институт управления и предпринимательства
Research Article Print ISSN 3105-8884 Online ISSN 3105-8892 January 2026, Vol. 2, No. 1
https://doi.org/10.65231/ijmr.v2i1.123

Big Data Analysis Of Common Grammar Errors Among EFL Learners: A Basis For Improving English Writing Instruction Aligned with SDG 4 (Quality Education)

Ivy Derla[1], Dr. Sanjayan T.S.[2], Dr. Avelino D. Bitang[3],Yuanyuan Li[4]

[1].Shinawatra University, 99 M10 Bangtoei Subdistrict, Samkhok District, Pathumthani 12160, Thailand

[2]College of Education, Goa University, Goa India

[3]University of Mindanao, Bolton Street, Corner Bonifacio Street, Davao City 8000 Philippines

[4].International Institute of Management and Business,220086, Minsk City, Belarus

KEYWORDS

Corpus-based research;

EFL writing;

Grammar error analysis;

NLP-assisted tools;

Data-informed instruction;

Sustainable development Goal 4

ABSTRACT

The increasing availability of digital learner data has opened new possibilities for enhancing the quality of English as a Foreign Language (EFL) instruction. This study investigates recurring grammar error patterns in EFL learners' writing using a corpus-based, big data-informed approach. Grounded in Error Analysis Theory and Interlanguage Theory, the study analyzes a large digital corpus of learner- written texts to identify systematic grammatical difficulties, including errors in verb tense, article usage, prepositions, subject-verb agreement, and sentence structure.

The findings reveal these grammar errors are not random performance mistakes but consistent indicators of learner's developing interlanguage system. Unlike traditional grammar instruction, tends to look at small scales or small samples like a few essays, maybe one class and limited classroom samples while digital corpora and NLP-assisted analysis enables to see and identify thousands of texts and writings. Big data does not work as a substitute for teachers, but as a tool that strengthens instructional insight through large-scale evidence. To allow us to move from assumptions to evidence and from guessing to knowing.

By informing data-driven curriculum design, supporting AI-assisted feedback practices, and enhancing teacher decision-making, the study demonstrates how large-scale learner data analysis can contribute to more effective and inclusive EFL writing instruction.

In alignment with Sustainable Development Goal 4 (Quality Education), this research highlights the potential of evidence-based and learner-centered teaching approaches to promote more effective, inclusive, and equitable EFL writing instruction across diverse educational contexts.

INTRODUCTION

In the 21st century, education has entered the era of data. Every sentence written by a student becomes part of a larger story that can reveal how language is acquired and developed. Yet, in the world of teaching specifically language teaching, we are often guided with intuition and limited classroom observation rather than empirical evidence. This study was inspired by a simple and real classroom experience: behind every grammar mistake is a learning opportunity. However, it gives the idea and realization that there are some failed strategies in traditional teaching. So, big data analysis enters the classroom, not to replace the teachers but one that allows educators to empower insights

* Corresponding author. E-mail address:ivy.t@siu.ac.th

Received date:January 10, 2026; Revised manuscript received date: January 20, 2025; Accepted date: January 25, 2025; Online publication date:January 30, 2026.

drawn from thousands of learner output. Traditional approaches to grammar instruction usually rely on small classroom samples and teacher intuition. While these approaches are valuable, they may not capture broader tendencies across learners and tasks. With the increasing use of digital writing platforms, it has become possible to test learner writing in larger numbers. This creates an opportunity to move beyond isolated examples and toward a more comprehensive understanding of grammar errors.

In many EFL classroom, students are able to generate meaningful ideas but struggle to present them accurately and comprehensively in written English. Teachers often spend considerate time correcting grammar, yet similar errors continue to appear ibn students' work. This situation raises an important question: are grammar errors simply mistakes, or do they reflect deeper patterns in how learners develop written language?

Learning English as a foreign language can be challenging, especially in writing. Many students make similar grammar mistakes that can affect their ability to communicate clearly. Furthermore, it looks at the most common grammar errors made by EFL learners in their writing. Understanding these mistakes will help teachers improve how they teach English writing, making learning easier and more effective.

Many students learning English as a second language struggle with grammar when writing. These errors can slowdown their progress and make communication vague. By analyzing common grammar errors found in EFL learners' writing, this research aims to support and improve better teaching methods that address these challenges effectively.

Writing can be a challenge filled with many grammar errors. These errors often result from misunderstanding of English rules. It collects and looks at those frequent errors to guide educators and language advocates in developing clearer and more helpful writing instruction.

This paper examines how data-based analysis can be used to identify common grammar error patterns among EFL learners and how these insights can help improve English writing instruction. By focusing on recurring errors rather than isolated mistakes, the study aims to provide a clearer basis for improving English writing instruction. In doing so, it seeks to support teaching practices that are more responsive to learner needs and aligned with the objectives of SDG 4, particularly in promoting effective, inclusive, and quality education.

1.THEORETICAL FRAMEWORK

This research is anchored on three major theories that explain why learners make grammar errors and large-scale data can provide deeper insight into their writing performance. The framework integrates Corder's Error Analysis Theory, Selinker's Interlanguage Theory, and the Big Data Analytics Model. Together, these theories guide the analysis, interpretation, and application of findings for improved English writing instruction.

1. Corder's Error Analysis Theory (1967) main idea that leaners' errors are not sign of failure but valuable evidence of learning in progress. It points out that classifying errors into systematic categories such as verb tense, articles, prepositions, S-V agreement, and sentence structure. Big data strengthens Corder's theory by allowing researchers to examine thousands of errors at once, making error patterns more reliable and visible.
2. Selinker's Interlanguage Theory (1972) introduced thew concept of interlanguage-which is influenced by both their first language (L1) and the target language (L2). Interlanguage is a temporary, evolving system that changes as learners gain more exposure and feedback. Interlanguage contains:

(1) The learner's first language (L1)
(2) General learning strategies
(3) Rules invented by the learner to fill the gaps in their knowledge
(4) Errors are signs that the interlanguage system is developing and self-adjusting over time.

This research, Interlanguage Theory helps explain:

(1) Why certain errors are common across L1 groups?
(2) Why some grammar forms develop earlier or later?
(3) How learners' internal rules influence their writing?
(4) Why predictable patterns appear consistently in the data?

By analyzing errors of writing samples, big data allows us to observe how interlanguage forms and evolves on a large scale.

Connection between the two theories:

Corder's theory tells us why it's important to study learners' errors, while Selinker's theory explains how those errors

form and evolve. Together, they provide the theoretical foundation for using big data analysis to trace and understand patterns of grammar errors among EFL learners.

2.PURPOSE OF THE STUDY

The purpose of this study is to examine recurring grammar error patterns in the written work of EFL learners using a big data-informed analytical approach. By identifying common areas of grammatical difficulty across learner texts, the study aims to provide a clearer basis for improving English writing instruction. Specifically, the study seeks to move beyond learners' level error correction and toward a deeper understanding of how grammar errors reflect learners' language development. The findings are intended to support more effective and inclusive writing pedagogy, in line with the goals of quality education under SDG 4.

3.RESEARCH QUESTIONS

1. What types of grammar errors commonly appear in the written texts of EFL learners?
2. Which grammar error categories occur most frequently across leaner text samples?
3. In what ways can a big data-informed approach contribute to more effective and inclusive EFL teaching practices aligned with SDG 4?

4.RELATED LITERATURE

Many experts believe that technology will not replace teachers, but rather will transform the role of teachers. In the past, teachers were the main source of information and knowledge. The future teacher will be more like a coach, helping students to find and use information for themselves (*Ashrafimoghari,2022)*. It is through helping educators and designers redesign and enhance their instructional insights and lesson accordingly. Learning analytics is a data-driven approach to understand how people learn, so that you can help them learn better (*Ashrafimoghari,2022*). It is design to analyze thousands of students writing and outputs.

Errors are the deviations or wrong forms of a language reflecting the competence of the learner. There can be various causes of errors. One of the causes of errors is the ignorance of appropriate rule in the foreign language. (Dr. Neupane cited Corde, 1999) Errors are not merely a deviation but rather, a reflection of growth and cognitive development. Error analysis is the systematic study and analysis of the errors committed by second language learners (Richards & Selinker, 2008). Grammatical error for language learners has recently attracted increasing interest in the Natural Language Processing (NLP) community. Grammatical error has the potential to create commercially viable software tools for the large number of students around the world who are studying a foreign language, in particular the large number of students of English as a Foreign Language (EFL) (Dahlmeier et al). NLP is a tool use detect large scale of student's writings and can help teachers detect errors which are in five categories: verb tenses, subject – verb agreement, preposition, articles, and sentence structure. The results show that verb tense was the most frequent. Article omission and preposition errors were also common, especially among learners whose first language differ structurally from English. As suggested by Corder (1967), mistakes are related to problems in performance, just like a slip of the tongue or pen. However, errors are systematical complications that indicate competence-related problems, contributing to the learner's progress (Gazioglu et al., 2024 cited Corder,1967). The most important is that these patterns were consistent across thousands of texts, and it is systematic, predictable, and teachable.

The biggest obstacle for grammatical error correction has been that until recently, there was no large, annotated corpus of learner text that could have served as a standard resource for empirical approaches to grammatical error correction. (Dahlmeier cited Leacock et al.,2010). Through hard work and consistency, they were able to create the UNS Corpus of Learner English (NUCLE), a large, annotated corpus of learner texts that freely available for research purposes. which can also be used as a tool. NUCLE can examine over thousands of students essays with a total of over one million words which are completely interpreted with error tags and corrections. Though, NUCLE has been there and available for two years now, there has been no reference paper that describe the details of the corpus.

Natural Learning Processing (NLP) tools handle tasks like text analysis, tagging, and pattern recognition in large datasets, such as learner text for grammar error detection. In EFL research, they automate error identification in writing samples, far beyond manual counts. Tools like AntConc or Wmatrix exemplify this by generating words and frequency stars from big data. Some common options for conceptual

works include:

1. AntConc – free software for words, collocations, and keyword extraction; ideal for spotting verb tense issues in EFL texts (Gazioglu, M., & Aydin, S. (2024)
2. Wmatrix – web interface with semantic tagging (USAS/CLAWS); compares sub-corpora to track learner progress. (Wmatrix7 onwards)
3. Sketch Engine – commercial tool for word sketches and grammar profiling; supports multilingual error analysis (Kilgarriff, A et al., 2004)

These are some tools which detect students' writing errors and can enhance teacher's insights at scale and teaching strategies. Another tool is spaCy designed for industrial -strength natural language processing (NLP), excelling in tasks like tokenization, part-of-speech tagging, dependency parsing, and named entity recognition. spaCy automates grammar profiling via dependency parsing (Honnibal & Montani,2017), pairing with AntConc for hybrid EFL analysis.

Research indicates that the use of digital learning tools such as e-books can enhance language learning achievement by overcoming physical barriers between teachers and students and fostering a deeper understanding of learners' learning contexts (Songkhro et al., 2025). These findings support the present study's argument that technology -assisted and data-informed approaches can promote and enhance more inclusive and effective EFL instruction. Though e-books are not classified as big -data, their use in digital learning gives large-scale learner interaction data, which can be analyzed to inform data-driven and EFL teaching methods and practices.

5.METHODOLOGY

This study adopts a qualitative descriptive research design supported by a corpus-based, big data-informed analytical approach. The design focuses on identifying recurring and systematic grammar error patterns in EFL learners' written texts rather than evaluating errors one by one.

This approach is for examining real learner language and aligns with Error Analysis Theory (Corder,1967) and Interlanguage Theory (Selinker,1972), which emphasize systematic patterns of learner errors as evidence of language development.

DATA COLLECTION

The data consisted of a large collection of learner-written texts which include essays, reports, reflections, and academic papers produced and collected from EFL university learners at Shinawatra University, Thailand representing a range of proficiency levels. Specifically, the data is comprised approximately 300 essays written by Chinese EFL students by open argumentative topics like "Technology Role in Education" to encourage learners to produce original output. The use of a large dataset allows the study to move beyond isolated classroom samples and toward a more comprehensive understanding of grammatical difficulties experienced by learners. All learner texts were collected from regular coursework, anonymized prior to analysis, and used only for research purposes, with no impact on students' academic evaluation and performance.

ERROR CATEGORIES

Grammar errors were identified and classified into five major categories based on frameworks in Error Analysis Theory (Corder,1967) and Interlanguage Theory (Selinker,1972). The analysis focused on the following:

1. Verb tense
2. Subject -verb agreement
3. Article usage
4. Prepositions
5. Sentence structure and word order

These categories were selected because they are widely recognized as persistent problem areas in EFL writing and frequently associated with first language transfer effects.

ANALYTICAL PROCEDURE

We used Python- based natural language processing (NLP) tools, including spaCy assist in identifying grammatical errors, which were interpreted within established error analysis frameworks.

The analytical procedure involved are the following steps:

1. Compilation of learner -all learner written text were compiled into a digital corpus suitable for computational analysis.
2. Automated Error Detection – Python -based Natural Language Processing (NLP) tools, particularly spaCy were used to process corpus, it performed tasks such as

tokenization, part-of-speech tagging, and dependency parsing to identify grammatical deviations related to the error categories.

3. Error Classification – identified grammatical patterns and deviations were grouped according to five categories.
4. Manual interpretation – automated results were manually reviewed to ensure pedagogical relevance and theoretical alignment. This helped identify systematic errors from occasional mistakes and ensured accurate interpretation of findings.

6. DATA ANALYSIS

Although frequency tendencies were observed, the study does not aim to provide statistical generalization but rather qualitative pattern identification across large-scale learner data. This section presents the analysis of grammar errors identified in the EFL learners' written texts using a corpus -based, big data-informed approach. Rather than examining individual learner performance, the analysis focuses on recurring and systematic grammar error patterns across a large collection of learner texts.

Digital learner corpus was examined using NLP – assisted corpus analysis tools to identify frequently occurring grammatical forms and deviations into predefined categories based on established error analysis frameworks. The analysis emphasizes patterns, frequency tendency, and persistent errors, rather than precise numerical measurement.

The results confirmed what many teachers observe, but now supported by strong data. Verb tense misuse was the most frequent error type, especially with perfect tenses. These errors included inconsistent tense usage within sentences and inappropriate tense use in written discourse. Such error patterns are common aspect which are often influenced by learners' first language structures.

Article omission was extremely common among learners whose native language have no article system like Thai and Chinese. Many learners have demonstrated difficulty in exercising definite and indefinite articles accurately, because they never use them in their first language. This finding s supports interlanguage theory that learners develop internal rules that may not fully align what is being used normally in English language.

Preposition confusion was another major challenge especially using "in," "on," and "at." Learners frequently used incorrect preposition and confused them entirely.

Errors related to subject-verb agreement were also observed, specifically in complex sentence structures.

Finally, sentence fragments and run-ons are also present which is due to difficulty in organizing ideas in a second language.

7.DISCUSSION

The findings of this study indicate that grammar errors in EFL learners' writing are systematic and consistent rather than random. The frequent occurrence of errors related to verb tense, article usage, prepositions, subject-verb agreement, and sentence structure suggests that these areas represent persistent challenges in English writing world. These results reinforce the view that grammar errors should be understood as part of learners' linguistic development rather than as isolated mistakes.

We need to shift our focus from simply identifying errors one by one to understanding them. When we hear our student say "she go to school every day", are not signs carelessness or laziness but it is a reflection of learners' evolving internal rule systems. Large-scale learner data makes it possible to observe how grammatical errors evolve over time, offering insights to learners' movement toward more target-like language use. This supports Corder's (1967) argument that errors provide valuable evidence of learning in progress.

How can we apply this?

The study supports **data-informed curriculum design**, enables educators to prioritize instructional content based on recurring grammar error patterns identified through large-scale learner data. Common grammar errors like verb tense, article usage, and sentence structure, can be integrated into writing curricula, ensuring that teaching content directly addresses learner's most persistent difficulties.

In addition, the integration of **AI-powered feedback tools** and NLP-assisted corpus analysis offers practical support for EFL writing instruction. Integrate writing platforms that analyze student errors in real time to help teachers and provide instant, personalized feedback. AI powered tools, such as ChatGPT, can generate immediate, detailed, and personalized feedback on student writing, potentially alleviating the workload of educators and providing timely assitstance to learners (Guo et al.,2024; Lee and Moore,2024)

The use of big data and NLP tools also contributed to teacher empowerment. Teachers can use error analytics dashboards to visualize which grammar topics cause the most difficulty enabling targeted remediation. Enhancing approach such workshops and ongoing support, aligning with SDG 4's teacher developing targets.

1. Corpus Training- teachers query sub-corpora for authentic examples, creating tailored drills example: depend on vs depend in.
2. AI Integration – freeing time for mentoring on cross-linguistic contrasts.
3. Collaborative Design – peer forums share Sketch Engine outputs, refining rubrics.

Cross- linguistic awareness, teachers should understand how a student's first language shapes their English grammar and address those predictable transfer errors explicitly. By understanding how learner's first language influences English writing, teachers can interpret errors more constructively and adopt inclusive, learner-centered strategies. This is what big data offers: clarity, precision, and personalization.

In alignment with Sustainable Development Goal 4 (Quality Education), the findings emphasize how data-informed approaches can support more effective, equitable and sustainable EFL teaching. Rather than replacing human insight, large-scale data analysis amplifies teachers' professional judgement by providing clearer, evidence-based perspectives on learner needs.

Big data is about amplifying teacher's strategies and methodologies. It helps us see learners as unique data stories, each mistake a clue, each sentence a step toward fluency.

Grammar instruction should be informed by authentic learner data rather than assumptions. By focusing on frequent and persistent errors, teachers can design more effective and targeted writing instruction. Big data can be of big help to improve writing instruction and improve common grammar errors of the learners.

CONCLUSION AND RECOMMENDATIONS

This study demonstrates that large-scale, corpus-based analysis provides a strong foundation for understanding grammar error in EFL learners' writing. Language learning generates substantial amount of analyzable data, including essays, digital submissions, and classroom writing tasks. Traditionally, such data has been examined through small samples, limiting the totality of findings. A big data-informed approach addresses this limitation by revealing patterns and tendencies that are not visible through only classroom observation.

In the context of EFL teaching, corpus-based analysis enables educator to identify common grammar errors across groups of learners examine their persistence, and relate them to linguistic development and first-language influence. Grammar error should therefore be viewed not as simple deviation, but as reflection of ongoing interlanguage development. It provides a scientific foundation for improving instruction because it allows educators to:

1. Identify common grammar errors across large groups of learners, not just within one classroom.
2. Understands the frequency and distribution of these errors, showing which grammar areas need greater emphasis.
3. Link learner errors to first language influence, proficiency level, or learning environment.
4. Develop evidence- based teaching strategies that target the most frequent and persistent problems.

In short, the basis of big data analysis in English teaching is its ability to transform qualitative observations into quantitative evidence, giving teachers and researchers a clearer, data-driven picture of how students learn, struggle, and improve in using English. Big data when used with empathy and insights, becomes more than just numbers and algorithms. It becomes a bridge between transformation and understanding. Learner errors should not be viewed simply as deviations, but as indicators of ongoing linguistic development and cognitive processing.

As researchers and educators, we must use technology not to standardize learning, but to humanize it. To connect with our student's journeys, recognize their challenges and guide them with compassion and evidence.

Based on the findings, big data analysis helps us understand grammar errors as part of the learning process and provides a strong basis for improving English writing instruction. EFL teachers should **adopt corpus-based and data -informed approaches** to identify and address common grammar errors in writing instruction, educational institutions should support the use of **NLP-assisted and AI -powered feedback tools** as complementary resources to develop writing skills, teacher training programs should emphasize **cross-linguistic awareness** to help educators

better understand learner error patterns, and future research may develop to examine how grammar error patterns evolve over time across proficiency levels.

Overall, when used thoughtfully, large-scale learner data analysis offers a powerful means of supporting more inclusive, effective, and evidence-based EFL writing instruction, contributing directly to **the goals of quality education under SDG 4**.

REFERENCE

1. Songkhro, J., Ali Mohsen, S., Wateh, M., Maseng, N., & Chedoloh, A. (2025). Developing effective service communication: The role of e-book in enhancing English speech acts in high vocational training. Journal of Modern Management, Shinawatra University, 3(2).
2. Gazioğlu, M., & Aydın, S. (2024). Identifying grammatical errors and mistakes via a written learner corpus in a foreign language context. Journal of Language Research (JLR), 8(2), 91–106. https://doi.org/10.51726/jlr.1553484
3. Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In Proceedings of the 11th EURALEX International Congress, EURALEX 2004 (pp. 105–115). Lorient, France.
4. Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/api/annotation#section-citation
5. Corpus Analysis. (n.d.). https://corpus-analysis.com/
6. Alnemrat, A. (2025). AI vs. teacher feedback on EFL argumentative writing. Frontiers in Education. https://doi.org/10.3389/feduc.2025.1614673
7. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550.
8. Shao, S. (2025). The role of AI tools on EFL students' motivation, self-efficacy and anxiety. Learning and Instruction. https://doi.org/10.1016/j.learninstruc.2025.101XXX
9. Yiakoumetti, A. (2006). A cross-linguistic approach to language awareness: Can English phonics benefit Greek learners of English? Language Awareness, 15(3), 137–157. https://doi.org/10.2167/la403.0
10. Woll, N., & Paquet, P.-L. (2025). Developing crosslinguistic awareness through plurilingual consciousness-raising tasks. Language Teaching Research. https://doi.org/10.1177/13621688211056544
11. Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching.
12. Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching, 10(1–4), 209–231.