Shajadul Alam Sweet*, Md. Nurul Kabir Emon,
Sadia Khandokar Mim, Fakhrul Islam Mahim, Mahfuj Hosen,
Partho Biswas, Easmin Sultana and Md. Mehedi Hasan Emon

Corpus-informed vocabulary instruction in Bangladesh: a practical framework for EFL contexts

Abstract: This study details the design and classroom implementation of a 36-week, corpus-informed vocabulary programme developed for Bangladeshi EFL learners. The study aimed to create a practical, replicable teaching framework to enhance academic vocabulary acquisition in resource-limited educational settings. The intervention combined corpus consultation tasks, guided collocation discovery, and production-focused writing activities integrated into the standard English curriculum. A single-group pre–post feasibility study was used to assess the practicality of classroom delivery, teacher workload, and observable language improvements. Participants completed writing tasks before and after the intervention, which were analysed for academic vocabulary coverage, collocation accuracy, and use of formulaic sequences. Descriptive analysis showed notable lexical and collocational gains, supported by teacher reflections and student focus groups. Rater agreement for coding was high (κ = 0.82), indicating reliability. Results suggest that corpus-informed vocabulary instruction can be feasibly incorporated into existing curricula when materials and teacher support are adapted to local contexts. Although causal conclusions cannot be drawn without a control group, this study offers an evidence-

---

*Corresponding author: Shajadul Alam Sweet, Department of English, University of Asia Pacific, Dhaka, Bangladesh, E-mail: shajadulalamsweet@gmail.com. https://orcid.org/0009-0005-0994-7502
Md. Nurul Kabir Emon, Sadia Khandokar Mim, Fakhrul Islam Mahim, Mahfuj Hosen, Partho Biswas and Md. Mehedi Hasan Emon, Department of English, University of Asia Pacific, Dhaka, Bangladesh, E-mail: nurulkabiremon@gmail.com (M.N.K. Emon), khandokarsadia63@gmail.com (S.K. Mim), fakhrulmahim@gmail.com (F.I. Mahim), 21216056@uap-bd.edu (M. Hosen), parthobiswas049@gmail.com (P. Biswas), mehedihasanemon2349@gmail.com (M.M.H. Emon). https://orcid.org/0009-0003-5680-5975 (M.N.K. Emon). https://orcid.org/0009-0006-9492-3544 (S.K. Mim). https://orcid.org/0009-0002-5073-9981 (F.I. Mahim). https://orcid.org/0009-0001-1867-6995 (M. Hosen). https://orcid.org/0009-0006-5980-104X (P. Biswas). https://orcid.org/0009-0003-7016-4190 (M.M.H. Emon)
Easmin Sultana, Department of English, Royal University of Dhaka, Dhaka, Bangladesh, E-mail: e.s.jafrin@gmail.com

based procedural model and transparent analytical framework for future large-scale, controlled research in EFL settings.

Keywords: vocabulary learning; corpus linguistics; EFL instruction; academic collocations; L1 transfer; Bangladesh

1 Introduction

The challenge of acquiring English vocabulary in Bangladesh cannot be fully tackled with traditional teaching methods. Despite years of studying formal English, most Bangladeshi students lack lexical sophistication, especially in academic settings (Rahman and Pandian 2018). This language barrier significantly impacts their academic performance and future career prospects in an increasingly English-dependent job market.

The problem is not with the learners themselves but with the lack of correlation between classroom instructional activities and learners' actual communicative needs. In Bangladeshi classrooms, vocabulary instruction is still based on memorisation, translation, and context-free grammar tasks rather than language and its use. These strategies fail to build the high level of vocabulary that students need to achieve academic success – that is, to know how to use language correctly and to mix words with others in different ways.

They can be addressed using corpus linguistics. The study of large amounts of original texts can help to identify the inherent tendencies of the English language that are not necessarily evident in traditional sources (Flowerdew 2012). Most corpus-sensitive pedagogical practice, however, is structured within well-resourced educational institutions of high technology and elaborate teacher-training courses. The question is: Can these strategies be applied in situations where resources are scarce but the need for effective vocabulary instruction is urgent, as in Bangladesh?

The paper will answer this question by designing and testing a corpus-informed vocabulary model tailored to Bangladesh's education system. We did not simply import existing practices; instead, we carefully adapted the corpus's principles to local requirements without diminishing their pedagogical impact. Three aspects of the corpus analysis, which can make the most significant impact, are described in the framework: the detection of locally relevant academic vocabulary, the prediction of systematic errors caused by the effects of the first language (L1), and the possibility to sustainably train a teacher who will not always have to resort to external resources.

Our practice departs from deficit-based explanations of learners' issues toward systematic, evidence-based solutions. We do not regard the difficulties students face as unsolvable problems, but rather as tendencies that can be addressed through a specific set of instructions. The perspective acknowledges learners' multilingual

skills as an asset rather than a barrier. It also provides concrete resources that could enable them to deal with the peculiarities of learning English vocabulary. The current research, thus, is not aimed at demonstrating effectiveness, but recording a repeatable and context-specific model of vocabulary teaching.

2 Literature review

2.1 Theoretical foundations in vocabulary learning

Vocabulary knowledge is more complex than simple knowledge of what the words mean. The vocabulary learning process, as described by Nation (2013), is multidimensional and involves form (the way words appear and sound), meaning (what words mean), and use (how words are used). The emphasis should be put on teaching vocabulary to address all these dimensions, not one or two.

The theory of task-induced involvement, proposed by Laufer and Hulstijn (2001), explains why some vocabulary-learning exercises are more effective than others. According to this theory, learning activities that engage learners in evaluating, searching, and actively applying target vocabulary items lead to greater retention than passive exposure or automatic repetition. The implications of this principle for corpus-informed activities are considerable, and these activities engage learners in meaningful work with the target vocabulary.

However, when applying these theoretical frameworks in an EFL context like Bangladesh, one must be aware of local factors. English is taught as a foreign language in Bangladesh, as well as a second language, in an environment referred to as the expanding circle, in which learners are exposed to English to a limited degree outside the classroom. In such situations, students face special difficulties in learning the language because they need to master higher stages of lexical development with minimal real input.

Furthermore, the assessment culture in education in Bangladesh has the potential to cause stress in pursuit of immediate, measurable results, which may be incongruent with the progressive nature of vocabulary development (Rahman 2020). The pressure on both teachers and students is to focus on aspects of English that are readily evaluated, which, in the vast majority of cases, are grammar rules and discrete vocabulary items, rather than patterns of natural language use.

2.2 Corpus linguistics and vocabulary instruction

Corpus linguistics has transformed our understanding of how vocabulary works in real English texts. Researchers have been able to define patterns by analysing

millions of words in authentic contexts, patterns that were previously invisible to language teachers and students (Sinclair 1991). These lessons have led to significant advancements in vocabulary education, particularly in mastering collocations, formulaic phrases, and academic vocabulary.

Word pairs that are natural combinations are known as collocations, which are an essential but frequently not well-studied part of vocabulary. Conventional vocabulary instruction generally focuses on individual words, whereas corpus studies reveal that proficient writers combine predictable word sequences to form natural-sounding prose (Nesselhauf 2005). In academic writing, in particular, certain collocations are necessary to achieve the accurate, formal register required for academic discourse (Ackermann and Chen 2013).

Learners processing corpus data directly through data-driven learning (DDL) techniques, which require identifying patterns in language, have demonstrated specific potential in vocabulary acquisition (Boulton and Cobb 2017). These approaches guide learners in developing metalinguistic awareness – understanding how language functions – which, in turn, facilitates long-term vocabulary acquisition. Nevertheless, the majority of DDL studies have been conducted in environments with well-developed technological infrastructure and comprehensive teacher preparation programs.

The difficulty lies in adapting these insights to resource-constrained environments. Conventional corpus-informed teaching typically presupposes access to large-scale corpora, specialized software, teacher training, and flexible curricula, which enable the use of innovative methods (Kennedy and Miceli 2017). These demands pose severe constraints in situations where basic educational materials are scarce.

2.3 L1 transfer in vocabulary learning

The vocabulary difficulties faced by Bangladeshi learners are often influenced by the transfer of Bangla, their first language. L1 transfer occurs when learners transfer patterns from their mother tongue to the target language, and this may lead to mistakes; however, it is also a valuable source of learning (Jarvis and Pavlenko 2008).

Among Bangladeshi learners of English, recurring transfer patterns were observed, particularly in the formation of collocations and the use of prepositions. As a case in point, when learners are asked to write to conduct research, they tend to write to make research, since the word in the Bangla language is a verb that is closer to "make" than "conduct." These are predictable, rule-governed patterns; hence, they can be taught systematically rather than corrected ad hoc.

However, conventional methods of transfer often have a deficit view of the transfer, and L1 influence is perceived as a nuisance that should be eliminated. Current studies indicate that transfer may become a beneficial phenomenon in

language learning when learners become aware of cross-linguistic differences and similarities (Ringbom 2007). This outlook enables a shift in emphasis on the removal of transfer to assist learners in utilising their multilingual resources strategically.

Cross-linguistic studies have shown that patterns of transfer are predictable in systematic ways by examining the structural variations between learners' first and target languages (Odlin 1989). This predictive ability enables the formulation of proactive teaching strategies that address potential challenges before they become entrenched.

2.4 Assessment and institutional contexts

Contemporary assessment practices should be encouraged to support vocabulary development initiatives in educational settings, as they will help demonstrate institutional support and keep students motivated. However, traditional knowledge-testing models are more likely to emphasise discrete knowledge rather than its application to context, which inherently breaks with the ideal practice of instruction (Read 2000). The accuracy of collocations, the correct application of the formula, and the degree of lexical sophistication are aspects of words-in-context measurement that are prioritised in modern vocabulary testing (Kremmel et al. 2017). Such practices are less inconsistent with instructional practice guided by the corpus since they focus on language patterns in real-life situations rather than knowledge outside of context.

In such a scenario, as in Bangladesh, the challenge is to implement new assessment innovations that focus on vocabulary without requiring extensive changes to existing systems. New teaching must be clearly connected to the organisational goals, and students and educators in the institution will not oppose new approaches, even though they may have pedagogical advantages (Wall and Alderson 1993).

2.5 Research questions

This study investigated three research questions (RQs):

RQ1: What specific vocabulary patterns characterize Bangladeshi learners' academic writing compared to target academic discourse?

RQ2: How can a corpus-informed vocabulary framework, designed from these findings, function effectively within Bangladesh's educational constraints while producing measurable learning improvements?

RQ3: What implementation protocols enable teachers to sustain corpus-informed vocabulary instruction independently, and how feasible is the approach for broader adoption across institutions?

3 Methodology

3.1 Research design

The research examined in the article used a single-group, mixed-methods feasibility design to assess the feasibility and classroom applicability of corpus-informed vocabulary instruction in Bangladeshi EFL classrooms. The quantitative part considered changes in the written language of the students before and after the intervention period of 36 weeks. In its qualitative part, teacher interviews and student focus group discussions were analysed to provide a picture of the perception of feasibility and pedagogical impact.

The design focused on documentation and implementation assessment, and not statistical generalisation. Lexical gaps and transfer-based errors were first defined in the corpus analysis of the learner's writing, which informed the intervention's instructional materials. Follow-up classroom results showed pre- and post-writing outcomes, supplemented by teacher and student comments. The combination of these sources ensured descriptive validity and provided an overview of how corpus-informed academic collocations were targeted in different conditions.

3.2 Participants and setting

The research was conducted in English-language classes at tertiary institutions. The selection of classes was done in consultation with the administrators, based on consent and logistical feasibility. Table 1 presents the total number of enrolled students and those who have completed the pre- and post-writing activities. The inclusion criteria were that the participant had to be enrolled in a mandatory English course at the time of the study; there was no random assignment or control group. The entire teaching was conducted during regular classroom sessions and in accordance with each institution's existing curriculum.

3.2.1 Ethical considerations

Implementation was done with institutional approval and teacher consent. The purpose of the research was explained to the students, and they were promised that

Table 1: Overview of participants and institutional contexts.

Institution type	No. of classes	Approx. No. of students (completed pre- and post-tasks)	Year level	Course type	Notes on context
Public university (urban)	3	85	1st year	Compulsory English	Large mixed-ability cohorts; limited computer access.
Private university (urban)	5	142	1st–2nd year	Compulsory English	Better ICT access; flexible curriculum allowed pilot implementation.
Affiliated college (semi-urban)	4	96	1st year	Foundation English	Traditional syllabus; emphasis on grammar and translation.
Technical institute (urban)	2	53	Diploma 1st year	English for Academic Purposes	Smaller classes; greater focus on practical writing.
Total	**14**	**376**	–	–	–

All classes participated voluntarily under institutional permission. Numbers represent approximate counts of students who completed both the pre- and post-writing tasks.

their grades would not be affected. To ensure confidentiality and ethicality, all the samples of the writing were anonymised before the analysis.

3.3 Corpus development

In this study, two complementary corpora were created: one a learner corpus indicating student writing, and the other a reference corpus indicating the academic language observed in the local instructional materials. The corpus of learners was estimated to contain 285,000 authentic student writing tokens gathered over 18 months during regular classroom activities. These were academic essays, exam answers, and writing tasks, the typical genres of Bangladeshi tertiary education.

The reference corpus consisted of approximately 420,000 tokens, selected from locally available textbooks, nationwide examination papers, and university-level assignments. This ensured that the corpus was realistic and reflected what the students are exposed to in their coursework. This localised corpus also emphasised contextual relevance, unlike foreign academic corpora, which would not align with local standards.

Preprocessing and analysis of all texts were performed using spaCy 3.7.2 (tokenization, lemmatization, and part-of-speech tagging) and AntConc 3.5.8 (frequency analysis and collocation analysis). A 5 % sample was manually validated to assess tagging accuracy. Several cross-checks were conducted during corpus construction to improve reliability. Although this style focused more on local authenticity, it also meant less access to broader global academic discourse. This weakness is recognised and elaborated in Section 5.4.

3.4 Framework development

The teaching model described in this paper integrated corpus-informed principles with class-based constraints typical in Bangladesh EFL settings. It comprised three components in relation to each other:

(1) **The Bangladesh Academic Collocation List (BACL).** As the lexical basis of the intervention, a localised academic collocation list was gathered. Based on the learner and reference corpus (Section 3.3), statistical filtering was used to identify high-frequency collocations (Mutual Information ≥ 3; minimum frequency = 5). The resulting BACL suggested 800 verb-noun and adjective-noun combinations commonly found in academic discourse, but they were somewhat challenging for local learners.

(2) **The predictive model of L1 transfer errors.** To address the prevalent lexical and collocational errors resulting from L1 influence, a predictive model was developed by comparing the repetitive patterns in the writing of Bangladeshi students with those of native English academic language. The categories of mistakes (e.g., omission of a preposition, miscollocation, overgeneralization) were hand-coded and used to create specific educational resources. This model allowed teachers to predict the types of errors and to draw attention to the instruction.

(3) **Protocol of Implementation and Teacher Support.** An instructional protocol helped teachers follow every instructional stage. It featured weekly lesson models, clear instructional objectives, and peer-reviewing tools for students. Teachers were given brief orientation training on how to use the corpus and on the pedagogical reasoning behind the tasks. The framework focused on manageability-the teachers were invited to match examples and pacing to their own syllabus without compromising the integrity of the core instructional cycle.

The combination of the three elements created a context-dependent system that could be replicated to integrate the teaching of corpus-informed language knowledge with regular curricula. The complete framework materials, including sample collocation lists, weekly guides, and coding templates, are presented in the Appendices.

3.5 Intervention procedures

The vocabulary programme based on the corpus was completed in 36 weeks, with two 45-minute lessons per week (about 54 hrs total). The teaching cycle incorporated corpus analysis, guided discovery, as well as production-based writing tasks. The Bangladesh Academic Collocation List (BACL), created for this study, was used to select 10 academic collocations per week.

3.5.1 Lesson sequence

The lessons every week were based on a five-stage cycle:
(1) **Retrieval (5 mins**): a brief quiz on collocation practised during the last week.
(2) **Guided discovery (15 mins):** students were asked to examine 6–8 concordance lines (per target item) to identify part-of-speech patterns and everyday-use contexts.
(3) **Patterning (10 mins):** learners were provided with collocation grids containing verb-noun and adjective-noun pairs, along with the most common grammatical partners.
(4) **Production (10 mins):** the students created short paragraphs or mini-tasks with at least six target collocations.
(5) **Feedback (5 mins):** The teacher checked the most frequent mistakes and spoke to the class about more natural examples of phrases.

3.5.2 Adaptation and preparation of teachers

The teachers had brief weekly instructions with lesson objectives, sample concordance lines, and answer keys. It was approximately 20 mins a week of preparation. The teachers were advised to revise sample sentences to fit class themes, while maintaining the emphasis on the specified collocations.

3.5.3 Homework and follow-up

Students were asked to write a 150–200-word paragraph, applying at least eight relevant collocations as target words, as homework. In the next session, they shared

draft and peer-review checklists focused on collocation and grammatical correctness. These paragraphs were given to teachers for brief written feedback.

All the participating classes used the same instructional model to ensure consistency, though they could make slight contextual adjustments. All students' detailed materials and worksheets are provided weekly in the Appendices to ensure transparency and replicability.

3.6 Outcome measures

Three outcome indicators were selected to report changes in students' written language after 36 weeks of intervention. Operationalization was used to define each measure, enhancing transparency and replicability.

(1) **Academic vocabulary coverage.** The student's writing was analysed in terms of the percentage of tokens selected from the known academic word lists (e.g., the Academic Word List and the Bangladesh Academic Collocation List). The coverage was measured as a percentage of total words, indicating the extent to which learners incorporated academic vocabulary during a lengthy writing task.

(2) **Collocation accuracy.** Correctness was measured using a predetermined set of verb-noun and adjective-noun collocations that were focused on during teaching. Each case was categorised as correct, wrong, or half correct based on grammatical and semantic correctness. The scores were reported as the number of accurate collocations per 1,000 words.

(3) **Formulaic sequence use.** The number of times multiword phrases and sets of academic words (e.g., as a result of, in terms of, it is essential to note that) appeared was divided by 1,000 words. This was an action that demonstrated productive application of ready-made academic expressions by the learners.

3.6.1 Coding and reliability

Reliability was assessed by coding a fifth of the writing samples by two trained raters, who showed high agreement ($k = 0.82$). Differences were resolved through discussion, and the remaining data were coded together.

3.6.2 Data treatment

Each indicator was computed based on descriptive statistics (means, standard deviations, medians, and change scores). There were no inferential analyses because the study aimed to evaluate feasibility and observable trends rather than test for

Table 2: Summary of outcome measures.

Outcome indicator	Operational definition	Unit of measurement	Data source	Analytical focus
Academic vocabulary coverage	Proportion of tokens from recognized academic word lists (e.g., AWL, BACL)	% of total words	Student pre-/ post-writing samples	Extent of academic vocabulary use
Collocation accuracy	Correct use of verb–noun and adjective–noun collocations targeted in instruction	Accurate collocations per 1,000 words	Student pre-/ post-writing samples	Precision of lexical combinations
Formulaic sequence use	Frequency of multi-word academic expressions (e.g., *as a result of*)	Occurrences per 1,000 words	Student pre-/ post-writing samples	Range of formulaic language use

All measures were analysed descriptively; no inferential statistics were applied due to the study's feasibility design.

statistical significance. Before analysing, all data were anonymised. In a bid to ensure transparency in the methods, Table 2 describes the operational definitions, data sources, and the analysis focus for each outcome measure.

3.7 Data analysis

A combination of thematic analysis and descriptive statistics was used to analyse data from student writing, teacher interviews, and student focus group discussions.

3.7.1 Quantitative data

One sample comprised pre- and post-writing and was analysed in relation to each of the three outcome measures: academic vocabulary coverage, collocation accuracy, and formulaic sequence use. Mean scores, standard deviations, and change values were computed on every indicator to reveal the apparent trends as time passed by. Due to the exploratory, single-group design, inferences were not based on inferential statistics. Instead, the focus was on discussing practical changes that could inform further large-scale testing.

3.7.2 Qualitative data

Teacher interviews and student focus group transcripts were coded thematically to examine perceptions of the intervention's practicality, classroom engagement, and perceived language development. There was an inductive derivation of codes and their refinement through repeated readings until they became saturated. The

representative quotes were chosen to reflect the main themes, including teachers' adaptation strategies, student motivation, and the observed lexical improvement.

Triangulation of quantitative and qualitative results was then used to provide a holistic picture of the programme's feasibility, instructional clarity, and possible pedagogical worth. All data sources and their analysis procedures are summarized in Table 3, which shows how the quantitative and qualitative strands were combined to achieve complete triangulation.

Table 3: Summary of data sources and analytical methods.

Data source	Description & scope	Type of data	Analytical method	Primary purpose
Student writing samples	Pre- and post-intervention essays collected from all participating classes (376 students)	Quantitative + Qualitative	Corpus-informed lexical analysis using *spaCy 3.7.2* and *AntConc 3.5.8*; manual coding of collocations and formulaic sequences	Examine observable lexical and collocational development.
Teacher interviews	Semi-structured interviews with participating instructors (12 teachers)	Qualitative	Thematic coding to identify perceptions of feasibility, workload, and instructional value	Explore pedagogical practicality and teacher perspectives.
Student focus-group discussions	Small-group discussions with representative students from each institution (48 participants)	Qualitative	Inductive thematic analysis of engagement, motivation, and perceived benefits	Capture learner experience and attitudinal change.
Institutional documents & course materials	Syllabi, lesson plans, and sample tasks from participating institutions	Documentary	Content analysis comparing local curricular language with corpus-derived targets	Ensure contextual alignment and material validity.
Researcher field notes	Observational notes recorded during implementation and feedback sessions	Qualitative	Reflective content analysis and cross-comparison with participant data	Document classroom dynamics and instructional feasibility

The integration of these quantitative and qualitative strands provided a triangulated understanding of classroom feasibility and language development within the corpus-informed instructional framework.

3.8 Ethical considerations

The research followed institutional and international ethics of education research. No formal permission was obtained from an Institutional Review Board (IRB) because the research posed little risk and was conducted in the regular classroom setting. The involvement was voluntary; no experimental interventions or sensitive personal information were gathered.

Participating institutions and course instructors gave their consent before taking part. Students were informed of the study's purpose, the confidentiality of their responses, and their right to withdraw from the study at any time without penalty. Before data collection, written consent from the teachers and verbal consent from the students were obtained.

Any analysis of all writing samples and interview transcripts was carried out anonymously, and any identifying information was omitted. The research was conducted in accordance with the ethical standards of the British Educational Research Association (BERA) Ethical Guidelines to Educational Research (2018) and conventional procedures of transparency, confidentiality, and voluntary involvement in a classroom-based research.

3.8.1 Supplementary materials overview

All supporting materials for this study are provided in the Appendices to enhance transparency and replicability. These materials offer in-depth illustrations of the teaching plan, educator support, qualitative tools, and language information that underlie the documented conclusions. In particular, the sample lesson framework utilised during the intervention is provided in Appendix A, and the materials for teacher orientation are discussed in Appendix B. The question guides to be used in interviews and focus groups are provided in Appendix C. Appendix D contains the sample excerpts on the learner corpus and the error-coding scheme applied in the analysis. Simultaneously, Appendix E presents a sample entry from the Bangladesh Academic Collocation List (BACL) compiled in this project. All these appendices, combined, provide the methodological openness of a corpus-informed classroom study and allow the study to be reproduced within the frameworks of other EFL settings.

4 Results

The findings of the research are shared in two sections: 1) quantitative observations of lexical development that could be observed regarding pre- and post-writing samples, and 2) qualitative observations representing the teacher interview and student focus-group discussion. Combined, the findings highlight the practicality and potential pedagogical importance of corpus-informed vocabulary instruction in Bangladeshi EFL classrooms.

4.1 Quantitative descriptions of language development

The pre- and post-writing samples were compared descriptively and showed improvement across the three measured indicators. There was a significant change in the percentage of academic-word tokens in written language; more have used target collocations correctly and more often used formulaic academic language. Even though no statistical testing was performed, the descriptive patterns indicate a slow increase in academic vocabulary control and collocational competence over the 36 weeks.

These findings were supported by teachers' classroom notes and student feedback, which indicated a better understanding of collocational patterns and greater confidence in applying academic vocabulary. Teachers also claimed they used fewer direct translations of Bangla and less repetitive or generic vocabulary.

This descriptive trend was confirmed in a review of sample essays. Indicatively, the common combinations of lexical constructions, e.g., make research or give suggestions, were used in typical sentences before the intervention. Post-intervention samples, in contrast, also included more target-like utterances, such as conducting research and offering tips. Moreover, learners generated a greater number of discipline-appropriate phrases (e.g., "play a vital role", "conclude"), providing evidence of exposure to and internalisation of more academic collocations.

4.2 Qualitative insights from teacher and student feedback

Thematic analysis of teacher interviews and student focus-group discussions found a few common themes:

(1) **Relevance of activities of corpus perceived:** Both groups ranked the authenticity of examples based on a real academic situation as more significant than the list of examples in a textbook.

(2) **Greater learner engagement:** Students said the discovery-based teaching method made vocabulary learning less robotic and prompted them to pay closer attention to collocations when reading.

(3) **Practical deliverability:** Teachers outlined the materials to be manageable, with an average of 20 mins per week of preparation time, which they stated they could manage.

(4) **Evidence of writing improvement:** Teachers reported fewer inaccurate lexical choices and less unnatural phrasing in the final assignments, consistent with the writing analysis.

Teacher and student statements also supported these quantitative observations. As one teacher remarked, students begin paying attention to collocations even in the reading materials, pointing them out during the discussion. A student also found

Table 4: Summary of key results.

Outcome indicator	Descriptive trend	Representative evidence	Interpretation
Academic vocabulary coverage	Moderate increase in proportion of academic tokens in post-writing samples	Pre-sample: *Students often relied on everyday vocabulary (e.g., "do research").* Post-sample: *Students employed more academic verbs and nouns (e.g., "conduct research," "provide evidence").*	Indicates improved awareness and use of academic lexis through corpus exposure and guided discovery.
Collocation accuracy	Noticeable reduction in miscues and greater precision in collocation use	Teachers observed fewer Bangla-influenced collocational errors; peer-review logs showed more accurate verb–noun combinations.	Suggests learners internalized target patterns through repeated noticing and production tasks.
Formulaic sequence use	Increased frequency of multi-word academic expressions	Students integrated recurring bundles (e.g., *"as a result of"* and *"on the other hand"*) in written tasks.	Reflects enhanced fluency and awareness of discourse-organizing phrases.
Teacher perceptions	High acceptability and perceived manageability	Teachers reported ≈ 20 mins prep time per week; valued materials' adaptability.	Confirms pedagogical feasibility and alignment with local curricular constraints.
Student perceptions	Positive engagement and increased lexical confidence	Focus-group comments highlighted the enjoyment of "discovering words in context".	Demonstrates motivational impact and perceived learning relevance.

Trends are descriptive and based on aggregated classroom data; no inferential testing was performed due to the study's feasibility design.

himself saying, "Now I understand why we say conduct research, not make research, because I noticed that it was repeated in the corpus lines". These remarks imply increased metalinguistic awareness and enhanced involvement in natural language patterns.

The overall findings suggest that the intervention was pedagogically viable, popular among participants, and capable of generating observable lexical growth within current curricular limitations. A summary of the trends and qualitative findings is shown in Table 4, which indicates the main results of the intervention and their pedagogical explanations.

5 Discussion

The current study explored the practicality and teaching feasibility of a corpus-informed vocabulary instruction model for Bangladeshi EFL students. Instead of measuring statistical effectiveness, the project aimed to document and evaluate a practical implementation framework that could improve the teaching of academic vocabulary in resource-limited settings. The study sought to demonstrate how corpus-informed methods can be realistically applied in local institutional contexts by combining three elements: the Bangladesh Academic Collocation List (BACL), a predictive model of common L1-influenced errors, and a teacher development protocol. The results suggest that the intervention was manageable and pedagogically feasible. Descriptive data and participants' responses indicate visible improvements in vocabulary breadth, collocational awareness, and the use of formulaic language. These findings should be viewed as initial signs of feasibility and pedagogical potential rather than proof of causal effectiveness. The subsequent sections address how the research responds to the research questions, how the results align with existing literature, and the implications for future research and teaching practice.

5.1 Addressing the research questions

The findings directly address the three research questions guiding this study.

Regarding RQ1, the study identified specific vocabulary patterns that characterize the academic writing of Bangladeshi learners in contrast with target academic norms. The pre-intervention analysis revealed several recurring issues: the overuse of high-frequency, general-purpose vocabulary; frequent collocational errors influenced by L1 transfer; and the limited use of formulaic academic expressions. These

features collectively highlight the lexical gap between learners' writing and standard academic English. Such trends are consistent with previous findings in EFL contexts, underscoring the persistent challenge of developing lexical sophistication in education systems that emphasize results over language depth (Nation 2013).

For RQ2, the study outlined practical steps for implementing a corpus-informed vocabulary framework in Bangladesh's educational setting. The intervention demonstrated that corpus-informed tasks and materials can be feasibly and meaningfully designed and applied even in resource-limited environments. Teachers were trained to interpret collocational data and integrate it into instruction, while learners completed writing tasks derived from locally relevant corpora. Improvements in lexical coverage, collocational accuracy, and phraseological variety indicate that the framework appeared to support the intended learning outcomes – enhancing learners' ability to produce more academically appropriate writing.

With respect to RQ3, the investigation established implementation protocols that enable teachers to sustain corpus-informed vocabulary instruction independently. The training component played a pivotal role in building educators' confidence and competence in applying corpus insights in their own classrooms. The model required minimal additional resources and proved sustainable after initial training. Furthermore, the potential for inter-institutional scalability appears promising, though broader trials are necessary to validate its feasibility on a national scale.

5.2 Integration of quantitative and qualitative evidence

Although the findings are primarily descriptive, their validity is reinforced by convergence across multiple data sources, by the overlap between textual data and the study participants' participation. It was observed that students were often correcting their incorrect collocations, such as "make research", with standard words like "conduct research", and this was also reflected in the learner corpus. Students also said they felt equipped to organize their writing better through learning formulaic expressions, such as "on the other hand" and "it should be noted that". These observations correlate with the diversity of phrases in learners' texts.

The consistency of the triangulation of evidence based on text analysis, teacher observations, and student reflections enhances the reliability of the findings. Due to the subjectivity inherent in single-method analysis in applied linguistics research, methodological triangulation, as described by Dörnyei (2007), serves as a protective

measure against this subjectivity. This work addresses the limitations of failing to present detailed statistical indices by demonstrating convergence across several sources.

5.3 Relevance to existing research and pedagogical contribution

The findings of this feasibility study align with and extend prior research on corpus-informed pedagogy. The Academic Word List, created by Coxhead (2000), demonstrated the utility of corpus-informed lexical resources for academic literacy, and the Academic Vocabulary List by Gardner and Davies (2014) highlighted the need to update word lists to account for disciplinary variation. This work continues the tradition and further develops the Bangladesh Academic Collocation List (BACL), a localised resource that reflects the lexical patterns used in Bangladesh's academic environment.

This enhancement in collocational accuracy is consistent with Nesselhauf's (2005) claim that the primary issue for L2 learners with collocations is negative transfer and that collocational patterns cannot be improved without explicit exposure to them. Likewise, Durrant and Schmitt (2010) also emphasize the pedagogical significance of corpus-informed collocation teaching. This research paper contributes to the current body of research by demonstrating that such practices can be successfully translated in a low-resource setting.

The increase in the number of formulaic sequences used confirms the argument by Wray (2002) that formulaic language is the focus of fluency and discourse coherence. Although the concept of phraseology has been employed numerous times in the past, particularly in more resource-intensive or technology-intensive contexts, the current results suggest that even simple corpus-informed exercises can aid learners in acquiring multi-word phrases, which, in turn, facilitate the comprehension of academic texts.

The study demonstrates that corpus-informed pedagogy can be effectively localised without losing theoretical rigour, thereby expanding the global evidence base for context-sensitive CALL practices.

5.4 Limitations and directions for future research

The research has several limitations that cannot be ignored, despite its positive results. On the one hand, the reference corpus was primarily compiled from locally commissioned materials, including textbooks, examination scripts, and student

assignments. This was contextually appropriate, but it may have limited access to the broader scope of academic vocabulary found in international corpora. Therefore, students can remain unfamiliar with idiomatic or globally standardised patterns of use. The subsequent study may test the hybrid type of corpora, such as local and international sources, to balance relevance and introduction in the global academic discourse. Second, the researchers did not administer a delayed post-test to evaluate long-term retention. Although immediate gains were realised, it is unclear whether they would persist over time without constant exposure. A longitudinal study with a delayed post-test would also help determine the sustainability of corpus-informed teaching. Third, even though teacher training was practical in the institutions under analysis, the model has not been tested at scale. The implementation may be affected by variability across institutions in resources, teacher preparedness, and curriculum needs. Future research may investigate how flexible the framework is under various institutional conditions, such as rural schools with limited infrastructure. Lastly, the fact that this article qualitatively presents its findings is a drawback, as it lacks descriptive statistical measures. The goal of future work should be to present quantitative estimates of effect sizes and levels of significance to increase the external validity of the findings. Despite these constraints, the study provides a transparent foundation for future controlled trials that could quantify the observed descriptive patterns.

5.5 Pedagogical implications

Despite these limitations, the findings hold meaningful implications for the practice of EFL teaching in low-resource contexts. To begin with, the paper demonstrates that corpus-informed approaches do not require costly technology or specialized software. Properly selected resources, such as the BACL, can help a teacher guide learners to identify real academic patterns using classroom resources. This is particularly important in situations such as Bangladesh, where technological barriers often hinder the adoption of innovation. Second, the findings emphasise the key importance of teacher training. Teachers not only adopted practices with corpus information, but also maintained them without intervention. This implies that teacher education in corpus linguistics should be incorporated into professional development programs, enabling instructors to innovate in their classrooms. Third, the framework also works towards the overall objective of educational equity. The intervention enabled students to access academic language sources that were generally not available in their learning contexts by adapting corpus-informed pedagogy to their local needs. This can help decrease the gap in educational

performance and increase learners' preparedness for higher education and international communication.

6 Conclusions

In the study, a corpus-informed vocabulary-teaching model is proposed and tested in Bangladesh to address the long-standing problem of students' writing of academic texts. It combines the Bangladesh Academic Collocation List (BACL) with a predictive error model and sustainable teacher-training procedures, thereby offering a context-based, internationally applicable framework for vocabulary instruction. These findings demonstrate that learners who have undergone the intervention have recorded noticeable gains in the three domains: academic vocabulary coverage, collocational accuracy, and the sequence use of formula. The support for these improvements is secured through teacher observations and student reflections, which focus on both increased confidence and enlightenment regarding academic discourse practices. Putting the results together, they suggest that corpus-informed teaching can be practically implemented in low-resource environments, yields immediate positive outcomes for learners, and does not render the practice unfeasible for teachers.

Also, several limitations should be noted. This was offered through contextual relevance in the form of a locally-created reference corpus, but at the cost of exposure to international, standardised academic use. It also questions the long-term benefits of retention, since the delayed post-test does not measure long-term retention. Furthermore, although the framework was found to be effective in the institutions studied, further research is needed to determine whether it can be scaled to other, more diverse educational environments.

Although this research was exploratory, its transparent methodology and localised structure provide a template that can be replicated in the same EFL context to pursue evidence-based vocabulary teaching. Future studies can be based on a hybrid corpus that uses both local and international resources, longitudinal analysis to determine retention, and testing the framework's elasticity in rural or low-resource institutions. Such a work would contribute to understanding the extent to which scaling up corpus-informed pedagogy can be achieved without losing its effectiveness.

To sum up, the study shows that corpus-informed vocabulary teaching can help overcome the gap between learners' current practices and the needs of academic speech. It is the structure that offers a path towards greater academic literacy and

educational equity in Bangladesh and other EFL environments, providing teachers with valuable resources and students with natural lexical input.

Appendix A: Sample Lesson Framework (Corpus-informed Vocabulary Cycle)

The instructional model followed a five-stage weekly cycle integrating corpus analysis, guided discovery, and writing practice. Each 45-minute session targeted ten academic collocations drawn from the Bangladesh Academic Collocation List (BACL).

Stage 1 – Retrieval (5 mins)
Short quiz reviewing previous collocations.
Example: Complete: "_____ research," "provide _____," "play a _____ role".
Stage 2 – Guided Discovery (15 mins)
Students examine concordance lines to infer patterns.

conduct research on social issues
conduct further investigation
conduct a survey among respondents

Task: Identify grammatical role and common objects of *conduct.*

Stage 3 – Patterning (10 mins)
Students complete a collocation grid.

Verb	Noun	Example sentence
Conduct	Research	The team conducted research on academic writing.
Draw	Conclusion	The writer drew a valid conclusion.
Provide	Evidence	The report provides evidence of progress.

Stage 4 – Production (10 mins)
Students write a 150-word paragraph using at least six target collocations.
Stage 5 – Feedback (5 mins)
Teacher reviews typical errors and presents improved phrasing.

Appendix B: Teacher Orientation and Support Materials

Teachers received a concise 90-minute orientation including:
–   Rationale for corpus-informed vocabulary instruction
–   Demonstration of AntConc concordance use
–   Overview of the BACL list
–   Strategies for guided discovery and peer review
–   Distribution of support materials:
    –   Weekly plans and collocation lists
    –   Sample concordance printouts
    –   Peer-review/self-check forms
    –   Reflective teaching log

Average preparation time: **20 mins per week**.
Teachers rated the protocol as feasible for large, mixed-ability classes.

Appendix C: Interview and Focus-Group Question Guides

Teacher interview prompts
1)   How useful were the corpus-informed activities?
2)   What challenges arose during implementation?

3) How did students respond to collocation-focused tasks?
4) Did the approach affect lesson planning or time management?
5) Could this model be scaled up across institutions?

Student focus-group prompts

1) What was your experience using corpus-based materials?
2) Did the activities help you write more accurately or confidently?
3) Which part of the lessons was most/least helpful?
4) How did peer-review or feedback influence your writing?
5) Would you like similar tasks in future English courses?

Appendix D: Sample Learner Corpus Excerpts and Error-coding Scheme

1) Pre-intervention excerpts
– "We make research on the topic and find the solution".
– "The teacher give us suggestion to write better".

2) Post-intervention excerpts
– "We conducted research on the topic and identified possible solutions".
– "The teacher provided suggestions to help us revise the essay".

3) Error-coding categories

Code	Description	Example	Correction
L1T	L1 transfer error	*do research*	*conduct research*
COL	Collocational misselection	*give suggestion*	*provide suggestion*
FORM	Formulaic omission	Missing *as a result of*	Inserted bundle
GRM	Grammatical misformation	*students is improving*	*students are improving*

4) Reliability

Two raters independently coded 20 % of data (κ = 0.82). Discrepancies resolved through discussion.

Appendix E: Bangladesh Academic Collocation List (BACL) – sample extract

The BACL was built from combined learner and reference corpora (~705,000 tokens) using Mutual Information ≥ 3.0 and minimum frequency = 5. Manual filtering removed irrelevant items. The full list includes 800+ verb–noun and adjective–noun pairs frequent in Bangladeshi academic writing (Table E1).

Table E1: Sample extract from the Bangladesh Academic Collocation List (BACL).

Verb–noun collocations	Adjective–noun collocations	Example in context
Conduct research	Academic achievement	Students conducted research on social issues.
Draw conclusion	Educational outcome	A clear conclusion was drawn from the findings.
Provide evidence	Theoretical framework	The report provides evidence of improvement.
Play role	Linguistic feature	Teachers play a vital role in vocabulary development.
Address issue	Pedagogical approach	The study addressed the issue of L1 transfer.
Make recommendation	Grammatical pattern	The committee made several recommendations.
Propose model	Cognitive process	The paper proposed a model for effective feedback.
Collect data	Communicative competence	Data were collected through student essays.
Analyse result	Lexical choice	Results were analysed using corpus tools.
Design questionnaire	Instructional strategy	The researcher designed a questionnaire for feedback.
Offer insight	Academic discourse	Findings offer insight into learners' writing habits.
Reach agreement	Pragmatic competence	The class reached agreement on key terminology.
Establish link	Research gap	The study established a link between exposure and output.
Highlight importance	Writing proficiency	Teachers highlighted the importance of collocation use.
Overcome challenge	Curriculum design	The model helped overcome implementation challenges.

The complete BACL dataset is available upon reasonable request from the corresponding author. This extract demonstrates the contextual relevance used in instructional materials.

References

Ackermann, Kirsten & Yu-Hua Chen. 2013. Developing the academic collocation list (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12(4). 235–247.

Boulton, Alex & Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67(2). 348–393.

Coxhead, Averil. 2000. A new academic word list. *Tesol Quarterly* 34(2). 213–238.

Dörnyei, Zoltán. 2007. *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Durrant, Philip & Norbert Schmitt. 2010. Adult learners' retention of collocations from exposure. *Second Language Research* 26(2). 163–188.

Flowerdew, Lynne. 2012. *Corpora and language education*. London: Palgrave Macmillan.

Gardner, Dee & Mark Davies. 2014. A new academic vocabulary list. *Applied Linguistics* 35(3). 305–327.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. New York: Routledge.

Kennedy, Claire & Tita Miceli. 2017. Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language, Learning and Technology* 21(1). 114–130.

Kremmel, Bernhard, Tineke Brunfaut & J. Charles Alderson. 2017. Exploring the role of phraseological knowledge in foreign-language reading. *Applied Linguistics* 38(6). 848–870.

Laufer, Batia & Jan Hulstijn. 2001. Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics* 22(1). 1–26.

Nation, Paul. 2013. *Learning vocabulary in another language*, 2nd ed. Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Odlin, Terence. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Rahman, Mohammad Mosiur. 2020. English language teaching in Bangladesh today: Issues, outcomes and implications. *Language Testing in Asia* 10(1). 1–14.

Rahman, Mohammad Mosiur & Ambigapathy Pandian. 2018. A critical investigation of English language teaching in Bangladesh: Unfulfilled expectations after two decades of communicative language teaching. *English Today* 34(3). 43–49.

Read, John. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Ringbom, Håkan. 2007. *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Wall, Dianne & J. Charles Alderson. 1993. Examining washback: The Sri Lankan impact study. *Language Testing* 10(1). 41–69.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Bionotes

Shajadul Alam Sweet
Department of English, University of Asia Pacific, Dhaka, Bangladesh
shajadulalamsweet@gmail.com
https://orcid.org/0009-0005-0994-7502

Shajadul Alam Sweet is an undergraduate student in the Department of English at the University of Asia Pacific, Dhaka, Bangladesh. His research focuses on corpus linguistics, applied linguistics, vocabulary acquisition, and computer-assisted language learning. He has published several papers examining the interface between linguistic theory and pedagogy, especially within EFL contexts in South Asia. His current

research explores corpus-informed frameworks for improving academic literacy and vocabulary instruction in Bangladesh. He has also served as a peer reviewer for two international journals, including one indexed in Scopus, and remains committed to advancing equitable, data-driven language education.

Md. Nurul Kabir Emon
Department of English, University of Asia Pacific, Dhaka, Bangladesh
nurulkabiremon@gmail.com
https://orcid.org/0009-0003-5680-5975

Md. Nurul Kabir Emon is an undergraduate student in the Department of English at the University of Asia Pacific. His research interests lie in applied linguistics, second language vocabulary development, and the pedagogical application of corpus tools in EFL settings. He is particularly interested in how data-driven learning and collocational analysis can support the teaching of academic English in resource-limited contexts. His recent projects involve developing contextually relevant teaching materials aligned with corpus-informed principles to improve lexical awareness among tertiary-level learners in Bangladesh.

Sadia Khandokar Mim
Department of English, University of Asia Pacific, Dhaka, Bangladesh
khandokarsadia63@gmail.com
https://orcid.org/0009-0006-9492-3544

Sadia Khandokar Mim is an undergraduate student in the Department of English at the University of Asia Pacific, Bangladesh. Her academic work explores language pedagogy, vocabulary acquisition, and learner autonomy within the Bangladeshi EFL context. She is passionate about designing engaging classroom materials that promote contextual vocabulary learning and the use of authentic language input. Sadia has participated in several collaborative studies on corpus-informed instruction and lexical development. Her research also extends to sociolinguistic perspectives on English language education and gender inclusivity in language learning environments.

Fakhrul Islam Mahim
Department of English, University of Asia Pacific, Dhaka, Bangladesh
fakhrulmahim@gmail.com
https://orcid.org/0009-0002-5073-9981

Fakhrul Islam Mahim is an undergraduate student in the Department of English at the University of Asia Pacific, Dhaka. His research interests include corpus linguistics, EFL vocabulary teaching, and digital literacy in language learning. He has contributed to studies on corpus-informed materials design and its impact on learner engagement. His work focuses on developing innovative classroom approaches that balance linguistic accuracy with communicative fluency. Mahim is committed to advancing research-based teaching practices that support equitable access to high-quality English-language education in Bangladesh.

Mahfuj Hosen
Department of English, University of Asia Pacific, Dhaka, Bangladesh
21216056@uap-bd.edu
https://orcid.org/0009-0001-1867-6995

Mahfuj Hosen is an undergraduate student in the Department of English at the University of Asia Pacific. His primary research areas include applied linguistics, corpus-informed pedagogy, and vocabulary instruction for EFL learners. He is particularly interested in exploring how linguistic data can enhance teaching efficiency and learner independence. Mahfuj's recent projects focus on developing academic word lists tailored to the Bangladeshi higher education context. He aims to help bridge the gap between theoretical research and practical classroom implementation through innovative language-teaching methods.

Partho Biswas
Department of English, University of Asia Pacific, Dhaka, Bangladesh
parthobiswas049@gmail.com
https://orcid.org/0009-0006-5980-104X

Partho Biswas is an undergraduate student in the Department of English at the University of Asia Pacific. His research interests encompass corpus linguistics, second language acquisition, and the integration of technology in English language teaching. He has worked on projects focussing on data-driven learning approaches that enhance vocabulary and collocational competence among university students. His research aims to support pedagogical innovation by applying corpus data in low-resource educational contexts. Partho is also engaged in promoting collaborative teacher development and the adoption of modern language-learning technologies in Bangladesh.

Easmin Sultana
Department of English, Royal University of Dhaka, Dhaka, Bangladesh
e.s.jafrin@gmail.com

Easmin Sultana is a Lecturer in the Department of English at the Royal University of Dhaka, Bangladesh. Her teaching and research areas include applied linguistics, literature, gender studies, feminism, romanticism, curriculum design, and vocabulary pedagogy. She has been involved in several research projects exploring the role of corpus-based instruction in enhancing students' academic writing and lexical competence. Easmin advocates for bridging theory and classroom practice through localised pedagogical innovations. Her work aims to empower both teachers and learners by promoting data-informed teaching methods suited to the Bangladeshi EFL context.

Md. Mehedi Hasan Emon
Department of English, University of Asia Pacific, Dhaka, Bangladesh
mehedihasanemon2349@gmail.com
https://orcid.org/0009-0003-7016-4190

Md. Mehedi Hasan Emon is an undergraduate student in the Department of English at the University of Asia Pacific, Dhaka. His research interests include corpus linguistics, lexical studies, and applied linguistics pedagogy. He has contributed to collaborative work on vocabulary profiling and the use of learner corpora in language teaching. His research emphasizes the importance of contextually grounded instruction and the use of empirical linguistic data in improving learners' writing accuracy and vocabulary depth. He aims further to develop practical frameworks for sustainable corpus-informed teaching in Bangladesh.