ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

THE CORPUS-BASED ANALYSIS OF THE DEFINITE ARTICLE IN SELECTED PAKISTANI ENGLISH ESSAYS

[1] Khadija Tul Kubra, [2] Fatima Tuz Zahra
1. Student Researcher, Department of English, Minhaj University Lahore, Punjab, Pakistan.
2. Lecturer, Department of English, Minhaj University Lahore, Punjab, Pakistan.
Corresponding Author: khadija.shahid610@gmail.com

ABSTRACT
This study investigates the use of the definite article "the" in Pakistani undergraduate students' academic essays and explores how the first language (L1) influences and shapes article use. A self-compiled corpus of forty essays by undergraduate students was analyzed using AntConc (Version 3.5.8) to examine the frequency, distribution, and contextual patterns of "the". The British National Corpus (BNC) was used as a reference to establish native-English benchmarks. Quantitative analysis revealed that Pakistani English exhibits a higher frequency of "the" compared to Standard English. Quantitative analysis showed that Pakistani English demonstrates a higher frequency of 'the' than Standard English. In addition, through qualitative concordance analysis, several recurrent patterns in its usage were identified, including overuse in generic or non-specific contexts and omission in definite contexts. These tendencies reflect L1 transfer from Urdu, which lacks an article system, and highlight features of Pakistani English as an evolving localized variety. The findings add to the understanding of article usage in World Englishes and offer pedagogical insight into improving academic writing instruction in similar linguistic contexts.

Keywords: Pakistani English, definite article, corpus linguistics, AntConc, British National Corpus, World Englishes, L1 influence.

INTRODUCTION
English possesses a complex article system that plays the leading role in indicating definiteness, specificity, and shared knowledge between writers and readers. However, this system presents some serious difficulties with regard to using the definite article "*the*" correctly among second-language learners due to an interaction of linguistic, cognitive, and contextual factors. In Pakistan, English is taught as a second language, while Urdu functions as the first language for most learners. Because Urdu does not have an article system like English, Pakistani students usually depend upon transfer from their L1 in producing English academic text. As a result, their usage of the definite article does not always align with the Standard English norm but rather follows patterns shaped by the linguistic structures of Urdu and developing norms of Pakistani English.

Corpus linguistics provides a systematic way to analyze these patterns by allowing researchers to study authentic language use in real contexts. With the increasing recognition of Pakistani English as an emerging variety within the paradigm of World Englishes, the study of article usage becomes important to show how local linguistic habits mold English writing in Pakistan. Therefore, this study uses a corpus-based approach to analyze how the definite article "the" is used by undergraduate Pakistani students in academic essay writing and compares these patterns with the British National Corpus (BNC) to highlight similarities, differences, and emerging linguistic features.

The study focuses on frequency, concordance patterns, and contextual usage in providing empirical insights into article variation in Pakistani English. The findings not only contribute to the description of Pakistani English but also have pedagogical relevance for teaching English in Pakistan, teaching academic writing, and developing ESL curricula.

RESEARCH OBJECTIVES
This study aims to achieve the following objectives:

1. To analyze the frequency distribution of the definite article "the" in Pakistani English undergraduate essay writing.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

2. To compare the usage patterns of "the" in Pakistani English with those in the British National Corpus (BNC).

SIGNIFICANCE OF THE STUDY

This study holds a significant number of dimensions, both theoretical and practical. First, it adds to the empirical work on Pakistani English by documenting the use of the definite article "*the*" among undergraduate students, a topic that has often been neglected in favor of broader error analyses. Using a corpus-based approach, the research provides an objective, data-driven analysis that moves beyond subjective evaluations of grammatical correctness.

Second, the findings carry pedagogical implications for English language teaching and curriculum design in Pakistan. A clearer understanding of how Pakistani learners use "*the"* can help instructors tackle persistent problems with article usage, enhance academic writing instruction, and refine grammar and composition teaching strategies. These insights are valuable for English departments where writing proficiency is at the core of the program.

Third, this study contributes to the theoretical discussions of second language L2 varieties of English by demonstrating how article usage in Pakistani English reflects both L1 Urdu influence and processes of linguistic nativization. This supports the recognition of Pakistani English as a legitimate and evolving variety within the World Englishes framework rather than a set of learner deviations.

Finally, by comparing a locally compiled undergraduate student corpus with the British National Corpus, for instance, the present study highlights how Pakistani English aligns with or deviates from the norms of Standard English. This leads to further insight into local patterns and provides empirical evidence relevant to debates on standardization, linguistic legitimacy, and policies on the teaching of English in multilingual settings.

RESEARCH QUESTIONS

1. Does the frequency of the definite article in Pakistani English differ from its frequency in Standard English?
2. How do Pakistani English undergraduate students' usage patterns definite article in their writing compare to Standard English norms?

LITERATURE REVIEW

Research on English article usage has long established that articles pose significant challenges for second-language learners, most especially for those speakers whose first language lacks an equivalent system. Within the Pakistani context, a few studies have focused on recurring patterns in the use of the definite article "the," attributing them largely to the influence of L1, developmental factors, and the evolving norms of Pakistani English.

Early research focused primarily on error analysis. Ishaq (2016) reported that Urdu speakers commit persistent article-related errors, pointing to overuse and omission patterns in both spoken and written forms of English. Likewise, Maqbool, Hussain, and Azhar (2018) recorded frequent variations in article usage among Pakistani students and have claimed that such patterns result from negative transfer from Urdu, which does not grammatically encode definiteness.

More recent research moves away from error-focused approaches toward an understanding of Pakistani English as a localized variety. Bergström (2021) has investigated teachers' attitudes to non-standard article usage and suggested that many variations which are regularly found in Pakistani English writing may reflect systematic features rather than random errors. These findings are in line with the World Englishes approach, which acknowledges nativised forms of English used in particular sociolinguistic environments.

Corpus-based studies have contributed to this understanding. Buriro (2023) examined grammatical features of Pakistani English using corpus methods and observed striking variation in both placement and frequency of articles, which might indicate that article usage

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

in Pakistan is becoming stabilized as part of a local norm. Newspaper and fiction research, such as Sajjad (2023) and Areej (2024), also evidences patterns of article omission or extension different from those in Standard English but consistent within Pakistani discourse.

Similar trends come to light in studies specifically dealing with the definite article. Studies on Pakistani English media, Yasmin, for instance, and Richtmann Publications, record both overuse before generalized nouns and omission where Standard English requires definiteness. Acquisition studies, such as on Urdu L1 speakers learning English (2018), note that learners depend more on semantic and pragmatic clues than on grammatical ones, which leads to different patterns from the native English corpora.

Although previous studies have examined article usage in newspapers, fiction, and spoken discourse, only a few have used corpus-based methods to explore undergraduate academic writing, where article accuracy is crucial. Furthermore, there is limited research that compares Pakistani student writing with a reference corpus, such as the British National Corpus (BNC), which is necessary to evaluate and identify systematic differences. This study addresses this gap by analyzing and comparing the frequency and concordance patterns of "the" in Pakistani undergraduate essays using AntConc with BNC data for clearer insight into whether such patterns represent emerging norms or persistent grammatical errors.

RESEARCH GAP

Although previous studies have examined article use in Pakistani English, most of the available research involves newspapers, fiction, teacher attitudes, or spoken discourse. Only a few studies have used corpus-based approaches to investigate student writing, and fewer still have targeted the definite article "the." Moreover, many earlier studies rely heavily on error analysis, treating differences from Standard English as errors, rather than as potential features of an emergent localized variety.

There is also a significant gap in terms of methodological rigor: few studies contrast Pakistani English use with a standardized reference corpus like the British National Corpus. In the absence of such a comparison, it is difficult to tell whether any observed pattern represents error, developmental stage, or systematic feature of Pakistani English.

Another gap relates to undergraduate academic writing, which is an important yet under-explored genre. Undergraduate students majoring in English are regularly involved in academic essays, and this makes their writing an important source for exploring the ways in which article norms are internalized, negotiated, and expressed in Pakistani English. However, very few studies have built a student writing corpus for empirical analysis of article use.

This paper fills these gaps by (a) compiling a self-built corpus of forty undergraduate essays, (b) conducting a frequency and concordance comparison with the BNC, and (c) framing the findings in both L1 influence theory and the World Englishes framework that differentiates between learner errors and emerging linguistic patterns.

METHODOLOGY

This study employed a corpus-based research design that examined the use of the definite article "the" in Pakistani undergraduate essay writing. The methodology combines both quantitative and qualitative approaches to provide a comprehensive understanding of how article usage in Pakistani English differs from Standard English. In the following sections, describe the sample, materials, data collection procedures, and analytical methods used in the study

Sample

The sample consisted of forty essays written by undergraduate students of a Pakistani university. The purposive sampling method was used to select essays that reflect typical Pakistani English academic writing. These essays are authentic classroom-based compositions

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

and, therefore, provide reliable data for examining how Pakistani learners use the definite article.

Materials

The present study is based on two corpora: one was a self-compiled Pakistani English Corpus (PEC), made up of the forty undergraduate essays selected for analysis in this study, while the second was the British National Corpus (BNC), which served as the Standard English corpus. The AntConc application (Version 3.5.8) was the main tool for corpus analysis in producing word frequency lists, concordance lines, and keyword in context (KWIC) patterns. The Pakistani English self-compiled Corpus was analyzed on AntConc to retrieve a frequency list and concordance lines. On the other hand, the BNC was the reference corpus from which the corresponding frequency lists, concordance lines, and KWIC patterns, representative of Standard English, were extracted. All essays were first converted into plain text (.txt) format to ensure smooth processing in AntConc.

Data Collection Procedure

The research design was systematic. The collection of forty essays from Undergraduate students was the first step. These essays were cleaned by removing headings, page numbers, and any other non-relevant formatting. After cleaning, all the essays were converted to .txt format and loaded into AntConc as a single folder to form the Pakistani English Corpus. The corresponding sample from the British National Corpus (BNC) was used as the reference for Standard English. AntConc (Version 3.5.8) was employed only for analyzing the self-compiled Pakistani English Corpus (PEC), generating frequency counts and concordance lines for the definite article. The BNC data, on the other hand, provided reference frequency and concordance information for Standard English, which was used for comparison with the AntConc-generated results from the Pakistani corpus.

Data Analysis

The quantitative and qualitative analytical approaches were employed in this study. Quantitatively, the frequency of "the" in both the Pakistani English Corpus and the BNC was calculated and normalized (frequency per thousand words) to enable accurate comparison. Qualitatively, concordance lines generated in AntConc were analyzed for contextual patterns such as overuse, omission, generalized use, or structural variation. These patterns were interpreted with the help of two theoretical frameworks: the L1 Influence Theory to clarify how the lack of an article system in Urdu shapes article acquisition, and the framework of World Englishes, which considers linguistic variation in Pakistani English as a part of the natural development of a localized English variety rather than just being errors.

ETHICAL CONSIDERATIONS

This research followed the standard ethical principles of working with students' academic work that they themselves produced. The corpus essays were made available on the basis that they were only to be used for research purposes. No personal information, names, or other details that would identify students were included in the corpus to ensure strict anonymity and confidentiality. Data were processed responsibly, securely stored, and accessible only to the researcher and team. Because students did not have personal risk, the analysis was based solely on linguistic features, and no evaluation of students' performance was included. The study did not modify, grade, or comment on the work of students beyond its linguistic examination in order to preserve respect for their academic work and authorship.

RESULTS AND DISCUSSION

The analysis of the forty undergraduate essays revealed distinct patterns in the use of the definite article *"the"* in Pakistani English writing. These findings, when compared with the British National Corpus (BNC), these results underscore both quantitative and qualitative differences that help define Pakistani English as an emerging localized variety.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Corpus Type	Total Words	Frequency of the	Concordance Hits (the)
Pakistani English Corpus	17,108	1,054	1,054
Standard English Corpus	100 million (large national corpus)	143,476	143,476

Table 1: This table highlights the striking difference in how frequently Pakistani students use the definite article "the" compared to Standard English norms.

Frequency Findings

The quantitative analysis showed that the definite article "the" appeared more frequently in the Pakistani English Corpus than in the BNC reference sample. Even after normalizing for every thousand words, Pakistani students demonstrated a significantly higher rate of usage. This indicates a tendency toward overgeneralization, where "the" is used in contexts that do not require definiteness in Standard English. Such overuse aligns with prior studies on Pakistani English, which attribute such patterns to first-language (L1) influence and structural transfer from Urdu.

Concordance Findings

The qualitative concordance analysis identified several recurring patterns of variation:

1. **Generalized Use Before Non-Specific Nouns**

   Students frequently used "the" before plural or abstract nouns, even when referring to general or non-specific concepts (e.g., the people, the education, the women). This reflects transfer from Urdu, where definiteness is understood through context rather than grammatical markers.

2. Overextension in Generic Contexts

"The" was often used to generalize entire categories (e.g., the students should learn, the society must change). While acceptable in some cases, its high frequency suggests an evolving norm in Pakistani English that differs from Standard English preferences for omitting the article in generic contexts.

2. **Occasional Omission in Required Contexts**

   Although overuse was more common, omissions were also observed, particularly before unique nouns or specific references (e.g., "government should…"). These omissions likely result from Urdu's lack of article distinction, leading learners to depend on meaning rather than grammatical convention.

3. **Contextual Variation and Inconsistency**

   Some essays exhibited inconsistent patterns, alternating between correct Standard English usage and localized forms. This inconsistency suggests that students may be in a transitional phase of article acquisition, balancing learned rules with intuitive linguistic tendencies shaped by Urdu.

Interpretation of Findings

From the perspective of L1 Influence Theory, the patterns clearly demonstrate how Urdu's lack of an article system impacts English usage. Learners tend to overuse "the" as a strategy to maintain clarity, which leads to higher overall frequency rates.

Through the World Englishes framework, these patterns appear not merely as learner errors but as emerging linguistic norms. The consistency of these variations across multiple essays and writers indicates that Pakistani English is developing distinct grammatical tendencies.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Overall, the findings suggest that differences in article usage among Pakistani learners should not be viewed solely as mistakes but as indicators of a localized English variety shaped by sociolinguistic, educational, and linguistic influences.

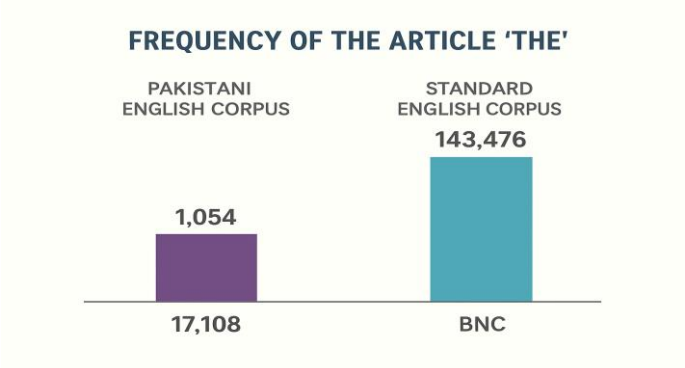

Figure 1: This figure visually showcases the significantly higher use of "the" in Pakistani English, revealing a clear pattern of overgeneralization.

CONCLUSION

This research explores how the definite article "the" is used in Pakistani undergraduate English essays through a corpus-based comparison with the British National Corpus (BNC). By analyzing forty essays using AntConc, it identified clear differences in both the frequency and contextual use of "the" between Pakistani English and Standard English.

The quantitative findings revealed a higher normalized frequency of "the" in the Pakistani English Corpus, suggesting a pattern of overgeneralization. Qualitative concordance analysis further showed recurrent tendencies such as generalized use before non-specific nouns, overextension in generic contexts, occasional omission in obligatory cases, and inconsistent use across essays. These patterns reflect the influence of Urdu, which lacks an article system, and illustrate how Pakistani learners depend more on meaning and context than on grammatical structure.

Viewed through the lens of the World Englishes framework, these patterns should not simply be classified as learner errors. Rather, they represent systematic and developing features of Pakistani English, shaped by local linguistic habits, educational settings, and cultural influences. In documenting these variations, this research adds to the growing recognition of Pakistani English as an evolving variety with its own linguistic norms.

The findings also offer practical implications for English language teaching in Pakistan. By understanding these recurring patterns, educators can develop more effective teaching materials, address common difficulties in writing, and adopt instructional approaches that balance adherence to Standard English with sensitivity to localized usage.

Future research could build on this study by including larger corpora, exploring other genres such as spoken or research writing, or conducting comparative analyses with other South Asian English varieties. Overall, this study provides a foundation for continued investigation into article usage and the broader grammatical characteristics of Pakistani English.

REFERENCES

Areej, A. (2024). *A corpus-based analysis of Pakistani English novels*. Pakistan Language and Humanities Review.

Bergström, A. (2021). *Non-standard article use in Pakistani English: An attitudinal study* (Bachelor's thesis). DIVA Portal.

British National Corpus (BNC). (2007). *BNC Consortium*. University of Oxford.

Buriro, A. (2023). *Grammatical features of Pakistani English: A corpus analysis*. National Journal of Arts & Social Sciences.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Hyland, K. (2009). *Academic discourse: English in a global context*. Continuum.

Ishaq, K. (2016). Persistent errors in English writing and speech made by Urdu speakers. *KJLR: Kashmir Journal of Language Research*.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Maqbool, M., Hussain, M., & Azhar, M. (2018). Error analysis of English language speakers in the use of articles in Pakistan. *Journal of Educational Research*.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Sajjad, U. (2023). *A corpus-based study of Pakistani English newspaper (COPENA)*. Journal of Development and Social Sciences.

Swan, M. (2005). *Practical English usage* (3rd ed.). Oxford University Press.

Trudgill, P. (2021). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Yasmin, M. (2015). A corpus-assisted study of linguistic features of Pakistani English media. *Pakistan Social Sciences Review*.

Yule, G. (2010). *The study of language* (4th ed.). Cambridge University Press.

Zhang, S. (2018). The acquisition of English syntactic structures by Urdu L1 speakers in Pakistan: The case of articles. *ResearchGate*.