ISSN: 2634-9221
European Journal of Applied Sciences; Vol. 14 No. 06 (2026) (34-49)
https://doi.org/10.14738/ejas.1406.12467

# Catapulting Vocabulary through Corpora and Corpus Tools

Roselin Jenifer, F. S. & S. J. Kala

PG & Research Department of English Fatima College, Madurai (Affiliated to Madurai Kamaraj University) Madurai, Tamil Nadu India

**Abstract:** The rapid advancement of Information and Communication Technology has led to the large-scale proliferation of data across educational domains. This accelerated development of ICT has paved the way for innovative language learning approaches such as Data-Driven Learning (DDL). Corpus-based tools allow learners and teachers to investigate recurring patterns in language, including frequency and collocation. This study analysed the effectiveness of corpus-based intervention in enhancing the vocabulary proficiency of beginner-level English learners, with particular focus on words from the Academic Word List (AWL). A pilot study was conducted to assess the existing vocabulary knowledge of 30 basic-stream English learners. The pilot study revealed a clear lexical gap and informed the design of corpus-based resources. The conventional prescribed instructional materials were then uploaded into VocabProfile, a corpus linguistic tool, to categorise vocabulary according to frequency bands and to identify AWL items. Targeted instructional materials were developed using these profiles featuring authentic language input and frequency-informed vocabulary activities. A pre-test–post-test design was used with the same cohort, with a two-week intervention between the tests. The paired t-test revealed a statistically significant improvement from pre-test to post-test, $t(29) = 10.00$, $p < .001$, with a large effect size (Cohen's $d = 1.84$), and the test instrument showed high reliability (KR-20 = 0.90). The results showed a statistically significant improvement in post-test scores, indicating that the corpus-informed materials enhanced learners' vocabulary acquisition, including AWL words. The study highlights the potential of corpus-based pedagogical materials to strengthen lexical competence among ESL learners.

**Keywords:** Corpus linguistics, frequency studies, Data-Driven Learning (DDL), vocabulary instruction, learner autonomy, corpus-based pedagogy

## INTRODUCTION

### Evolution of Pedagogical Methods

#### *Teacher-Centric Approach*

The historical analysis of the evolution of pedagogical materials is directly proportional to the needs and requirements of the society at a particular time. For instance, the industrial revolution that dominated during the 19th century demanded a larger educated workforce: "English was seen as a dangerous weapon" (Pennycook, 2007, p.1) and educational institutions during that time followed traditional pedagogical systems that resulted in large-scale production of learned masses. The social and political needs of that century favoured citizens who could read and understand basic text materials and contribute to the development of a nation. Therefore, schools and colleges adapted to an organised structure that could educate a large number of students efficiently.

The massification of education gradually permeated into India as a result of colonisation. The British government was in demand of a larger workforce that could

contribute to the British economy. Lord Macaulay's Minutes (1835) on the educational status of India calls out for an academic approach that creates an educated labour force. Macaulay opted for an educational system in which English was taught to Indians to appeal to the English government rather than providing a holistic proficiency to an individual. Even decades later, a pedagogical approach in which the teacher determines the needs of the students is widely popular throughout India.

The teacher-centric methodology of English language teaching proved to be efficient for the previous generations as it democratised education for larger populations: "The traditional teacher-centred approach focuses on the teacher as the expert in transmitting knowledge to the student, who is the novice" (Tengku Kasim, 2014, p. 200). The teacher-centric approach exhibited a larger student-teacher ratio as the flow of information has always been linear and unidirectional. It places the teacher at the centre of the educational process and as the primary source of knowledge: "Traditional methods, which are grammar-translation and direct instruction, place significant stress on rote learning and formal correctness, normally at the cost of communicative competence" (Hashim, 2025, p. 1). The passive receptive framework has been heavily relied upon for teaching English in India. Various pedagogical approaches like Grammar Translation Method, Audio-Lingual Method, Direct Method and Presentation, Practice, Production (PPP) method have been frequently used in the context of language learning.

### *A Shift towards Learner-Centric Approach*

The dawn of the twenty-first century along with its technological advancements had a profound impact on the prevalent educational materials as, "foreign language instruction is transitioning from traditional, teacher-centered approaches to more dynamic, learner-centric paradigms" (Alshraah, 2023, p. 191). The development in the communication and digital systems proliferated the distribution of information throughout the globe. The unbridled access of educational materials on the internet has altered the teacher-student dynamic: "…in English as a second or foreign language, the idea of a teacher as "facilitator" is considered to be more compatible than the teacher's role as a transmitter of knowledge to the students" (Gulnaz, 2015, p.935). The role of a teacher drastically transformed from being an authoritative figure to a facilitator who caters to the needs of the students.

English language teaching in India has gradually transformed from being teacher-centric to learner-centric owing to the technological progressions of the century. Learners of the current generation have a higher demand of communicative competence of a language rather than mere cognitive accumulation. Globalisation and digital communication require learners to collaborate and assimilate with the growing trends of contemporary educational landscape: "In India, even students at the university level often find it difficult to comprehend academic texts in English" (Barua, 2026, p. 1). A Learner-centric approach helps develop learner autonomy, encourage active learning and addresses individual differences. A student-centred approach is necessary in English language teaching as it assists in the acquisition of language through active use and meaningful interaction rather than by memorising rules.

Various types of learner-centric approaches have been experimented and successfully implemented for teaching English for ESL learners: "With regard to English language teaching, learner-centred pedagogy is very crucial for communicative language

teaching which stresses interaction, fluency, and application of language in a meaningful manner" (Nakazinga, 2026, p. 36). Some of these methods are Communicative Language Teaching (CLT), Task-Based Language Teaching (TBLT), Content and Language Integrated Learning (CLIL). These innovative teaching methods assist in language learning through multimodal contents that stimulate a holistic development of linguistic proficiency. Multimedia aids and digital platforms play a crucial role in promoting learner autonomy as they cater to the target needs of individual learners. In the context of contemporary education, "technology is no longer merely an optional supplement to classroom instruction; it forms part of the broader learning environment through which learners encounter language and communicate" (Orujlu, 2026, p. 19). The inclusion of Information and Communication systems (ICT) in education has demanded a curriculum that prepares students to be active participants of the digital era.

**ICT in Education**

ICT in education has provided opportunities for self-learning owing to its vast number of e-books, video tutorials and personalised learning applications: "The rapid expansion of this system has sparked a learning revolution, as new technological breakthroughs in education have required a re-examination of new teaching methodologies and instruments" (Hidayah, 2022, p. 40). The implementation of ICT enabled teaching materials in ESL classrooms, calls for an integrated development of English language learning. Several methods of ICT enabled teaching includes Computer-Assisted Language Learning (CALL), Mobile-Assisted Language Learning (MALL) and Project based Learning (PBL).

The ICT enabled tools help the students become active participants in classroom discussion and promote better cognitive development. The educational materials that are created using digital resources promote active learning habits owing to their multimodal contents:

> *They combine sounds, pictures/images and texts which draw the learners' attention or compel them to watch, listen and become engaged in the lesson. Similarly, multimedia use in classroom will provide the learners with opportunity for interacting with diverse texts (Sharndama, 2015, p. 35).*

The multimedia software programmes endorse flexibility in teaching grammatical rules by inculcating real-life situations that help student familiarise the language without any apprehension.

***Types of ICT Tools***

ICT tools encompass a wide range of digital technologies that facilitate the creation, storage and exchange of information. These utilities promote the learning process by enhancing the quality of language instruction:

> *Exploiting the web to its full potential for language learning is less a matter of technology than of methodology. Instead of being empowered by new technology, users may easily become overpowered, because of the vast range of options open to them. This is where learning management systems make a difference (Jager, 2004, p.45)*

Learning Management Systems (LMS) such as Moodle and Google Classroom enable language facilitators to manage and coordinate the process of learning through centralised online platforms.

Language learning applications such as Duolingo, Memrise, and Nibble provide interactive exercises through gamified learning experiences. These applications encourage self-paced learning and promote student engagement: “By incorporating game-like elements such as points, levels, and rewards, the platform encourages learners to practice consistently and maintain their motivation” (Jakob, 2026, p. 65). Digital assessment tools such as Google Forms, Kahoot! and Quizizz allow facilitators to create quizzes and receive immediate feedback on learner performance. Instant evaluation helps learners identify their strengths and weaknesses and encourages continuous improvement. Digital libraries and online repositories like National Digital Library (NDL) and JSTOR offer e-books and research articles, thereby enabling students to engage in extensive reading and academic research.

**Data-Driven Learning**

The final shift in the advancement of ICT in English Language Acquisition heralded with the advent of Data-Driven Learning (DDL). DDL tools use words from real world examples that are stored in corpora, the distinct databases. DDL is based on the principle that language learning is more effective when learners observe how language is actually used in real-life contexts: “Learner dictionaries, too, have been influenced by Corpora. Rather than using invented sentences, lexicographers began to turn to corpora to supply examples of real English to illustrate how words are used” (Dudeney, 2012, p.535). DDL models use morphological, syntactic and semantic rules to engage students’ interaction.

Tim Johns is credited with the coining of the term DDL and he asserts that, "At the heart of the approach is the use of the machine not as a surrogate teacher or tutor, but as a rather special type of informant” (2010, p.1). John is of the view that machines and software data acts not as teachers but as informants to the learners. He differentiates the teacher and an informant based on the flow of questions and answers. A teacher initiates the questions and conducts the answers, whereas, an informant is a passive listener and responds only when a question is asked by the learner. The usage of Corpora falls under the broader concept of Data-Driven Learning (DDL).The implementation of corpora has made educational resources accessible to both facilitators and learners. Corpus-based learning encourages students to discover linguistic patterns for themselves, thereby promoting learner autonomy and critical thinking.

**Corpus Linguistics**

A corpus is a systematically compiled electronic database of authentic language samples collected from books, newspapers and other real-world sources. Corpus linguistics uses computational tools to analyse large datasets and identify patterns that may not be evident through intuition alone: “Corpus linguistics can be defined as the study of the compilation and analysis of corpora, which are large collections of naturally occurring language texts that chosen to characterize a state or a variety of language” (Szudarski, 2018, p.4). The advent of digital technology and ICT has made corpora available to everyone, enabling their integration into ESL classrooms.

## REVIEW OF LITERATURE

### Innovative ICT Tools

The implementation of ICT tools in education sector has accelerated the learning process by simultaneously promoting innovative practices in digital technology. Arcade Nduwimana posits that,

> *The advances in ICT helped address the challenge of authentic materials necessary for autonomous learning. These advances provided a vast range of digital tools and resources … These tools and resources provide greater freedom and flexibility to learn at one's own pace and convenience" (2023, p. 697).*

The ICT tools thus ensure various methods of teaching and learning practices by which learners choose their own preferred modes of learning. It paves way for tailored and personalised teaching inputs that promote individual performance.

ICT enabled teaching materials prove to be fruitful as long as the facilitators have mastered the digital skills: "In this 21st century, there is a demand on everyone, including ELT teachers, to be technologically literate" (Abraham, 2022, p. 2). Students who are learning English as Second Language find greater comfort when they are taught using ICT enabled teaching materials: "teachers have begun adopting more innovative and learner-centered methods, such as real-life simulation activities, to develop the speaking confidence of learners" (Jasmih, 2026, p. 84). Nevertheless, advocating the digital tools proves to be challenging owing to the prevalent disparities in digital distribution:

> *Studying ICT's role in ELT is essential to identify effective strategies that empower teachers and students, ensuring technology benefits all learners, especially in resource-limited settings (Katel, 2024, p. 38).*

In an ICT enabled classroom, DDL resources like corpus tools enhance the interaction between the facilitator and the learner by providing curated digital contents that cater to the needs of all learners. The conception of DDL as an instructional approach encourages learners to operate as 'language detectives' who analyse the data presented: "Learners are not taught overt rules, but explore corpora to detect patterns among multiple language samples; this fits well with the constructivist paradigm for language learning" (Boulton, 2010, p. 535). Data-Driven Learning promotes pattern recognition skills among students through which they construct knowledge and experiment with the ways through which language is used in natural contexts: "DDL can refer to virtually any application of corpus linguistics to language learning, but it commonly involves use of keyword-in-context (KWIC) concordance lines, where each line shows the target word, phrase, or construction in context from authentic texts." (Inglis, 2026, 1). A learner who is accustomed to DDL masters the skills of predicting, observing and verifying the contents of the instructional material (Shirato, 2007, p. 394). The DDL approach therefore promotes autonomous language learning that expands beyond the horizons of classroom environment.

### Corpus Linguistics in Language Acquisition

Learners are exposed to 'real' language rather than a set of rules and regulations: "DDL draws from process teaching in that it sees grammar as a flexible system of recurring and

interrelated prototypes rather than a static set of rules" (Hadley, 2002, p. 107). This type of language learning becomes organic than the mechanistic regressions of conventional approaches. The usage of the Corpora and the concordances help students explore the linguistic and semantic usages of languages in corporeal situations: "DDL entails language learners working with written or spoken data, resulting in increased language sensitivity, noticing, induction, and ability to work with authentic data" (Pérez-Paredes, 2019, p.4). Corpora are the foundational tools of data-driven learning approach and are used in various levels of language learning and content preparation strategies. Corpora prove to be useful in grammar and vocabulary research, lexicography, translation studies, language change and historical studies, ethnographical studies and sociolinguistic research.

Corpus linguistics is thus defined as the systematic analysis of authentic texts that are collected and organised in specialised corpora in order to explain linguistic structure and promote the pedagogical performance. Lynne Flowerdew opines that, "Corpus linguistics is usually associated with a phraseological approach to analysis, which takes a syntagmatic, as opposed to a purely paradigmatic, view of language" (2009, p. 393). Corpus Linguistics acts as a catalyst in transforming linguistics and pedagogy into scientifically quantifiable field of studies: "In corpus linguistics, language study is always the study of written (or transcribed or quoted or otherwise recorded) texts or text pieces, i.e., language which can be reproduced, heard, read and interpreted repeatedly" (Teubert, 2005, p.3). Corpus Linguistics has made quantitative and qualitive approaches possible in language studies.

## Research Gap

A critical assessment into the existing literature on DDL and Corpus Linguistics substantiates corpus tools as efficient linguistic apparatus. It is found that corpus tools like Concordancers, readability analysers, complexity tools and syntactic analysers enhance language teaching. However, the implementation of corpus tools like Vocabulary Profilers and Frequency analysers to improve the vocabulary of students remain unexplored. This research paper, thus, focusses on enhancing vocabulary skills among ESL learners using corpus tools.

## Research Objectives and Questions

This research paper attempts to emphasise the significance of learner-centric teaching materials that enhance the active participation of students in classroom. It ventures to examine the impact of corpus linguistics in vocabulary acquisition and to identify its role in enhancing autonomous learning. This research paper endeavours to answer the following questions:

- **RQ1**- Do the instructional materials provided to the learners categorise and prioritise vocabulary based on their frequency
- **RQ2** –How familiar are the basic level learners with low frequency words
- **RQ3** – What are the ways to enrich the teaching materials to cater to the needs of the digital generation

The research has chosen a micro-corpus named the OpenSubtitles Corpora to analyse the needs of the students and current vocabulary proficiency of the learners.

### Research Hypothesis

This research paper is grounded in the assessment that vocabulary acquisition through corpus tools facilitate second language learning. This research hypothesises that ESL learners find it comparatively difficult to master certain words that are rarely used in their daily life. It also propounds that the educational materials used currently do not distinguish between what students know and what they need to know. Based on this premise, the study advocates that a more effective pedagogical approach designed with the help of corpus profilers would allow for prioritising low frequency words.

This research hypothesises that,

- **$H_0$:** There is no significant difference between the pre-intervention and post-intervention vocabulary scores of ELT learners induced by corpus linguistic tools
- **$H_1$:** There is a significant difference between the pre-intervention and post-intervention vocabulary scores of ELT learners induced by corpus linguistic tools.

### Research Methodology

This research paper employs a quantitative, longitudinal research and infers the data that has been collected through questionnaires. It aims at utilising statistical techniques and relies on observations implemented within controlled settings to ensure the accuracy and reliability of the gathered data. A Needs Analysis has been conducted in order to gauge the learners' lexical competence. The researchers have made use of the OPUS OpenSubtitles to frame objective questions from the movie that they have watched in the Language Lab, in order to analyse their prior knowledge of basic vocabulary. The research then uses VocabProfile to categorise the vocabulary introduced to the students.

## NEED FOR THE STUDY

### Purpose of Needs Analysis

Conventional pedagogical materials often rely on textual data that may not accurately reflect the actual language use. The intervention of Corpus linguistic tools provides empirical advantage over word frequencies, usage patterns and syntactic structure.

These enable the facilitators to create instructional materials that can identify the gaps between learners' current proficiency and target requirements. This analysis facilitates the development of the syllabi, selecting materials and creating learning activities that cater to the learners' specific needs.

### Participants

A Needs Analysis was conducted to understand the vocabulary proficiency of ESL learners from Fatima College, with the help of corpus-based instructional materials. A sample of 30

students from section 'A', the basic stream was selected in order to test the pre-existing knowledge of the learners. The participants were purposively selected due to their limited exposure to diverse English vocabulary in their everyday use.

### Material Development

The sample of 30 learners were made to watch the movie, ***Harry Potter and the Philosopher's Stone*** in the Language Lab. After watching the movie, the learners were given a questionnaire based on the script pf the movie procured form ***OPUS OpenSubtitles Corpora***. It consists of 15 questions with the words that were used in the movie. The questionnaire was organised as multiple-choice questions with four choices for each item. A total of 15 words were selected based on the categorisation of the corpus tool named ***VocabProfile***. The script of the movie is uploaded in the profiler and they were categorised into three major groups

- **K1 Words** – High frequency words that are used regularly in daily life
- **K2 Words** – Moderate frequency words that are comparatively less frequent
- **AWL Words (Academic Word List)** – Low frequency words that are mainly used in academic settings

### Data Collection

| | Families | Types | Tokens | Percent | Current profile | |
|---|---|---|---|---|---|---|
| | | | | | % | Cumul. |
| K1 Words (1-1000): | 619 | 822 | **4879** | **91.44%** | 91.44 | 91.44 |
| Function: | ... | ... | (2755) | (51.63%) | | |
| Content: | ... | ... | (2124) | (39.81%) | | |
| > Anglo-Sax | ... | ... | (1774) | (33.25%) | | |
| K2 Words (1001-2000): | 132 | 139 | **196** | **3.67%** | 3.67 | 95.11 |
| > Anglo-Sax | ... | ... | (132) | (2.47%) | | |
| 1k+2k | | | ... | (95.11%) | 0.51 | 95.62 |
| AWL Words: | 22 | 23 | **27** | **0.51%** | 4.39 | 100.00 |
| > Anglo-Sax | ... | ... | (6) | (0.11%) | | |
| Off-List Words: | ? | 171 | **234** | **4.39%** | | |
| | 773+? | 1153 | 5336 | 100% | | |

**Figure 1: Result of the movie script sorted by VocabProfile**

The words that have been uploaded in the profiler have produced results as shown in Figure 1. It is found that the movie script has 91.44% of K1 words, 3.67% of K2 words and 0.51% of AWL words. A test was conducted based on these groupings and the results are as follows.

### Findings of Needs Analysis

**Table 1: Scores of the Needs Analysis Questionnaire**

| N | LEVEL | MEAN | SD | NEEDS REPORT |
|---|---|---|---|---|
| 30 | K1 Words | 3.2 | 1.14 | Low Need |
| | K2 Words | 2.6 | 1.22 | Moderate Need |
| | AWL Words | 1.7 | 1.11 | High Need |

It is inferred from the results that the learners have scored well in high frequency words. However, there is a comparative decline in their knowledge of K2 and AWL words. The needs analysis test has proven that extensive priority should be given to teach lower frequency words.

## DATA ANALYSIS

### Background

The integral purpose of this study is to investigate the effectiveness of the instructional intervention on English as a Second Language (ESL) learners and to evaluate its impact on their language learning outcomes. The effectiveness of such interventions requires systematic examination to determine their potential contribution to language learning and teaching practices. This study seeks to assess whether the instructional intervention facilitates meaningful improvements in learners' performance by comparing their achievement before and after participation in the programme.

A research was conducted among the same group of students from 'A' section of basic streamlining to validate the hypothesis. The learners were asked to read the prescribed textbook, ***3 In 1 Tenali Raman: Value of Freedom & Other Stories***. The contents of the book were uploaded in the profiler tool and the results are as follows:

| | Families | Types | Tokens | Percent | Current profile | |
|---|---|---|---|---|---|---|
| | | | | | % | Cumul. |
| K1 Words (1-1000): | 779 | 1232 | 13306 | 85.21% | 85.21 | 85.21 |
| Function: | ... | ... | (7167) | (45.90%) | | |
| Content: | ... | ... | (6139) | (39.31%) | | |
| > Anglo-Sax | ... | ... | (4576) | (29.30%) | | |
| K2 Words (1001-2000): | 300 | 413 | 987 | 6.32% | 6.32 | 91.53 |
| > Anglo-Sax | ... | ... | (429) | (2.75%) | | |
| 1k+2k | | | ... | (91.53%) | 1.48 | 93.01 |
| AWL Words: | 98 | 129 | 231 | 1.48% | 6.99 | 100.00 |
| > Anglo-Sax | ... | ... | (30) | (0.19%) | | |
| Off-List Words: | ? | 589 | 1092 | 6.99% | | |
| | 1177+? | 2363 | 15616 | 100% | | |

**Figure 2: Result of the Textbook sorted by VocabProfile**

It is found that the contents of the textbook have 85.21% of K1 words, 6.32% of K2 words and 1.48% of AWL words. The learners were given a test on the vocabulary extracted from the profiler. The questionnaire consisted of 30 words, 10 words each from K1, K2 and AWL words and circulated via Google Forms to facilitate the process.

The study was conducted in two phases: Pre-test and Post-test phases. In the Pre-test phase, the learners were asked to answer the questionnaire based on their already prevalent proficiency. Whereas, in the post-test, the facilitators used profiling tools to categorise the words into the frequency groups and priority was given to low frequency words. Vocabulary practices like Word mapping, Total physical Response (TPR), Substitution and Interactive drills were introduced in the classroom and a post-test was conducted after all these pedagogical interventions.

**Reliability Analysis**

The researchers prepared the pre-test and post-test questionnaires based on the above-mentioned parameters. In order to ensure that the assessment produces consistent result, it is essential to conduct a reliability test for the instrument. The reliability of the present study was assessed by Kuder–Richardson formula (KR-20) to measure the test stability and guarantee consistent grading.

**Table 2: Reliability Analysis**

| Test | Participants | No of Items | Σpq | Total Test Variance ($\sigma^2$ | KR-20 Coefficient | Interpretation |
|---|---|---|---|---|---|---|
| Pre-test | 30 | 30 | 6.65 | 20.55 | 0.89 | 0.9 Acceptable |
| Post-test | 30 | 30 | 4.13 | 11.58 | 0.86 | 0.9 Acceptable |

The internal consistency reliability of the instrument was assessed using the Kuder–Richardson Formula 20 (KR-20), which is appropriate for dichotomously scored items. The analysis yielded a KR-20 coefficient of **0.9** for the pre-test questionnaire and **0.9** coefficient for the post-test questionnaire, indicating an acceptable internal consistency reliability for the questionnaire.

**Normality Testing**

Before proceeding with the descriptive and inferential statistics, normality test is conducted to determine whether the scores are distributed in a way that allows further statistical analyses. This study uses Skewness, Kurtosis scale to interpret the normality of the test results.

**Table 3: Normality Test**

| Variable | N | Mean Difference | SD | Skewness | Kurtosis | Interpretation |
|---|---|---|---|---|---|---|
| Post-test − Pre-test | 30 | 8.63 | 4.68 | 0.52 | 0.44 | Approximately Normal |

The obtained skewness value is **0.52** and kurtosis value is **0.44**, which determine that the test results are approximately normal in distribution.

### Descriptive Statistics

The descriptive statistics are used to provide a clear summary of the test results. The table shows the comparative analysis of the students' performance in pre-test and post-test phases.

**Table 4: Descriptive Statistics**

| Variable | N | Mean | SD | Min | Max | Mode |
|---|---|---|---|---|---|---|
| Pre-test | 30 | 16.16 | 4.53 | 6 | 27 | 15 |
| Post-test | 30 | 24.8 | 2.88 | 19 | 30 | 24 |

The average of the pre-test and post-test results are 16.16 and 24.8 respectively. This indicates a distinct improvement from the pre-test and post-test. The standard deviation (SD) denotes how varied the answers are. The pre-test SD is 4.53 and 2.88 which means the post-test results are very close to the mean and that the data is consistent and highly predictable.

### Inferential Statistics

#### *Hypothesis Testing*

A t-test is used to statistically validate the hypothesis stated in the introduction of this research paper. It determines that the gain score obtained from the pre-test and post-test are mathematically valid and are not produced by random chance.

**Table 5: Hypothesis testing using t-test**

| Variable | N | Mean | SD | Mean Difference | t | df | p |
|---|---|---|---|---|---|---|---|
| Pre-test | 30 | 16.16 | 4.53 | | | | |
| Post-test | 30 | 24.73 | 3.4 | 8.63 | 10.08 | 29 | < .001 |

A paired-sample t-test was conducted to examine the effect of the instructional intervention of corpus linguistic tools on learners' performance. The results revealed a statistically significant difference between the pre-test and the post-test scores**. $t(29) = 10.00$, $p < .001$**. This indicates that the instructional intervention has significantly improved the learners' performance. Since the findings indicate a statistically considerable improvement between the pre-test and post-test scores, the study thus rejects the null hypothesis and supports the alternate hypothesis.

#### *Effect Size Analysis*

Cohen's d is used to measure how large or significant the gain scores are. Large effect size that exceeds d=0.9 indicates that the improvement is large and efficient. The present study has produced a Cohen's d of **1.84** which denotes that the intervention has produced a practically meaningful and substantial improvement in learners' achievement.

**Table 6: Effect Size using Cohen's d**

| Variable | N | Mean Difference | SD DIFFERENCE | Cohen's d | Interpretation |
|---|---|---|---|---|---|
| Post test - Pre test | 30 | 8.63 | 4.68 | 1.84 | Large |

### Summary of Findings

The pre-test score highlights that out of 30 questions, the students scored well in K1 words and the score comparatively declined in K2 and AWL words. The pre-test result demonstrates that the learners are familiar with high frequency words and it advocates that the facilitator should give importance to low frequency words. The comparative results from the pre-test and post-test are elucidated in the pie diagrams below.

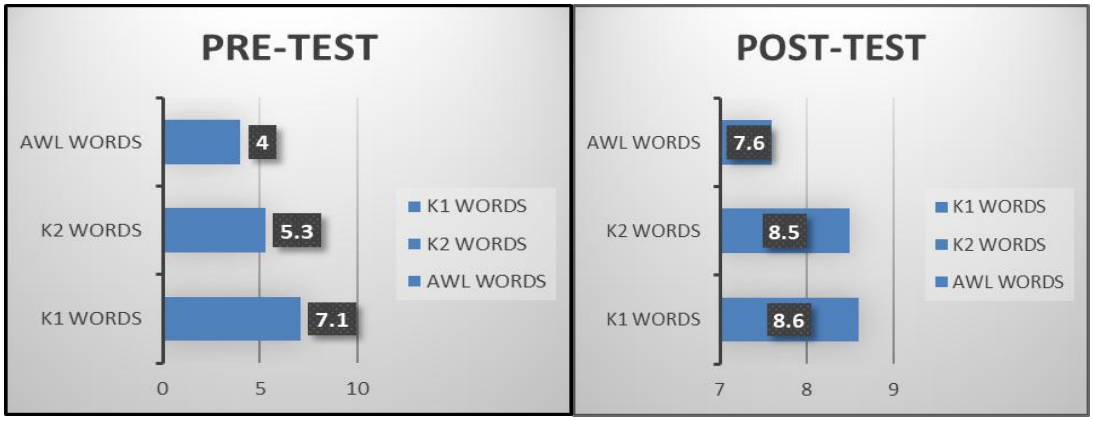



**Figure 3: Comparative Analysis of Pre-test and Post-test scores**

## DISCUSSION

### Observation

The descriptive and inferential analyses highlight the efficacy of the instructional strategy employed in this study. The increase in average, a very close SD range and large Cohen's d indicate that the hypothesis was successfully validated. It indicates that the pedagogical aid significantly enhanced students' English language abilities. The improvement in mean scores suggests that curated teaching methods can produce measurable gains in language proficiency. The study provides a clear demonstration of how instructional aids designed with the help of corpus tools yield substantial benefits for learners.

### Research Findings

The present study has attempted to evaluate the significance of corpus tools in enhancing students' lexical aptitude. The data collected through pre-test and post-test assessments

have provided clear evidence of the impact of the hypothesis. It is found that learners from basic streamline are not exposed to a variety of vocabulary due to a lack of exposure to English. Hence, it becomes an obligation of language facilitators to embed new and sporadic words in the instructional materials.

The research posits that the instructional materials provided to students are not equipped with the categorisation of vocabulary based on their frequency, which assists in acquisition of new words **(RQ1**). It is also found that the students are familiar with high frequency words that they use in everyday life. However, it is observed that the students find it difficult to master low frequency words and academic words that are less commonly used in daily life **(RQ2).** The research proposes that instructional materials should transform from being teacher-centric to learner-centric to enhance maximised growth of vocabulary acquisition among ESL learners. Implementing corpus tools in pedagogical materials will provide the expected shift towards a learner-friendly environment **(RQ3).**

## Research Recommendations

Based on the observations made from the acquired statistics, this research recommends that the instructional materials provided to English as Second Language (ESL) learners must include data-driven learning approaches, through which the learners directly interact with the authentic source materials. The outcomes of this study stipulate that educational institutions must consider expanding learners' corpora by including spoken and written data from students of various educational backgrounds, age group and proficiency levels. These customised corpora will assist the facilitators in identifying the needs of the learners and in catering to their demands.

The observations from the pre-test and post-test results infer that the learners are more accustomed to K1words that are frequently used in their day-to-day conversations. Thus, the English language facilitators must conduct vocabulary drills for the retention of K1 words. It is also recommended that the facilitators introduce K2 words, thereby achieving the acquired linguistic proficiency. This paper recommends that proper focus and training should be provided to the facilitators on integrating DDL tools in classrooms. Developing training modules for corpus-informed teaching could enhance the practicality of language learning.

## Limitations

One of the major limitations of Corpus-informed pedagogy is the lack of corpus that are exclusively curated for educational purposes. The relatively smaller size of the available corpus does not fully represent the broader population of ESL learners. A larger dataset could provide more authentic linguistic sources. Corpus linguistics often requires the use of specialised software such as concordances and frequency analysis tools. These tools may be difficult for facilitators and leaners of beginner levels to understand and operate the corpora without proper training. As a result of this, the implementation of corpus-based activities in classrooms can become challenging. Though corpus data provide authentic examples, they often present language in short, fragmented contexts. These limited contexts may make it difficult for learners to fully understand the meaning, pragmatic use, or discourse function of certain vocabulary items.

**Scope for Further Study**

Future research could involve creating a larger learner corpus with a wider range of learners from different educational backgrounds and proficiency levels. These corpora could enable facilitators to introduce new and necessary vocabulary to the learners. Further studies can also investigate the relative effectiveness of corpus-based vocabulary learning for ESL beginners, intermediate and advanced learners. These studies can reveal whether corpus linguistics can help ESL learners at different proficiency levels. Most vocabulary studies carried out using corpora have concentrated only on written texts. Future studies could include spoken corpora of learners' language to investigate how vocabulary is used in real-time interaction. This would give a better understanding of how learners use vocabulary in their day-to-day activities.

**Conclusion**

Corpus linguistics has proved to be an effective pedagogical tool to enhance the vocabulary skills. The implementation of corpus-based curriculum in ESL classrooms has exposed the learners to linguistic data derived from real-life situations. This in turn enables students to get familiarised with the usage of words in their natural contexts. Corpus tools can gradually encourage learners to become independent language explorers. With proper guidance, beginners can learn to use simple corpus-based resources to check word usage, frequency and examples.

Corpus linguistics has emerged as a powerful and practical approach to vocabulary instruction for beginner-level ESL learners. Corpus-based instructional materials have addressed the individual learning needs of learners through authentic language examples. As a result of this innovative intervention, vocabulary learning has become more meaningful, engaging and efficient. The integration of corpus-based resources has enabled leaners to observe vocabulary in authentic contexts. This exposure has strengthened learners' lexical competence by promoting contextualised language learning. The incorporation of corpus linguistics into vocabulary learning has contributed not only to lexical development but also to the implementation of learners' overall communicative performance

**Disclosure Statement:**

No potential conflict of interest was reported by the author(s).

**Funding Details**

This research is funded by the Senior Research Fellowship under the University Grants Commission

## REFERENCES

Abraham, M., Arficho, Z., & Habtemariam, T. (2022). Effects of training in ICT-assisted English language teaching on secondary school English language teachers' knowledge, skills, and practice of

using ICT tools for teaching English. Education Research International, 1–10. https://doi.org/https://doi.org/10.1155/2022/6233407

Alshraah, S. M., & Nishat, A. S. (2023). Scrutinizing the relationship between language proficiency and the use of mitigated devices by Saudi learners. Research Journal in Advanced Humanities, 4(3). https://doi.org/10.58256/rjah.v4i3.1324

Barua, N. M. (2026). AI-generated graphic organisers and L2 reading comprehension: evidence from a tertiary elt classroom in India. Journal of English Language Teaching, 68(2), 38–44. https://doi.org/10.66121/z192tc30

Boulton, A. (2010). Data-driven learning: taking the computer out of the equation. Language Learning, 60(3), 534-572. https://doi.org/10.1017/S0958344009000068

Dudeney, G., & Hockly, N. (2012). ICT in ELT: How did we get here and where are we going? ELT Journal, 66(4), 533–542. https://doi.org/10.1093/elt/ccs050

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. International Journal of Corpus Linguistics, 14(3), 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Gulnaz, F., Alfaqih, A. M., & Mashhour, N. A. (2015). Paradigm shift: A critical appraisal of traditional and innovative roles of an English teacher in Saudi ELT classrooms. Theory and Practice in Language Studies, 5(5), 934–946. https://doi.org/http://dx.doi.org/10.17507/tpls.0505.07

Hadley, G. (2002). An introduction to data-driven learning. RELC Journal, 33(2), 99–124. https://doi.org/10.1177/003368820203300205

Hashim, A. A. (2025). Innovative approaches in English language teaching: A comparative analysis of traditional and modern methods. South Asian Research Journal of Arts, Language and Literature, 7(01), 1–7. https://doi.org/10.36346/sarjall.2025.v07i01.001

Hidayah, J., & Prihantoro. (2022). ICT-based learning in new normal era: viewed from practice and impact in ELT classroom . ENGLISH FRANCA: Academic Journal of English Language and Education, 6(1), 37–63. https://doi.org/10.29240/ef.v6i1.4116

Inglis, B. (2026). Doing DDL right: Principles and practical tips for the classic use of corpora in the language classroom. JALTCALL Trends, 2(1), 103471. https://doi.org/10.29140/jct.v2n1.103471

Jager, S. (2004). Learning Management Systems for Language Learning. In A. Chambers, J. E. Conacher, & J. Littlemore (Eds.), ICT and Language Learning: Integrating Pedagogy and Practice (pp. 33–48). essay, Bloomsbury Academic.

Jakob, J. C., Persulessy, S. I., Nikijuluw, R. P., Le-Hong, P., & Afdaliah, N. (2026). Integrating duolingo and QUIZIZZ in English Language Learning for Civil Engineering Students: A Literature Review. Epigram, 23(1), 64–81. https://doi.org/10.32722/epi.v23i1.8348

Jasmih, J. I., & Ngui, W. (2026). Enhancing ESL learners' speaking confidence through real-life simulation: A qualitative case study. Jurnal Pemikir Pendidikan, 14(1), 83–94. https://doi.org/10.51200/jpp.v14i1.6880

Johns, T. (2010). Should you be persuaded - Two samples of data-driven learning materials ELR Journal, 4, 1–16.

Katel, K. P. (2024). Transforming ELT with ICT: Empowerment or encumbrance. Solukhumbu Multiple Campus Research Journal, 6(1), 36–53. https://doi.org/https://doi.org/10.3126/smcrj.v6i1.74522

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., & Volodina, E. (2013). Corpus-based vocabulary lists for language learners for nine languages. Language Resources and Evaluation, 48(1), 121–163. https://doi.org/10.1007/s10579-013-9251-2

Macaulay's minute on education, February 2, 1835. https://home.iitk.ac.in/~hcverma/Article/Macaulay-Minutes.pdf

Nakazinga, R., Waiswa, M. M., Kalanda, E. D., & Tumuhek, P. B. (2026). Teachers' understanding of the learner-centred approach in English language classrooms in selected secondary schools in Central Uganda. Journal of Research Innovation and Implications in Education. https://doi.org/10.59765/vpr8

Nduwimana, A., & Ndoricimpa, C. (2023). The use of ICT tools in learning English autonomously. JOLLT Journal of Languages and Language Teaching, 11(4), 696–708. https://doi.org/https://doi.org/10.33394/jollt.v

Orujlu, A. (2026). The transformation of English language teaching in the digital era: Autonomy, assessment, Multimodal Literacy, and Artificial Intelligence. Acta Globalis Humanitatis et Linguarum, 3(4), 19–29. https://doi.org/10.69760/aghel.026004002

Pennycook, A. (2007). ELT and colonialism. International Handbook of English Language Teaching, 13–24. https://doi.org/10.1007/978-0-387-46301-8_2

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in call research during 2011–2015. Computer Assisted Language Learning, 35(1–2), 36–61. https://doi.org/10.1080/09588221.2019.1667832

Sharndama, E. C. (2013). Application of ICTs in teaching and learning English (ELT) in large classes. Journal of Arts and Humanities, 2(6), 34–39.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. Language Teaching Research, 11(4), 393–412. https://doi.org/10.1177/1362168807080960

Szudarski, P. (2018). Corpus Linguistics for vocabulary: A guide for research. Routledge.

Tengku Kasim, T. S. (2014). Teaching paradigms: An analysis of traditional and student-centred approaches. Journal of Usuluddin, 40, 199–218. https://doi.org/10.22452/usuluddin.vol40.8

Teubert, W. (2005). My version of corpus linguistics. International Journal of Corpus Linguistics, 10(1), 1–13. https://doi.org/10.1075/ijcl.10.1.01teu