INTERNATIONAL MULTI DISCIPLINARY JOURNAL FOR RESEARCH & DEVELOPMENT
eISSN 2394-6334 Impact factor: 8.854 Volume 13, issue 01 (2026)

A CORPUS-BASED ERROR ANALYSIS OF WRITTEN ESSAYS BY UZBEK ENGLISH LEARNERS

Aminjonov Avazjon

PhD, Senior Lecturer at Kokand University.
Email: aaaminjonov@kokanduni.uz

ORCID: 0009-0000-2138-9842

Abstract

This study reports a corpus-based error analysis (EA) of 25 short written essays produced by Uzbek learners of English as a foreign language (EFL) under a CEFR-aligned Task 1 prompt (functional writing). Following the EA procedure described in Kutlimuratova's (2021) corpus-based methodology, all identifiable errors were classified and quantified using a synthesized 13-tag scheme adapted from Divsar and Heydari's (2017) IELTS learner-corpus coding model. The scheme is theoretically grounded in (i) Chuang and Nesi's (2006) hierarchical error coding, (ii) Dagneaux et al.'s (1998) computer-aided error analysis approach, and (iii) Hou's (2016) ten-category system for learner writing. Descriptive statistics show 211 errors in total (M = 8.44 errors per essay, range = 1–16). Verb-related errors (V) were the most frequent (n = 40, 19.0%), followed by article errors (A) and sentence structure errors (SS) (n = 23 each, 10.9%). Spelling (S) and deletion (D) errors were also prominent (n = 19 each, 9.0%), and confusing/unclear statements (CU) accounted for 18 instances (8.5%). Findings are discussed in relation to Uzbek learners' developing interlanguage and the genre demands of short functional writing, and pedagogical implications are proposed for form-focused instruction, corpus-informed materials, and targeted feedback.

Keywords

error analysis; learner corpus; Uzbek EFL learners; CEFR writing; grammatical accuracy; corpus-based pedagogy.

Introduction

Second language (L2) writing remains one of the most demanding skills for EFL learners because it requires the simultaneous coordination of lexis, grammar, cohesion, and genre conventions. In CEFR-aligned functional writing tasks (e.g., short letters or emails), the communicative goal is often clear and the expected length is modest; however, accuracy and clarity still shape the perceived quality of the message. For Uzbek learners, whose first language differs typologically from English in morphology, article systems, and certain syntactic patterns, written production frequently reveals systematic deviations that can be examined as evidence of developing interlanguage.

Error Analysis (EA) offers a principled way to identify, classify, and quantify such deviations. Since Corder's seminal work on the significance of learner errors, EA has been widely used to describe recurring problems and to generate pedagogically actionable profiles of learner needs. In the context of learner corpus research, EA can be strengthened by adopting consistent tagsets that make patterns visible across multiple texts and allow comparisons across studies.

Recent corpus-based EA studies have proposed increasingly detailed error taxonomies and annotation procedures. Dagneaux et al. (1998) introduced computer-aided error analysis (CEA) as a method for compiling and exploring error-tagged learner corpora. Chuang and Nesi (2006) further demonstrated how hierarchical coding can capture both broad and fine-grained error categories in academic learner writing. Hou (2016), working with Taiwanese EFL students, employed a ten-category system focusing on frequent morphosyntactic and lexical errors in academic writing. Building on these approaches, Divsar and Heydari (2017) proposed a 13-aspect coding scheme for IELTS essays and demonstrated how frequency-based profiles can inform instruction.

In Uzbekistan, Kutlimuratova (2021) provided a corpus-based EA model specifically targeting Uzbek EFL learners' writing and operationalized a 13-tag set aligned with Divsar and Heydari's scheme. Her study illustrated the value of combining corpus tools with descriptive statistics to identify the most frequent error types and to link them to plausible sources (e.g., limited grammatical control, lexical gaps, or task-driven constraints).

The present study adopts Kutlimuratova's (2021) methodological logic of systematic identification, coding, and quantification of errors in learner writing while applying it to a new dataset: a corpus of 25 essays written by Uzbek EFL learners for a CEFR Task 1 prompt. To ensure comparability and pedagogical interpretability, the study uses a synthesized tagging system of 13 categories inspired by Divsar and Heydari (2017) and aligned with the categories operationalized in Kutlimuratova (2021).

This paper addresses the following research question: What are the most frequent error categories in Uzbek EFL learners' CEFR Task 1 essays, as captured by a corpus-based 13-tag EA scheme, and what pedagogical implications follow from the observed distribution? To answer this question, the study reports descriptive results and discusses them against prior learner-corpus findings, with illustrative examples (Appendix A) drawn from the dataset's genre context.

Literature Review

EA has evolved from early descriptive accounts of learner deviations to corpus-informed, systematic methodologies that treat errors as data about learners' developing linguistic system. In learner corpus research, errors are typically operationalized as departures from target-like usage in context, and they are annotated with tags that enable frequency analysis and pattern discovery. Dagneaux et al. (1998) argued that computer-aided procedures can greatly increase the efficiency and reproducibility of error analysis, especially when the goal is to obtain error lists and frequencies across a corpus.

One challenge in corpus-based EA is balancing granularity with usability: highly detailed taxonomies can capture nuance but may reduce reliability and increase annotation time, whereas broad categories may obscure patterns important for pedagogy. Chuang and Nesi's (2006) hierarchical coding approach addresses this tension by coding errors at multiple levels, broad 'parent' categories with more specific subcategories, allowing researchers to analyze both macro-patterns and finer distinctions. Similarly, Hou (2016) used a focused ten-category system that targets frequent, instructionally salient errors (e.g., articles, prepositions, verb forms), enabling a clear mapping from findings to teaching priorities.

Divsar and Heydari (2017) developed a 13-aspect coding scheme for IELTS essays and reported that word choice and verb-form errors were particularly frequent in their dataset. Their study is notable for linking quantitative profiles to teaching implications, including the design of materials aligned with observed weaknesses. Drawing on the same family of categories, Kutlimuratova (2021) examined Uzbek EFL learners' essays in a corpus-based design and reported that article errors were the most frequent in her corpus, while linking/connecting word errors and sentence structure errors occurred relatively rarely. These findings suggest that the relative prominence of error categories may vary by prompt type, learner cohort, proficiency distribution, and annotation decisions.

Although Uzbek EFL learner writing has begun to receive attention through corpus-based EA (e.g., Kutlimuratova, 2021), more datasets are needed to establish whether observed error profiles are stable across tasks and genres. CEFR Task 1 functional writing (short letters/emails) places distinct demands on learners, including pragmatic clarity, conventional openings/closings, and concise complaint or request structures. By analyzing 25 Task 1 essays with a comparable 13-tag scheme, the present study contributes (i) a task-specific error frequency profile for Uzbek EFL learners, (ii) a comparison point to prior Uzbek and IELTS-based corpus studies, and (iii) an example set illustrating how common errors map to the tagging system used in analysis.

Methodology

The dataset comprises 25 essays written by pre-intermediate Uzbek EFL learners as part of a CEFR-aligned Task 1 writing activity. The texts represent short functional writing (e.g., letter/email style) and were treated as a small learner corpus for error profiling. Because the present paper focuses on the linguistic error distribution rather than individual learner trajectories, essays are anonymized and referenced only by numeric IDs (1–25) in the results tables.

The study follows a corpus-based EA methodology in the spirit of Kutlimuratova (2021), which operationalizes EA as systematic identification, classification, and frequency analysis of learner errors in a corpus of essays. At the level of coding philosophy, the approach aligns with Dagneaux et al.'s (1998) CEA orientation (consistent tagging enabling automated counting) and is compatible with Chuang and Nesi's (2006) hierarchical view of error categories (macro-categories that can be subdivided where needed). In addition, Hou's (2016) ten-category emphasis on frequent morphosyntactic error types informed the selection and interpretation of the most salient categories in the dataset.

Errors were coded using a synthesized set of 13 tags inspired by Divsar and Heydari (2017) and consistent with the list described in Kutlimuratova (2021). The tags used in the present study are: S (Spelling), Pun (Punctuation), A (Article), WC (Word Choice), I (Insertion), D (Deletion), N (Noun: singular/plural), V (Verb: tense/inflection), P (Preposition), WF (Word Form), SS (Sentence Structure), CU (Confusing/Unclear), and O (Linking/Connecting). Operational definitions follow the 'surface' manifestation of the error in context (e.g., missing article = A; wrong tense/inflection = V; missing preposition = P), while allowing multiple tags for a single segment when errors co-occur (e.g., a sentence may be both structurally ill-formed and contain a missing function word).

Each essay was inspected for departures from target-like usage. Identified errors were assigned one or more tags from the 13-category scheme. Counts were aggregated per essay and per error type, and the resulting frequency table was exported to a spreadsheet for descriptive analysis. Following Kutlimuratova (2021), the study reports totals, percentages, and basic

descriptive statistics across essays. Because the available dataset did not include word counts per essay, error rates are reported as raw counts rather than normalized per 100 words. The analysis is therefore best interpreted as a profile of relative error distribution within the corpus rather than as a precise measure of error density across differently sized texts. The corpus was anonymized, and no identifying information about learners is reported. Examples in Appendix A are presented only to illustrate the tag categories and do not disclose personal data.

Results

The corpus of 25 CEFR Task 1 essays contained 211 coded errors in total. At the individual-text level, learners produced an average of 8.44 errors per essay (SD = 4.51), with a median of 8. Error totals ranged from 1 to 16 per essay, indicating substantial variability in accuracy across writers. When essays were grouped by overall error density, 6 texts fell in the low-error band (1–5 errors), 9 texts in the mid band (6–10 errors), and 10 texts in the high-error band (11–16 errors). This spread suggests that within the same instructional context, some learners can meet the communicative demands of functional writing with comparatively few form-based problems, whereas others struggle to maintain grammatical and discourse control even in short texts.

Across the 13-category tagset, errors clustered most strongly in morphosyntax. Grammatical categories (V, A, SS, P, D, I, N, WF, and O combined) accounted for 147 instances (69.7% of all errors), whereas orthographic accuracy (S and Pun) contributed 32 cases (15.2%). Lexico-semantic choice (WC) represented 14 cases (6.6%), and meaning-level breakdowns coded as CU occurred 18 times (8.5%).

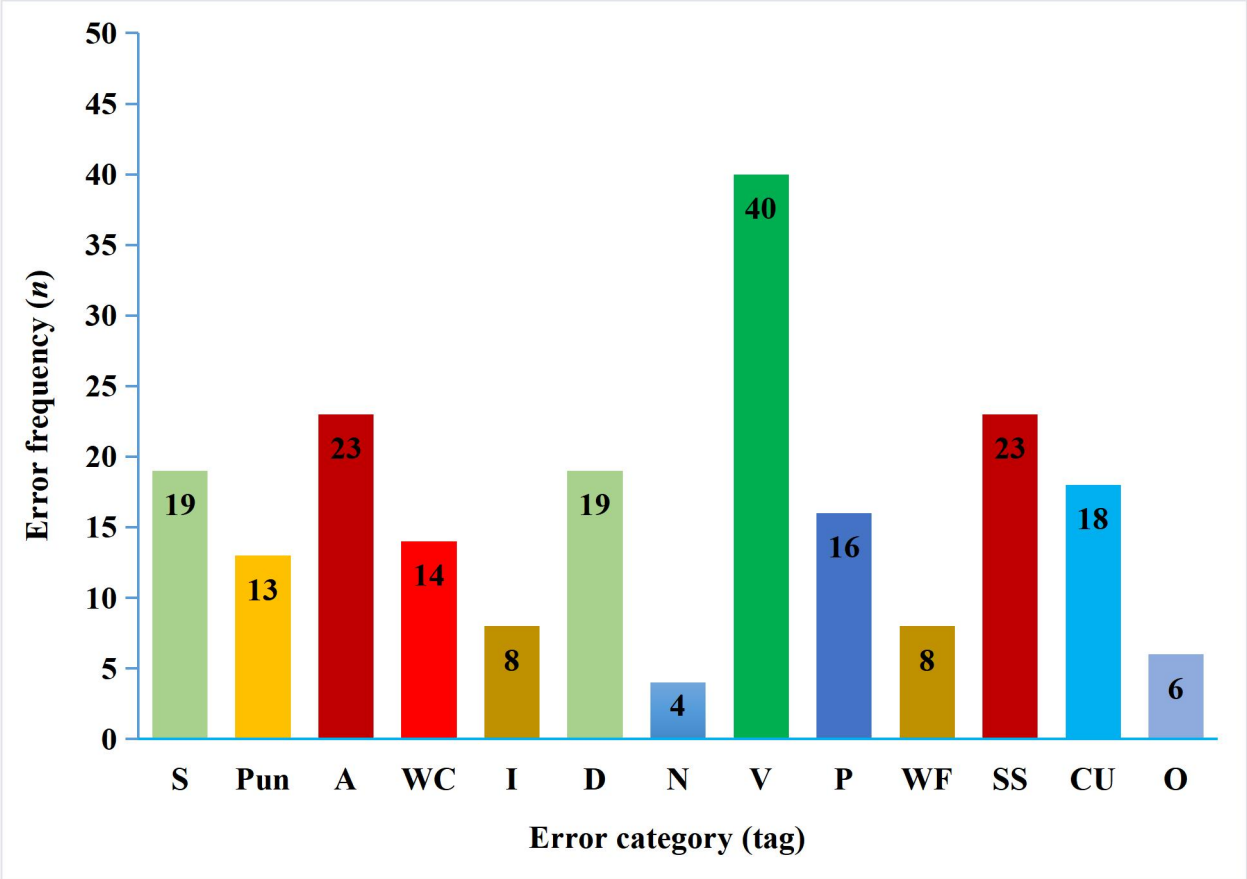

Figure 1. Frequency of errors by category (tag)

As shown in Figure 1, verb-related problems (V) were the single largest category (n = 40, 19.0%), followed by article errors (A) and sentence structure errors (SS) (n = 23 and 23, 10.9% and 10.9%, respectively). Spelling (S) and deletion/omission (D) were also frequent (n = 19 and 19, 9.0% each). Confusing or unclear statements (CU) made up 18 cases (8.5%), and preposition errors (P) accounted for 16 cases (7.6%). Lower-frequency categories included word-form errors (WF) and insertion errors (I) (n = 8 and 8, 3.8% and 3.8%), with noun-number errors (N) being rare (n = 4, 1.9%). Linking/connecting issues (O) occurred 6 times (2.8%). Looking beyond totals, the breadth of categories across texts also matters for pedagogy. Verb errors appeared in 19 of the 25 essays, while SS and D errors were present in 15 and 14 essays, respectively. In contrast, noun-number problems (N) appeared in only 4 essays. This pattern suggests that, for many writers, accuracy breakdowns are not isolated slips but recurring difficulties that spread across multiple grammatical subsystems.

Because the current dataset is functional writing (complaints, requests, and service evaluations), certain linguistic moves recur: describing past experiences, evaluating quality, and requesting action. The most frequent error categories map directly onto these moves. Below, representative excerpts illustrate how each tag was operationalized in context. In several cases, one sentence legitimately receives multiple tags, reflecting the fact that form problems often cluster rather than occurring in isolation.

a) Verb errors (V) commonly involved tense control, auxiliary use, and negation in infinitival constructions. For instance, “I'm trying very hard to don't get angry” combines non-target negation and an incorrect infinitival complement (V, SS). Similarly, “Also, restaurant service quite slow” omits the required copular verb (“is”), yielding a structurally incomplete clause (V, D, SS).

b) Article errors (A) were typically omissions in environments requiring an indefinite or definite determiner. Examples include “I have to stay hotel that is near the airport” (A, P) and “Today I took email from the manager of that hotel” (A). Such omissions reduce the formality of the message and can make reference tracking harder for the reader.

c) Sentence-structure problems (SS) often took the form of fragments, unstable word order, or ill-formed relative clauses. “In a nutshell, very bad” is a fragment lacking a subject and finite verb (SS, D, CU), while “Did you take a letter from hotel’s manager too which we stayed?” shows relative clause misformation and awkward embedding (SS, P, A, CU).

d) Spelling errors (S) were frequent enough to be pedagogically relevant, even though they were not the dominant category. Typical examples include “plasant” for pleasant and “condititioning” for conditioning. In a few cases, the spelling choice resulted in a different real-word form (e.g., “vocation” for vacation), which can be more disruptive than a simple typo because it misleads the reader.

e) Deletion errors (D) captured missing obligatory items such as prepositions or auxiliaries. For example, “And I’m looking forward you to meet me again soon” is missing the preposition “to” after look forward, and it also mis-selects the complement form (P, V, D). Similarly, “Next Monday Would you go new hotel?” lacks both the comma after the time adverbial and the preposition “to” in “go to a hotel” (Pun, V, P, A).

f) Word-choice errors (WC) reflected non-target collocations and pragmatic inappropriateness in complaint writing. “First of all, it is not very bad Wifi and unclear kitchen” includes a vague and unnatural collocation (“unclear kitchen”), which is better expressed with evaluative adjectives such as dirty or poorly equipped (WC, A, SS, CU). Another example, “exchange the staffs,” illustrates lexical mis-selection alongside a noun-form issue (WC, N).

g) Word-form errors (WF) were often derivational: an adjective was used where a noun was needed or vice versa. For instance, “safe instead of safety” and “I didn't satisfaction breakfast” illustrate incorrect derivational choice and, in the latter case, a missing copular structure (“I wasn’t satisfied”) (WF, V, SS).

h) Linking/connecting problems (O) were relatively infrequent in raw counts, but they were consequential when they occurred because they affected coherence. “Because hotel is massive mistake service” contains a subordinator without a main clause and an ill-formed predicate, producing a coherence breakdown (O, SS, A, CU).

Taken together, the results show that the strongest constraints on communicative effectiveness in these essays are not limited to surface mechanics. Instead, the majority of errors involve the grammar needed to narrate events and formulate polite requests, which aligns with the functional demands of CEFR Task 1 writing.

Discussion

The most salient outcome of the corpus analysis is the predominance of verb-related errors. In functional writing, learners must manage time reference (typically recounting past experiences), evaluation (often using copular structures), and interpersonal stance (requests, suggestions, and advice). Each of these communicative moves relies heavily on the English verbal system including tense/aspect choices, auxiliary support, and appropriate complementation patterns. Consequently, small disruptions in verb form quickly cascade into broader sentence-level problems. Several recurring V patterns in the corpus resemble well-attested developmental and L1-influenced difficulties. First, learners frequently omit the copular verb in evaluative clauses (e.g., “restaurant service quite slow”), a pattern that is likely encouraged by the fact that Uzbek does not require an overt copula in the present tense in the same way English does. Second, learners show instability in negative and interrogative constructions that require auxiliary do-support in English (e.g., “I’m trying very hard to don’t get angry”; “Next Monday Would you go new hotel?”). Such examples suggest that learners may possess the intended meaning and even know individual word forms, but they have not fully proceduralized the morphosyntactic routines needed for fluent sentence assembly under time pressure.

From a pedagogical standpoint, this profile supports prioritizing verb patterns as high-yield targets. Rather than treating verb errors as a single broad category, instruction can focus on the small set of constructions most frequently required by CEFR Task 1: past-tense narration (I stayed, I received), copular evaluation (The service was…, The room was…), and request frames (Could you…, I would suggest…, Please…). Short, repeated practice with these frames, combined with feedback that explicitly links each error to the communicative function it disrupts, can plausibly reduce both V errors and downstream SS and CU errors.

Article errors were among the most frequent categories and are consistent with the broader SLA literature on learners whose L1 lacks an article system. In the corpus, article omissions

were common in precisely the environments that matter for reader interpretation: introducing new referents ("a hotel"), maintaining shared reference ("the hotel"), and using fixed expressions ("in good health"). When articles are missing, the text remains generally interpretable, but it often sounds less formal and can force the reader to infer whether a noun is being introduced or referred to again.

The presence of noun-number (N) and countability-related problems, although small in raw frequency, reinforces this point. Examples such as "a water" (countability) and "guest" for "guests" (plural marking) indicate that determiner choice and noun morphology interact in learner production. Targeted teaching that treats determiners and countability as a connected system, rather than separate grammar points, may be more effective for Uzbek EFL learners, especially in functional writing where words like *staff, accommodation, luggage,* and *information* appear frequently.

Sentence structure errors and deletions together account for a substantial share of the corpus. Importantly, these categories often overlap with CU coding: when a sentence loses a finite verb, a required preposition, or a coherent clause link, meaning can become difficult to reconstruct. Fragments such as "In a nutshell, very bad" are understandable in informal speech, but in formal letters, they reduce rhetorical completeness because the writer does not explicitly state what is being evaluated. Complex sentences were a particular pressure point. Learners sometimes attempted relative clauses or embedded structures to increase formality ("…the hotel … which we stayed"), but the resulting clause integration was unstable. This suggests that part of the SS profile may be an unintended consequence of learners trying to 'sound academic' or 'sound formal' without full grammatical control. A practical implication is to teach accuracy-first alternatives: shorter clauses, clear subject–verb structure, and simple but correct connectors (because, so, therefore) before encouraging elaboration through embedding.

Although lexical and word-form categories are smaller in frequency than grammar-based categories, they carry disproportionate pragmatic weight in complaint letters. Non-target collocations such as "unclear kitchen" or "exchange the staff" can make a complaint sound vague, overly harsh, or simply unnatural. Similarly, derivational errors (safe/safety; satisfaction/satisfied) often occur in precisely the formulaic politeness moves that CEFR Task 1 expects writers to master. Prepositions sit between grammar and vocabulary, and they appeared repeatedly across many of the essays. The patterns we saw, such as leaving out 'to' in movement phrases ("go to a hotel") and using the wrong preposition after common adjectives or verbs ("satisfied with", "look forward to"), suggest that learners may be learning these multiword expressions in an incomplete way. Teaching implications follow naturally: instruction should be organized around high-frequency chunks rather than isolated preposition lists, and feedback can use the P tag to direct learners to revise the entire phrase (look forward to + V-ing) instead of changing a single word in isolation.

In Kutlimuratova's (2021) corpus-based study of Uzbek learners' IELTS writing, the most common error types reported were spelling, article, punctuation, and word choice, whereas sentence structure, linking word, and confusing-expression errors were rare. The present CEFR Task 1 corpus partially converges with that pattern in that article and spelling problems are frequent and linking issues remain comparatively infrequent. However, the current dataset differs in the prominence of verb-related errors and in the substantial share of sentence-structure and omission-related problems. One plausible explanation is genre and task demand. IELTS academic essays often reward lexical sophistication and careful editing, which may foreground orthographic and lexico-grammatical choices. By contrast, CEFR Task 1 functional writing

places heavier, immediate demands on routine interpersonal grammar: narrating a service experience, describing what went wrong, and requesting remediation politely. These moves naturally activate tense control, auxiliary structures, and fixed request frames, which may explain why V and SS errors are especially visible here. Another explanation concerns proficiency distribution: the spread from 1 to 16 errors per essay indicates that the corpus likely includes writers at noticeably different levels of procedural control, which can magnify grammar-based differences in a small dataset.

Finally, the synthesized tagset used in this study is designed for pedagogical interpretability, drawing on hierarchical and corpus-based coding traditions (e.g., Chuang & Nesi, 2006; Dagneaux et al., 1998; Hou, 2016) while remaining compact enough for classroom feedback. The results demonstrate that such a tagset can highlight both high-frequency grammar targets (V, A, SS, P) and lower-frequency but high-impact problems (CU, O), supporting focused remedial instruction without overwhelming teachers or learners with overly granular labels.

Pedagogical implications for Uzbek EFL writing instruction. The clearest instructional implication is prioritization. If teachers must choose a small number of targets with the greatest payoff for communicative success in CEFR Task 1 writing, the present corpus indicates that verb patterns, article use, and sentence completeness should come first. In practice, this means (a) routine drilling of past narration and copular evaluation frames, (b) explicit work on a/the in common service contexts (a hotel, the manager, the service), and (c) revision routines that check for a subject + finite verb in every sentence. At the same time, surface accuracy should not be treated as independent of discourse. Many clarity problems arise because learners attempt complex clause linkage without stable sentence control. A staged approach is therefore recommended: build a bank of short, correct complaint/request sentences, then teach combining strategies (because, so, however) with clear punctuation and paragraphing. Using the 13 tags as feedback shorthand (e.g., ‘V’ for verb form, ‘A’ for article) can make this staged approach transparent to learners and can support longitudinal tracking of improvement over repeated writing tasks.

Limitations and directions for further research. Two limitations should be acknowledged. First, the corpus is relatively small (25 essays), and it is tied to a single task type; larger and more diverse corpora would allow more robust generalization. Second, the current dataset summarizes error counts by category, which is appropriate for frequency profiling but does not directly model co-occurrence patterns at the level of individual error instances. Future research could extend the analysis by adding inter-rater reliability reporting for the tagging scheme and by exploring how particular constructions (e.g., request frames) concentrate specific error types over time.

Conclusion

This study applied a corpus-based EA approach, following Kutlimuratova’s (2021) methodological logic and a synthesized 13-tag coding scheme inspired by Divsar and Heydari (2017), to profile errors in 25 CEFR Task 1 essays written by Uzbek EFL learners. Across the corpus, 211 errors were identified, with verb-related errors (V) emerging as the most frequent category, followed by article (A) and sentence structure (SS) errors. Spelling (S), deletion (D), and confusing/unclear statements (CU) were also substantial contributors to the overall error load.

The distribution suggests that instruction for Uzbek EFL writers in functional genres should emphasize (i) verb tense/inflection control in high-frequency communicative functions, (ii) article use as part of noun-phrase construction, and (iii) sentence completeness and word order through genre-based frames and guided rewriting. Using a concise tag set in classroom feedback may help learners recognize recurring patterns across drafts and build a more accurate interlanguage system.

Several limitations should be noted. The corpus is small and represents a single task genre; results therefore should not be generalized to all Uzbek EFL writing contexts without further evidence. Because word counts were not available, errors were not normalized per text length. Additionally, the present report is based on the aggregated frequency table rather than full error-annotated text, limiting deeper qualitative analyses of error sources. Future research should expand corpus size, include multiple prompts and proficiency bands, and report inter-rater agreement for coding. Combining error profiles with learner interviews or classroom observations could also clarify how instruction, exposure, and L1–L2 contrasts shape the observed error patterns.

References

1. Chuang, F.-Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. Corpora, 1(2), 251–271.
2. Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, 26(2), 163–174.
3. Divsar, H., & Heydari, R. (2017). A corpus-based study of EFL learners' errors in IELTS essay writing. *International Journal of Applied Linguistics & English Literature*, 6(3), 143–149.
4. Hou, H. I. (2016). Learner corpus and academic writing: Identifying the error patterns of Taiwanese EFL students. *Journal for the Study of English Linguistics*, 4(1), 19–30.
5. Kutlimuratova, B. (2021). Uzbek students learning English as a foreign language: Error analysis using corpora (Master's thesis, Universidade da Coruña).