

Corpus Based Error Analysis of Article Usage in Sri Lankan ESL Learners' Writing

Dilhan Rajapakse¹

Abstract

English functions as a second language in Sri Lanka, where it is crucial for higher education, professional advancement, and global communication. Despite long-term exposure, Sri Lankan English as a Second Language (ESL) learners continue to produce recurrent grammatical errors, particularly in the use of the English article system (a, an, the). Previous research has highlighted article errors as one of the most persistent difficulties for L2 learners, especially for those whose first languages (Sinhala and Tamil) lack a comparable system of definiteness. This study adopts a corpus-based approach to investigate article errors among Sri Lankan English learners as a second language. A learner corpus of 25 answer scripts (approx. 20,000 words) written in response to a literature question on William Blake's A

Poison Tree was compiled and analyzed using AntConc 3.5.9. Article errors were annotated into three categories: omission, overuse, and wrong choice. Quantitative analysis revealed that "the" was disproportionately overused, often before proper nouns and literary titles (e.g., "The William Blake"), while "a/an" was frequently omitted before singular countable nouns. Wrong choice errors included misapplication of a before vowel sounds. Qualitative concordance evidence further indicated that learners memorized and misapplied article noun clusters. These findings align with global research that highlights the complexity of articles in SLA and the influence of L1 transfer, overgeneralization, and fossilization. The study advocates for the development of a Sri Lankan learner corpus to enable systematic research and targeted interventions in ESL pedagogy.

Keywords: AntConc, articles, ESL, error analysis, interlanguage, learner corpus

1. Introduction

1.1 English in the Sri Lankan Context

English occupies a unique position in Sri Lanka as both a second language (ESL) and a link language. While Sinhala and Tamil function as the official national languages, and English continues to play a central role in education, commerce,

and administration. Its importance is particularly pronounced in higher education, where English serves as the medium of instruction for many academic disciplines, including the sciences, and professional fields such as medicine and law (Liyanage, 2023). Moreover, proficiency in English is strongly associated with socioeconomic

¹ Instructor in English, Department of Languages, Bhiksu University of Sri Lanka.

E-mail: dilhan@busl.ac.lk

mobility, access to white-collar employment, and participation in globalized networks of communication.

Despite its institutional prominence, English remains a foreign language for the majority of Sri Lankans, acquired in school settings rather than as a natural medium of communication. This learning environment often leads to incomplete mastery, with learners developing an interlanguage that reflects both transfer from their first languages (Sinhala and Tamil) and strategies of simplification. As a result, persistent grammatical errors are a common feature of Sri Lankan ESL learners (Meyler, 2007). Hence, the research is conducted involving Diploma in English (external) students, the Bhiksu University of Sri Lanka.

1.2 Persistent Grammatical Challenges in ESL

Among the many grammatical difficulties faced by ESL learners in Sri Lanka, three areas stand out: prepositions, articles, and morphemes. Wijeratne and Jeyaseelan (2015) observed that learners frequently misapplied morphemes when presenting entities and events, while Abeywickrama (2010) identified recurrent article overgeneralization, such as “I went (the) home and spoke to (the) Perera”. These issues are not merely stylistic but signal deep structural challenges in mastering the morphosyntactic system of English.

Articles, in particular, represent one of the most complex areas of English grammar for second language learners worldwide (Master, 1997; Ionin, Ko &

Wexler, 2004). Unlike prepositions or verb tenses, articles do not always carry overt meaning, making them difficult for learners to acquire. Furthermore, because neither Sinhala nor Tamil has an equivalent article system, Sri Lankan learners lack a direct L1 reference point. This results in frequent omission, overgeneralization of “the,” and confusion between “a” and “an”.

1.3 Error Analysis in Second Language Acquisition

Error analysis (EA) provides a systematic method for studying these difficulties. Pioneered in the 1960s and 1970s (Corder, 1967; Dulay, Burt & Krashen, 1982), EA focuses on learner errors as evidence of language development rather than mere deficiencies. Errors reveal underlying interlanguage systems and developmental stages. In the Sri Lankan context, EA has proven valuable in highlighting the persistent grammatical difficulties that remain despite years of formal instruction (Abeywickrama, 2010; Wijeratne & Jeyaseelan, 2015).

1.4 Corpus Linguistics and Error Analysis

While traditional EA studies relied on small samples and manual coding, the advent of corpus linguistics has revolutionized error research. Learner corpora; systematic collections of learner-produced texts, enable large-scale, empirical investigation of error patterns (Granger, 2008; Gilquin, 2020). Tools such as AntConc (Anthony, 2022) provide powerful functionalities, including concordance searches, frequency counts, collocation analysis,

and keyword comparison. These methods allow researchers to not only detect errors but also to quantify them, analyze their distribution, and identify recurring collocational patterns.

Globally, learner corpora such as the International Corpus of Learner English (ICLE) and the International Corpus Network of Asian Learners of English (ICNALE) have yielded valuable insights into article usage, tense/aspect marking, and prepositional errors. However, Sri Lanka lacks a comparable learner corpus. Existing corpora, such as the International Corpus of English, Sri Lanka (ICESL), the Sri Lanka English Newspaper Corpus (SLENC), and SAVE 2020, primarily capture institutional or professional English (Jayaweera & Dias, 2014), not learner interlanguage.

1.5 Research Gap

The lack of Sri Lankan learner corpora can be observed as a gap that has significant consequences for both pedagogy and research. Without systematic learner corpora, Sri Lankan ESL educators rely on anecdotal observations rather than empirical data when addressing learner difficulties. Similarly, SLA research in Sri Lanka remains underdeveloped compared to global trends. Article usage, in particular, has not been examined corpus-based in Sri Lankan contexts, despite being one of the most persistent sources of learner error.

This study aims to address this gap by conducting a corpus-based error analysis of article usage among students

of English in Diploma level.

2. Literature Review

2.1 Error Analysis and its Relevance in SLA

Error Analysis (EA) has long been recognized as a central tool in the study of second language acquisition (SLA). Pioneering work by Corder (1967) argued that learner errors are not “bad habits” to be eradicated but rather evidence of the learner’s interlanguage system, the transitional competence that reflects underlying strategies of language learning. Dulay, Burt, and Krashen (1982) later emphasized that errors are natural indicators of developmental stages and provide researchers with crucial insight into the mechanisms of L2 learning.

EA involves identifying, categorizing, and analyzing learner errors with the purpose of explaining their causes and proposing pedagogical interventions. Richards et al. (1992) suggest that EA is valuable not only for descriptive linguistics but also for applied purposes, including syllabus design, remedial teaching, and material development. Erdogan (2005) further highlighted that EA helps teachers to predict problem areas and adapt instruction accordingly.

In the Sri Lankan ESL context, error analysis has been especially useful in identifying recurring difficulties across generations of learners. Abeywickrama (2010) observed that article omission, prepositional misuse, and verb tense confusion remain persistent even at university level, reflecting fossilization

of errors due to insufficient input and inadequate instructional focus. Similarly, Wijeratne and Jeyaseelan (2015) identified morpheme misapplications, suggesting that learners struggle with bound inflectional morphology.

Together, these studies show that EA remains relevant in Sri Lanka for both diagnostic and prescriptive purposes. By revealing learner difficulties systematically, EA provides a foundation for data-driven pedagogy.

2.2 The English Article System

Among all areas of grammar, the English article system has been repeatedly identified as one of the most challenging for L2 learners (Master, 1997; Ionin, Ko & Wexler, 2004). English articles, a, an, the, and zero article, are function words that encode definiteness, specificity, and countability. Their correct use depends on complex semantic and pragmatic distinctions rather than purely syntactic rules.

2.2.1 Definite Article (the)

The definite article signals that a referent is identifiable to both speaker and listener (Lyons, 1999). For example, the book implies that the interlocutors share knowledge of the specific referent. Learners often overgeneralize the, using it even with proper nouns (e.g., the William Blake), or with generic nouns where no article is required (e.g., the nature is beautiful).

2.2.2 Indefinite Articles (a/an)

Indefinite articles mark a singular, countable noun as nonspecific. The

choice between a and an depends on phonology, not spelling, which adds another layer of complexity (Huddleston & Pullum, 2002). Learners frequently omit indefinite articles altogether (e.g., He is teacher), or misuse them before uncountable nouns (e.g., a knowledge). Wrong choice errors are also common, such as a English poet instead of an English poet.

2.2.3 Zero Article

English requires no article before plural and uncountable nouns in generic reference (Swan, 2016). However, many learners fail to differentiate zero article contexts, inserting unnecessary articles (e.g., the books are useful when referring to books in general).

2.2.4 Complexity in Acquisition

Research has consistently shown that article acquisition is influenced by both universal principles and L1 transfer. Master (1997) proposed that learners progress through predictable stages, beginning with article omission and moving toward overgeneralization of the. Ionin et al. (2004) argued that learners rely on semantic notions of definiteness and specificity, but may misinterpret these depending on their L1. Since Sinhala and Tamil lack an article system, Sri Lankan learners cannot rely on L1 transfer and instead must acquire article rules as entirely new grammatical features.

2.3 Global Research on Article Errors

A substantial body of research has examined article errors across learner

populations. Huebner (1983), studying Vietnamese learners of English, noted persistent overuse of the due to reliance on shared knowledge assumptions. Robertson (2000) found that Korean learners omitted indefinite articles in obligatory contexts, influenced by the absence of articles in Korean. Ionin, Ko, and Wexler (2004) demonstrated that Russian and Korean learners often confuse definiteness and specificity, leading to systematic misuses of a and the.

Across contexts, three major error categories recur:

1. Omission—learners leave out required articles (She bought book).
2. Overuse – learners insert unnecessary articles (The happiness is important).
3. Wrong choice – learners select the incorrect article (a English poet).

These error types appear to be universal, but their frequency and distribution vary by L1 background. Since Sinhala and Tamil lack articles altogether, Sri Lankan learners often rely on transfer strategies that result in omission and overgeneralization.

2.4 Corpus Linguistics and Error Analysis

Corpus linguistics has transformed the way researchers study learner errors. Kennedy (2001) and Biber et al. (1998) emphasize that corpora provide authentic, empirical evidence of language use, moving beyond intuitionbased descriptions. Corpus tools enable researchers to

examine frequency, distribution, and collocational patterns with precision.

2.4.1 Concordance and Frequency Analysis

Concordancers such as AntConc (Anthony, 2022) allow researchers to generate keyword in context (KWIC) displays, showing how learners use a particular form in varied contexts. This makes it possible to detect systematic patterns of error, such as repeated omission of a before singular count nouns. Frequency lists further highlight which words or forms are disproportionately used, revealing learner reliance on certain structures.

2.4.2 Error Annotated Learner Corpora

Annotation of corpora adds a structured layer of information, marking errors according to categories such as omission, overuse, or substitution. Tetreault and Chodorow (2008) argue that annotation is essential for developing “gold standard” data sets in computational linguistics. Leacock et al. (2010) highlight that annotated corpora form the basis for automated error detection systems. Granger (2008) and Gilquin (2020) emphasize that learner corpora specifically provide unique insights into interlanguage development, as they foreground systematic deviations from target norms.

2.4.3 Existing Learner Corpora

Internationally, several learner corpora have advanced SLA research:

- International Corpus of Learner

English (ICLE) – essays by advanced EFL learners across multiple L1s. (Granger, 2003)

- ICNALE (International Corpus Network of Asian Learners of English) – learner writing across Asian countries, useful for crosslinguistic comparison. (Ishikawa, 2013)
- Lang 8 Learner Corpus – learner output from language exchange platforms, often annotated.

These resources have yielded numerous studies on article acquisition, collocation, and lexical bundle use. However, none include Sri Lankan learners as a distinct category.

2.5 Error Analysis in the Sri Lankan ESL Context

Research on Sri Lankan learners has documented recurring errors, but usually through small scale, non-corpus-based studies. Abeywickrama (2010) noted omission of articles and misuse of prepositions in undergraduate writing. Wijeratne and Jeyaseelan (2015) highlighted morpheme misapplications in essays. More recently, Karunanayake et al. (2022) described Sri Lanka's limited corpus resources, such as the Sri Lanka English Newspaper Corpus (SLENC), the International Corpus of English, Sri Lanka (ICESL), and SAVE2020. These corpora, however, represent edited professional English rather than learner interlanguage.

Sri Lankan learners face unique challenges with articles because:

1. L1 Transfer: Sinhala and Tamil lack articles, leading to overreliance on zero article contexts.
2. Overgeneralization: Learners tend to overuse the in contexts of specificity, even when generic reference is intended.
3. Pedagogical Gaps: Grammar teaching in Sri Lanka often emphasizes rules in isolation rather than authentic usage.
4. Fossilization: Certain article errors persist across learner levels, becoming entrenched in interlanguage.

Despite these issues, no large-scale corpus-based study has yet been conducted on article usage by Sri Lankan learners. This creates a significant gap that the present study aims to address.

3. Aims and Research Questions

The overall aim of this research is to investigate the persistent difficulties Sri Lankan ESL learners face in using the English article system and to demonstrate how a corpus-based approach can provide systematic, empirical insights into learner interlanguage. By focusing on writing in English Literature, the study highlights both the linguistic and pedagogical dimensions of article errors.

Specific Aims

1. To identify and categorize article errors in Sri Lankan ESL learner writing, specifically focusing on omission, overuse, and wrong choice.
2. To quantify the frequency and

distribution of article errors across a corpus of learner answer scripts.

3. To analyze article errors contextually through concordance evidence, highlighting recurring collocational patterns and memorized phrases.

4. To interpret error causes in light of SLA theories, particularly L1 transfer, overgeneralization, and fossilization.

5. To demonstrate the pedagogical value of corpus tools like AntConc in diagnosing learner difficulties and shaping targeted teaching practices.

Research Questions

1. What types of article errors are most frequent in Sri Lankan ESL learners' academic writing?

2. How do these errors distribute across learner writing, and what quantitative patterns emerge from corpus analysis?

3. What contextual and collocational patterns can be observed through concordance lines of article usage?

4. To what extent can article errors be explained by L1 influence (Sinhala), overgeneralization, and fossilization?

5. How can corpusbased approaches such as AntConc enhance the validity of error analysis in Sri Lankan ESL research?

4. Methodology

The methodological framework of this study is grounded in corpus-based error

analysis. Following best practices in learner corpus research (Granger, 2008; Gilquin, 2020), the study involves data collection, corpus compilation, and analysis using AntConc.

4.1 Research Design

This study adopts a descriptive and analytical design, aiming to capture authentic learner output and systematically analyze article errors. The approach is both quantitative (frequency counts of errors, distribution analysis) and qualitative (contextual interpretation of concordance lines).

4.2 Data Collection

The primary data set consists of 25 student answer scripts on Literature written by Diploma in English external students at Bhiksu University of Sri Lanka. The scripts, which collectively amount to approximately 20,000 words, were responses to a literature question on William Blake's poem *A Poison Tree*.

Several factors make these scripts particularly suitable:

- **Authenticity:** The texts represent natural learner production under examination conditions, free from teacher correction.
- **Homogeneity:** All learners were responding to the same question, ensuring thematic consistency while still allowing variation in grammatical accuracy.
- **Relevance:** The scripts belong to ESL learners who are from all

walks of life in North Central province. This makes the sample a balanced one with both male and female and representing students from 18 to 55 years of age.

Ethical guidelines were observed: student identities were anonymized, scripts were digitized, and only linguistic features relevant to analysis were retained.

4.3 Corpus Compilation

The collected texts were scanned and digitalized using an Optical Character Recognition (OCR) software and then converted into plain text (.txt) format for compatibility with AntConc. Metadata such as student number, word count, and date were recorded in a spreadsheet for reference.

The final product is a small learner corpus that can be expanded in future research. While modest in size compared to international corpora, it provides sufficient data to identify systematic article error patterns.

4.4 Annotation of Article Errors

Article errors were annotated manually using a two level coding scheme:

1. Error Category

- Omission – failure to use an article where required.
- Overuse – insertion of an article where none is needed.
- Wrong Choice – inappropriate article selection.

2. Error Context

- Noun type: proper noun, common countable noun, uncountable noun.
- Collocational environment: recurring lexical bundles or memorized phrases.

Annotation was conducted by the researcher and validated through peer review with another trained annotator, enhancing reliability.

4.5 Analytical Tools

The primary tool for analysis was AntConc 3.5.9 (Anthony, 2022), a free, open-source concordancer widely used in learner corpus research. AntConc offers several key functionalities relevant to this study:

1. Concordance Tool – displays key word in context (KWIC) lines for articles, making it possible to observe local usage environments and detect patterns of omission or overuse.
2. Word List Tool – generates frequency lists, allowing comparison between article frequencies in the learner corpus and reference corpora.
3. Collocation Tool – identifies recurring lexical partners, useful for detecting memorized or formulaic misuse.

4.6 Analytical Procedures

The analysis proceeded in five steps:

1. Baseline Frequency Analysis: Word lists for a, an, the were generated to

Error Type	Frequency	Percentage	Example from Corpus
Omission	172	44.8%	Blake was visionary poet
Overuse	139	36.2%	The William Blake wrote the A Poison Tree
Wrong Choice	73	19.0%	a English poet
Total	384	100%	

determine raw frequency of article usage.

2. Concordance Analysis: Concordance lines were examined to detect local patterns of misused punctuations.

3. Error Classification: Errors were coded into omission, overuse, and wrong choice, with contextual notes.

4. Quantitative Distribution: Frequencies of each error type were tabulated across all scripts.

5. Qualitative Interpretation: Errors were interpreted in light of SLA theory, considering L1 transfer, overgeneralization, and fossilization.

4.7 Limitations of Methodology

While corpus-based analysis offers systematic insights, certain limitations remain:

- **Corpus Size:** The dataset (25 scripts, ~20,000 words) is relatively small, limiting generalizability.
- **Manual Annotation:** Error coding required human judgment, which may introduce subjectivity despite validation.

- **Genre Specificity:** Since the corpus is based on literature exam scripts, patterns may differ from other genres.

- **Lack of Spoken Data:** The study focuses exclusively on written learner production, excluding spoken interlanguage.

5. Analysis and Findings

This section presents the results of the corpusbased error analysis of article usage among Sri Lankan ESL learners. Findings are divided into three parts: overall distribution of article errors, detailed analysis of error types, and qualitative insights from concordance evidence.

5.1 Overall Distribution of Article Errors

The learner corpus contained approximately 20,000 running words across 25 answer scripts. Within this dataset, all three articles were used. However, their distribution revealed marked irregularities when compared to standard English norms.

A total of 384 article errors were identified, representing a significant

proportion of grammatical errors in the scripts. The distribution across error types is shown below:

Table 1. Distribution of Article Errors

The results show that omission errors were most frequent, followed closely by overuse of “the.” Wrong choice errors, while less frequent, still appeared consistently.

5.2 Error Type 1: Omission of Articles

Omission errors accounted for nearly half of all mistakes. Learners frequently dropped the indefinite article a/an before singular countable nouns, particularly in descriptions of Blake as a poet or philosopher.

Examples (from learner corpus):

- William Blake was English poet of 18th century.
- He had visionary power to see nature deeply.
- Speaker in poem is angry man who hides feelings.

Such omissions reflect L1 transfer. Sinhala do not require determiners before nouns, so learners often perceive articles as unnecessary. Additionally, fossilization appears to play a role, since these errors persisted across multiple scripts.

Concordance Sample (AntConc, keyword = “poet”):

Blake was English poet who challenged...

Blake an English poet and painter of...
the great English poet of Romantic period

Here, we see both omission (English poet) and correct use (an English poet) occurring in the same corpus, suggesting inconsistent mastery.

5.3 Error Type 2: Overuse of “the”

Overuse errors were the second most common category. Learners frequently inserted the before proper nouns, titles, and generic terms.

Examples (from learner corpus):

- The William Blake is famous for symbolism.
- The A Poison Tree shows anger of the man.
- The nature is very powerful in Blake’s poetry.

The repeated collocation “The A Poison Tree” occurred in 11 different scripts, suggesting it may be the result of classroom instruction or memorized notes.

Concordance Sample (keyword = “The” before names/titles):

The William Blake is visionary poet...
The A Poison Tree describes consequences...
The Songs of Innocence and Experience show...

This suggests learners are overgeneralizing the rule that titles

Article	Frequency in Learner Corpus	% of Total Tokens	Approx. Frequency in Native Corpora*
The	1,438	7.2%	~5.8% (BNC Academic sub-corpus)
A	294	1.5%	~2.3%
An	82	0.4%	~0.8%

require capitalization and treating them as necessarily accompanied by the.

Overuse also occurred with generic plural nouns:

- The trees give us life.
- The people should control anger.

In these cases, learners appear to transfer definiteness from their L1s, where generic nouns often carry a determiner.

5.4 Error Type 3: Wrong Choice

Wrong choice errors involved misuse of a/an and occasional substitution of the where a/an was required.

Examples:

- Blake was a English poet.
- He was an poet of revolution.
- He became angry with the man (intended: a man in general).

These errors often arose in contexts where learners attempted to apply phonological rules but misinterpreted the relationship between spelling and sound. The phrase a English poet was found in 7 different scripts, suggesting fossilized misunderstanding.

5.5 Quantitative Comparison of Article Use

A frequency analysis of article tokens showed disproportionate reliance on the.

Table 2. Raw Frequencies of Article Tokens in Corpus

A frequency analysis of article tokens showed disproportionate reliance on the.

Table 2. Raw Frequencies of Article Tokens in Corpus

*Reference values based on BNC (Leech, Conrad & Culpeper, 1994).

The results show that learners overused “the” and underused “a/an” compared to native norms. This imbalance reflects the error types discussed earlier.

5.6 Collocational Patterns

Collocation analysis in AntConc revealed systematic misuse in certain recurring word clusters:

1. “The William Blake” – 14 occurrences
2. “The A Poison Tree” – 11 occurrences
3. “A English poet” – 7 occurrences
4. “The nature” (generic reference) –

19 occurrences

These collocations indicate that article misuse is not random but rather formulaic, suggesting that learners acquire erroneous chunks from classroom exposure, peer interaction, or rote learning.

5.7 Qualitative Insights

Beyond raw frequencies, qualitative analysis showed the following trends:

- **Memorized Expressions:** Learners often repeated phrases such as the A Poison Tree and the William Blake, implying transfer of teacher-provided notes into writing without grammatical adjustment.
- **Genre Influence:** Because these were literature scripts, learners often wrote about authors, texts, and abstract concepts, contexts where article rules are less transparent.
- **Inconsistency:** Some learners alternated between correct and incorrect usage within the same script, indicating partial mastery rather than complete ignorance.
- **Pragmatic Simplification:** Learners occasionally omitted articles in complex noun phrases to reduce processing load (He is poet who shows anger), reflecting cognitive strategies described in interlanguage theory.

5.8 Summary of Findings

1. Omission was the most frequent error (44.8%), reflecting L1 influence

from Sinhala/Tamil.

2. Overuse of “the” (36.2%) was widespread, particularly before proper nouns, titles, and generic terms.

3. Wrong choice errors (19%) revealed gaps in phonological awareness of a/an usage.

4. Collocation evidence showed that errors are often formulaic, not random, suggesting pedagogical origins.

5. Frequency imbalance confirmed overreliance on the and underuse of a/an, aligning with global research on article acquisition.

6. Discussion

The results of the corpus-based analysis confirm that Sri Lankan ESL learners face persistent difficulties in mastering the English article system. By categorizing article errors into omission, overuse, and wrong choice, and by quantifying their distribution across a learner corpus, the study sheds light on both universal tendencies in SLA and Sri Lanka-specific challenges. This section discusses the implications of these findings in relation to interlanguage theory, error causation, pedagogical practice, and the need for corpus-informed instruction.

6.1 Article Errors as Evidence of Interlanguage

The findings highlight that learners’ use of articles reflects a developing interlanguage system rather than random mistakes. Corder (1967) argued that learner errors should be seen as windows into the learning process. In

this corpus, repeated errors such as the William Blake or a English poet are not accidental slips but systematic features of learner interlanguage.

- Omission errors suggest that learners have not yet fully internalized the obligatory role of determiners before singular countable nouns.
- Overuse errors reveal learners' strategies of overgeneralization: assuming that all proper names and titles require an article.
- Wrong choice errors show confusion in applying phonological rules, reflecting partial rather than complete rule acquisition.

In all three cases, learner output demonstrates systematicity, aligning with interlanguage theory's claim that errors represent developmental stages rather than deficiencies.

6.2 Causes of Article Errors

Three major factors explain the persistence of article errors in Sri Lankan ESL learners:

6.2.1 L1 Transfer

Neither Sinhala nor Tamil has a grammaticalized article system. Sinhala uses demonstratives (*eka* for singular, *vala* for plural) to mark definiteness in certain contexts, while Tamil uses case endings and particles, but neither language has obligatory determiners

equivalent to *a*, *an*, *the*. As a result, learners often transfer zero-article usage from their L1s, leading to omission (*He is poet*) and overuse of the in inappropriate contexts (*The nature is beautiful*).

6.2.2 Overgeneralization

Learners frequently extend partially acquired rules to contexts where they do not apply. The repeated phrase *the A Poison Tree* shows overgeneralization of the rule that literary works and proper names require *the* (as in *the Iliad*). Similarly, learners apply *the* to all nouns in generic reference, overgeneralizing from contexts where definiteness is required.

6.2.3 Fossilization

Some errors, appear across multiple scripts, suggesting fossilization, when incorrect forms become entrenched in learner interlanguage due to insufficient corrective feedback (Selinker, 1972). Because such errors are rarely noticed in classroom contexts, they persist into advanced levels.

Conclusion

This study has investigated article usage among Sri Lankan ESL learners through a corpus-based error analysis of 25 English literature scripts. By focusing on omission, overuse, and wrong choice errors, and by applying concordance and frequency analysis in AntConc, the research has yielded insights of both theoretical and pedagogical value.

References

- Abeywickrama, R. (2010). An Analysis of Errors in English Writing of Sinhala Speaking Undergraduates. *Sabaragamuwa University Journal* Volume 9 Number 1, 97–114.
- Anthony, L. (2022). AntConc (Version 3.5.9) [Computer software]. Waseda University. <http://www.laurenceanthony.net/software/antconc/>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Brown, P., & Chua, S. (2019). Lexical bundles in Singaporean business communication: A corpusbased analysis. *Asian Englishes*, 21(2), 145–161.
- Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, 5(1–4), 161–170.
- Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. Oxford University Press.
- Erdogan, V. (2005). Contribution of error analysis to foreign language teaching. *Mersin University Journal of the Faculty of Education*, 1(2), 261–270.
- Gilquin, G. (2020). Learner corpora and second language acquisition: The proof of the pudding is in the eating. *International Journal of Learner Corpus Research*, 6(2), 123–145.
- Granger, Sylviane. (2003). *The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research*.
- Granger, S. (2008). Learner corpora in foreign language education. *Language Teaching*, 41(3), 441–454.
- Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.
- Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Karoma Publishers.
- Ishikawa, Shin'ichiro. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*. 1. 91–118.
- Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, 12(1), 3–69.
- Jayaweera, A., & Dias, N. (2014). Unknown words analysis in pos tagging of sinhala language. *Machine Learning and Applications: An International Journal (MLAIJ)* Vol.1, No.2.
- Karunanayake, T., Perera, D., & Silva, M. (2022). Corpus resources for Sri Lankan English: Past, present and future directions. *South Asian Englishes Journal*, 2(1), 55–74.
- Kennedy, G. (2001). *An introduction to corpus linguistics*. Routledge.
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated grammatical error detection for language learners*. Morgan & Claypool.
- Leech, G., Conrad, S., & Culpeper, J. (1994). *The British National Corpus: A skeletal grammar and pragmatics*. Longman.
- Liyanage, I. (2023). Language Policy and Medium of Instruction in Asian Higher Education: Neoliberalism and Nationalist Responses. In *The Palgrave Handbook of Asian Higher Education Policy* (pp. 105–126). Palgrave Macmillan.
- Lyons, C. (1999). *Definiteness*. Cambridge University Press.
- Master, P. (1997). *The English article system*:

-
- Acquisition, function, and pedagogy. *System*, 25(2), 215–232.
- Meyler, M. (2007). *A Dictionary of Sri Lankan English*. Mendis Brothers.
- Nelson, M. (2000). *A corpus-based study of the lexis of business English and business English teaching materials* (Doctoral dissertation, University of Manchester).
- Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2), 135–172.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10(1–4), 209–232.
- Swan, M. (2016). *Practical English Usage* (4th ed.). Oxford University Press.
- Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 865–872.
- Wijeratne, W., & Jeyaseelan, S. (2015). Morphological errors in ESL writing of Sri Lankan undergraduates. *Sri Lanka Journal of Linguistics*, 19(2), 23–45.