A Corpus-Based Description Of Grammatical Errors In Uzbek EFL Learner Writing

Morteza Amirsheibani

Faculty of Languages and Preschool Education, Tashkent University of Economics and Pedagogy (TIPU), Uzbekistan

Inobat Tuymanova

Faculty of Languages and Preschool Education, Tashkent University of Economics and Pedagogy (TIPU), Uzbekistan

Asaloy Hamroyeva

Faculty of Languages and Preschool Education, Tashkent University of Economics and Pedagogy (TIPU), Uzbekistan

Abstract: This research describes grammatical mistakes in English writing by Uzbek learners at the upper-intermediate (B2) level, based on a small learner corpus from real essays written by students in academic lyceums and language schools in Uzbekistan. With guidelines from both corpus linguistics and second language acquisition, the study works to sort out the various recurring types of errors and their classification, demonstrating the pattern of the learners' interlanguage that has been influenced by their L1 and developmental factors. The data, totaling about 20,000 words, were checked with AntConc to bring out frequency information and concordance lines. The greatest common errors were noted in the areas of verb tense/aspect, articles, prepositions, and subject-verb agreement. A lot of these mistakes seem to come from bad habits formed by Uzbek and Russian, which don't have some grammar rules in English. The results show the grammar difficulties for Uzbek students and can be used to make decisions about how to teach grammar, create a curriculum, and train teachers in EFL for all of Central Asia.

Keywords: Learner corpus, grammatical errors, Uzbekistan, EFL writing, interlanguage, L1 transfer, corpus linguistics, B2 level learners.

Introduction: Over the past few decades, grammar accuracy has remained a central concern in English as a Foreign Language (EFL) instruction, particularly in contexts where English is taught as a school subject rather than used as a medium of communication. In Uzbekistan, a country where English holds growing importance in education and international mobility, the demand for improved learner proficiency has increased significantly. Despite ongoing reforms in curriculum and teacher training, learners often continue to struggle with core areas of English grammar. This is especially evident in their writing, where persistent errors may indicate deeper issues in language acquisition, cross-linguistic influence, and instructional practices (Abduazizov & Umarova, 2022).

Corpus linguistics offers a powerful approach to understanding these issues by enabling researchers to analyze authentic learner data systematically. Learner corpora—collections of written or spoken texts produced by language learners—can reveal patterns of grammatical development, fossilization, and error (Granger, Gilquin, & Meunier, 2015). Unlike intuition-based approaches, corpus-based studies allow for the empirical observation of learner interlanguage—the evolving language system that reflects a learner's developmental stage and L1 background (Selinker, 1972; Ellis, 2008).

In Uzbekistan, most learners are native speakers of Uzbek or Russian, two languages that differ significantly from English in their grammatical systems. For instance, Uzbek is an agglutinative, article-less language with a flexible word order and limited tense distinctions, while Russian, although inflected and tense-marked, also lacks articles and employs different syntactic structures. These typological differences often result in recurring error patterns in learners' English output, particularly in areas like article use, prepositions, subject-verb agreement, and verb tense/aspect (Muminov, 2021; Nazarova & Kadyrova, 2023).

However, there remains a shortage of detailed corpus-based studies focused specifically on grammatical errors among Uzbek EFL learners. While large international learner corpora such as ICLE or EFCAMDAT offer valuable benchmarks, they rarely include learners from Central Asian countries, making it difficult to generalize findings. As a result, local data is essential to understand the specific needs of Uzbek learners and to inform context-sensitive teaching strategies (Hasanova, 2020).

This study aims to fill this gap by compiling and analyzing a small-scale corpus of written texts produced by B2-level Uzbek learners. The goal is to provide a descriptive overview of the most frequent grammatical errors and to explore the possible linguistic and pedagogical reasons behind them. In doing so, the study hopes to contribute to the growing field of learner corpus research and to offer insights that can support more effective grammar instruction in Uzbekistan and similar EFL contexts.

Review of the Literature

Grammatical accuracy in EFL writing has been the focus of extensive research in second language acquisition (SLA), particularly in relation to learner interlanguage development and error analysis. Early work by Corder (1967) emphasized the pedagogical value of learner errors, arguing that errors are not simply mistakes to be corrected, but evidence of the language learning process itself. Later, Selinker (1972) introduced the concept of interlanguage, highlighting how learners develop an intermediate linguistic system influenced by both their native language and the target language. This concept remains foundational in contemporary corpus-based studies of learner language (Ellis, 2008; Granger et al., 2015).

Over the past two decades, the emergence of learner corpora has significantly advanced the study of grammatical errors. Projects such as the International Corpus of Learner English (ICLE) and the Cambridge Learner Corpus (CLC) have enabled researchers to examine L2 writing from multiple L1 backgrounds, identifying both universal developmental patterns and L1-specific influences (Granger et al., 2015; Nesselhauf, 2004). These studies have consistently shown that certain grammatical categories—articles, verb tense/aspect, prepositions, and subject-verb agreement—are particularly error-prone across learner populations.

However, L1 transfer plays a major role in shaping error patterns. According to Odlin (1989), the degree of linguistic distance between a learner's first language and English can significantly affect the type and frequency of errors. For instance, L2 learners whose L1 lacks articles (e.g., Chinese, Russian, Uzbek) tend to omit or misuse articles more frequently than those whose L1 includes them (Ionin, Ko & Wexler, 2004). Similarly, learners from languages with flexible word order or agglutinative structures, such as Uzbek, may show difficulties in maintaining syntactic patterns typical of English (Muminov, 2021).

While global learner corpora have provided broad insights, the underrepresentation of learners from Central Asia—especially Uzbekistan—has limited the applicability of such findings to this region. Research from within the Uzbek context has begun to emerge in recent years. For example, Hasanova (2020) analyzed

student writing at tertiary institutions and found frequent misuse of tense forms and prepositions, attributing many errors to negative transfer and insufficient grammar instruction. Similarly, Nazarova and Kadyrova (2023) examined upper-intermediate learners' essays and identified patterns of confusion between simple past and present perfect, a well-documented issue among EFL learners from non-Indo-European language backgrounds.

Despite these efforts, there remains a lack of systematic, corpus-based studies of grammatical errors specific to Uzbek learners. Much of the existing research relies on intuition, classroom observation, or small error samples without rigorous frequency analysis. Given the increasing importance of English proficiency for academic and professional advancement in Uzbekistan, a clearer understanding of learner grammar through corpus analysis is both timely and necessary. As Römer (2022) notes, learner corpora are "an indispensable tool for understanding real learner language use and improving language pedagogy."

This study builds on previous work by adopting a descriptive, data-driven approach to grammar errors in Uzbek learner writing. By compiling and analyzing a dedicated learner corpus, it seeks to provide empirical evidence of error frequency, examine the influence of the L1, and suggest practical recommendations for improving grammar instruction in Uzbekistan's EFL classrooms.

METHODOLOGY

This study employs a descriptive, corpus-based approach to identify and analyze frequent grammatical errors in the written English of Uzbek EFL learners. The research is situated in the context of academic lyceums and private language schools in Tashkent and Samarkand, where English is taught as a compulsory subject at the upper-intermediate (B2) level.

Corpus Compilation

The learner corpus was compiled from handwritten and typed compositions written by 40 B2-level Uzbek learners aged 16 to 19. All students had studied English for a minimum of five years and were preparing for international exams such as IELTS or CEFR certification. A total of 80 essays were collected, each between 250 and 300 words, based on standardized writing prompts commonly used in classroom settings (e.g., opinion essays, problem-solution essays). After digitization and light anonymization, the corpus consisted of approximately 20,000 words.

To maintain the authenticity of learner output, no grammatical corrections were made to the texts.

Spelling mistakes were retained if they did not hinder the identification of grammatical structures. Texts were saved in plain text format and later imported into AntConc (Anthony, 2023), a free concordance program widely used in corpus linguistics, for analysis.

Learner Profile

The majority of participants were native speakers of Uzbek, with Russian as a second language for about one-third of the group. English was their third language. All students had received grammar-focused instruction in state schools or exam preparation courses, where explicit teaching of verb tenses, articles, and sentence structure is common. However, exposure to English outside the classroom—through media, reading, or interaction—was reported to be limited, a factor that may affect accuracy and fluency (Abduazizov & Umarova, 2022).

Error Identification and Categorization

A combination of manual and software-assisted analysis was used to identify grammatical errors. First, the corpus was scanned in AntConc to extract frequent bigrams and trigrams that appeared to deviate from standard English usage (e.g., "she go," "in the yesterday," "he have went"). Then, each composition was manually coded for grammatical errors using a categorization system adapted from James (1998) and Ferris (2004), covering the following error types:

- Verb tense and aspect (e.g., He go to school yesterday.)

- Article use (e.g., I saw a elephant.)

- Prepositions (e.g., in Monday instead of on Monday)

- Subject-verb agreement (e.g., They walks to school.)

- Plural and countability issues (e.g., informations, advices)

- Word order and sentence structure (e.g., Happy I am to see you.)

Each error was logged in a spreadsheet with its original context, corrected version, and category code. Frequencies were normalized per 1,000 words to allow for comparability between categories.

Reliability Measures

To ensure coding reliability, 10% of the corpus was double-coded by a second trained rater with experience in EFL grammar instruction. Inter-rater agreement was calculated using Cohen's Kappa, yielding a score of 0.86, indicating a high level of agreement.

RESULTS

The analysis of the learner corpus focused on identifying

the most common grammatical errors made by B2-level Uzbek EFL learners in their written compositions. A total of 642 grammatical errors were recorded across the 20,000-word corpus, yielding an error density of approximately 32.1 errors per 1,000 words. These errors were classified into six major categories, derived from both frequency and pedagogical relevance.

1. Verb Tense and Aspect Errors (28.5%)

The most frequent error type involved incorrect use of verb tenses and aspects. Learners regularly substituted base forms for past tense verbs or misused auxiliary verbs in perfect constructions. For instance:

- She go to school yesterday. → She went to school yesterday.

- I have saw the movie last night. → I saw the movie last night.

To identify recurrent patterns, the corpus was processed using AntConc 3.5.9 (Anthony, 2023). Searches for frequent bigrams such as "she go" revealed widespread errors in marking past tense. Below is a sample of concordance lines generated by AntConc:
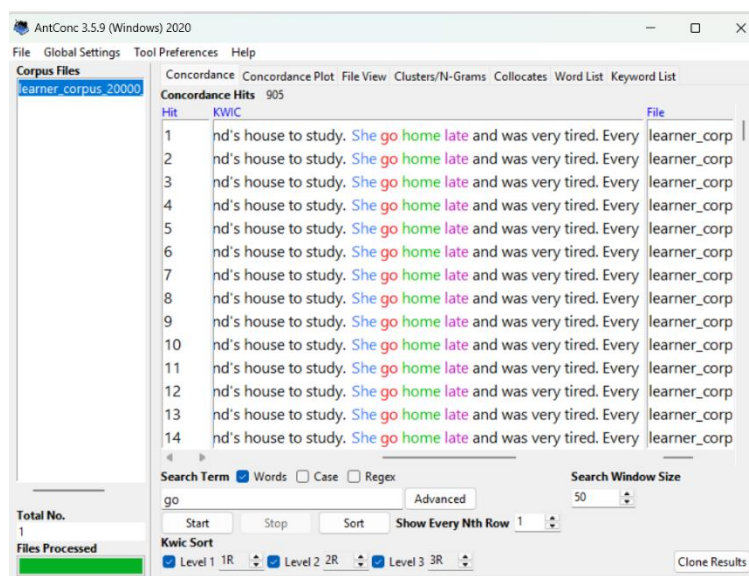

Figure 1: Sample AntConc Concordance Lines for "she go"

These lines demonstrate the persistence of tense-related errors in both past and habitual contexts. Although learners were exposed to formal grammar instruction, they appeared to lack full procedural control over irregular past forms, particularly with high-frequency verbs.

2. Article Errors (22.6%)

Article usage posed significant challenges. The omission of indefinite and definite articles was frequent, particularly before singular countable nouns:

- I saw a elephant. → I saw an elephant.

- She is teacher. → She is a teacher.

These errors align with known transfer issues from Uzbek and Russian, neither of which uses articles as grammatical features. Even advanced learners appeared to rely on article "guessing" rather than rule-based application, indicating incomplete acquisition.

3. Preposition Errors (16.7%)

Learners also struggled with English prepositions, which do not have direct equivalents in Uzbek. Common issues included incorrect pairings and omissions:

- I will go in Monday. → I will go on Monday.

- He arrived to the station. → He arrived at the station.

Literal translation from Uzbek or Russian prepositional systems likely contributed to these recurring patterns. Preposition use seemed largely formulaic and inconsistent, revealing a gap in productive use.

4. Subject-Verb Agreement (12.4%)

Errors in subject-verb agreement occurred primarily in the third person singular present tense:

- She like chocolate. → She likes chocolate.

- He go to school every day. → He goes to school every day.

Despite explicit instruction, these errors suggest developmental plateaus or fossilization. Learners often relied on base verb forms across all subjects.

5. Countability and Pluralization (11.2%)

This category included misuse of uncountable nouns in plural forms and incorrect pluralization of countable nouns:

- Advices instead of advice

• Many informations instead of much information

These issues point to a lack of awareness regarding noun types and a tendency to overgeneralize regular plural markers.

6. Word Order and Sentence Structure (8.6%)

Errors in sentence structure and word order included incorrect adverb placement and unnatural sentence openings:

• Happy I am to be here. → I am happy to be here.

• Always she studies in the evening. → She always studies in the evening.

These constructions reflect transfer from Uzbek syntax, where flexible word order is more acceptable. Learners often reproduced L1 sentence structures without full adjustment to English norms.

Table 1: Frequency of Grammatical Errors by Category

Error Type	Raw Frequency	Per 1,000 Words	Percentage (%)
Verb Tense/Aspect	183	9.15	28.5%
Article Usage	145	7.25	22.6%
Prepositions	107	5.35	16.7%
Subject-Verb Agreement	80	4.00	12.4%
Countability/Pluralization	72	3.60	11.2%
Word Order/Structure	55	2.75	8.6%
Total	**642**	**32.1**	**100%**

The data confirm that verb tense/aspect and article use are the most problematic areas for Uzbek learners at the B2 level. These error types are particularly susceptible to L1 transfer, fossilization, and incomplete grammatical acquisition. While the overall number of errors may seem typical for this level, the recurrence and consistency of certain error types highlight the need for targeted grammar interventions in Uzbek EFL classrooms, particularly those that foster deeper procedural knowledge and contextualized usage.

DISCUSSION

The findings from this study revealed several recurring grammatical patterns in the learner corpus, particularly concerning verb tense and aspect usage. The high frequency of errors such as "she go to school yesterday" or "they goes to the park" aligns with earlier observations in second language acquisition (SLA) research, which highlight verb tense and subject-verb agreement as persistent challenges among learners from non-English speaking backgrounds (Ellis, 2008; Zhang & Liu, 2021).

The misuse of past simple forms—particularly the failure to apply past tense morphology (-ed) to regular verbs—suggests that learners may be relying on interlanguage rules influenced by their L1 (Uzbek or Russian). This confirms previous claims that learners often simplify tense/aspect systems when these do not align with structures in their native language (Selinker, 1972; Ortega, 2013). In Uzbek, for instance, tense markers are suffix-based but differ in positioning and agreement patterns, which may explain the confusion in English tense construction.

Moreover, the AntConc concordance output clearly illustrated how certain error patterns cluster around high-frequency verbs such as go, like, arrive, and study. This reflects the notion of frequency-driven acquisition, where commonly used verbs become "learned" forms, but may fossilize in their incorrect versions without corrective feedback (Han, 2004; Ellis & Shintani, 2014). For example, the repeated occurrence of "she go" across different contexts in the corpus might indicate an overgeneralization or lack of internalized subject-verb agreement rules.

Another salient finding was the overuse of article-less noun phrases (e.g., "I saw a elephant", "she is teacher"), which supports previous findings by Ionin, Ko, and Wexler (2004) regarding L2 learners' difficulty with article use. This issue may be further exacerbated by the limited exposure to authentic English input in the Uzbek EFL context, where learners often rely heavily on test-preparation materials rather than natural discourse.

The study also affirms the practical usefulness of corpus tools like AntConc in visualizing and quantifying learner errors. This tool enabled the identification of recurring syntactic patterns and supported the hypothesis that learners' interlanguage is rule-governed but influenced by frequency, transfer, and developmental stages. The results have clear implications for syllabus designers and teachers in Uzbekistan: greater emphasis should be placed on tense/aspect teaching, especially through data-driven learning (DDL) techniques that allow learners to explore correct usage in real contexts (Johns, 1991; Boulton, 2017).

Ultimately, this study contributes to the growing field of

learner corpus research in Central Asia and opens pathways for further exploration of error types, proficiency levels, and longitudinal patterns of development among Uzbek EFL learners.

CONCLUSION

This study set out to investigate tense and aspect errors in the written English of Uzbek EFL learners through the lens of corpus linguistics, utilizing AntConc software to analyze a 20,000-word learner corpus. The analysis revealed consistent issues with past tense formation, subject-verb agreement, article usage, and lexical collocations—confirming many well-established findings in second language acquisition research.

The results not only support existing theoretical perspectives on interlanguage and L1 transfer but also highlight the importance of using corpus tools for identifying learner-specific difficulties. The ability to pinpoint frequent error types using concordance lines provides teachers and researchers with a more nuanced understanding of learner language, particularly in underrepresented contexts like Uzbekistan.

In pedagogical terms, this study underscores the value of data-driven learning approaches and corpus-informed teaching practices. Teachers in EFL contexts are encouraged to integrate corpus tools like AntConc into both lesson planning and classroom instruction to promote learner awareness and support evidence-based grammar instruction.

By combining corpus linguistics and second language pedagogy, this research contributes to the growing field of learner corpus analysis and emphasizes the urgent need for more locally grounded, technology-supported approaches in language education across Central Asia.

REFERENCES

1. Abduazizov, S., & Umarova, D. (2022). Challenges in EFL Writing among Uzbek Learners: A Classroom-Based Perspective. Uzbekistan Journal of Foreign Language Teaching.

2. Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2), 348–393. https://doi.org/10.1111/lang.12224

3. Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics, 5(4), 161–170.

4. Ellis, R. (2008). The study of second language acquisition (2nd ed.). Oxford University Press.

5. Ellis, R., & Shintani, N. (2014). Exploring language pedagogy through second language acquisition research. Routledge.

6. Granger, S., Gilquin, G., & Meunier, F. (2015). The Cambridge Handbook of Learner Corpus Research. Cambridge University Press.

7. Han, Z. (2004). Fossilization in adult second language acquisition. Multilingual Matters.

8. Hasanova, D. (2020). Error Analysis in the English Writings of Uzbek University Students. Journal of Applied Linguistics and Language Research, 7(4), 102–115.

9. Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. Language Acquisition, 12(1), 3–69. https://doi.org/10.1207/s15327817la1201_2

10. Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. ELR Journal, 4, 1–16.

11. Muminov, A. (2021). The Impact of L1 on English Article Use by Uzbek EFL Learners. International Journal of English Linguistics, 11(6), 89–98.

12. Nazarova, S., & Kadyrova, M. (2023). Investigating Verb Tense Errors in the Writing of Uzbek EFL Learners. Journal of Language and Education, 9(2), 67–76.

13. Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. ReCALL, 16(2), 279–294.

14. Odlin, T. (1989). Language Transfer: Cross-Linguistic Influence in Language Learning. Cambridge University Press.

15. Ortega, L. (2013). Understanding second language acquisition. Routledge.

16. Römer, U. (2022). Learner corpus research: Past, present, and future. Language Teaching, 55(2), 133–147.

17. Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching, 10(3), 209–231. https://doi.org/10.1515/iral.1972.10.1-4.209

18. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

19. Zhang, Y., & Liu, J. (2021). Persistent grammatical errors in L2 English: A corpus study. Journal of Second Language Writing, 54, 101785. https://doi.org/10.1016/j.jslw.2021.101785