

JEDLIT

Journal of Education and Literature

e-ISSN: 3063-0533

<https://ejurnal.unival-cilegon.ac.id/index.php/jedlit>

The Role of Digital Corpus-Based Applications in Enhancing English Collocation Mastery: A Conceptual Review for ICT-Integrated English Language Education

Maya Amelia¹, Pilda Nugraha Firdaus²

Program Studi Tadris Bahasa Inggris, Fakultas Hukum dan Agama Islam,
Universitas Al-Khairiyah (UNIVAL)

mayaamelia250@gmail.com , pildanugrahafirdaus@gmail.com

Abstrak

Penguasaan kolokasi bahasa Inggris—seperti *make a decision* atau *heavy rain*—masih menjadi tantangan tersendiri bagi pembelajar bahasa Inggris sebagai bahasa asing (EFL), termasuk calon guru bahasa Inggris. Kesulitan ini bersumber dari sifat kolokasi yang probabilistik dan sulit dijelaskan hanya melalui kaidah tata bahasa. Salah satu solusi yang berkembang pesat dalam beberapa tahun terakhir adalah pemanfaatan aplikasi digital berbasis korpus (seperti SKELL, COCA, dan Just The Word) melalui pendekatan Data-Driven Learning (DDL). Artikel ini bertujuan mengkaji secara konseptual bagaimana aplikasi digital berbasis korpus dapat berkontribusi terhadap penguasaan kolokasi bahasa Inggris, serta merumuskan implikasinya bagi integrasi ICT dalam pendidikan bahasa Inggris. Kajian dilakukan melalui studi kepustakaan (library research) terhadap literatur primer terbitan sepuluh tahun terakhir yang membahas kolokasi, linguistik korpus, dan pembelajaran berbasis data. Hasil kajian menunjukkan bahwa penggunaan aplikasi korpus digital secara konsisten dilaporkan berkontribusi positif terhadap penguasaan kolokasi peserta didik, dengan syarat disertai pelatihan literasi korpus yang memadai dan keterlibatan yang terarah, bukan sekadar eksposur singkat. Artikel ini merekomendasikan integrasi pelatihan literasi korpus digital ke dalam kurikulum pendidikan bahasa Inggris, khususnya bagi program pencetak calon guru.

Kata-kata kunci: kolokasi; korpus digital; ICT; data-driven learning; pendidikan bahasa Inggris

Abstract

*Mastering English collocations—such as *make a decision* or *heavy rain*—remains a persistent challenge for learners of English as a foreign language (EFL), including pre-service English teachers. This difficulty stems from the probabilistic nature of collocation, which is difficult to explain through grammatical rules alone. One promising response that has developed rapidly in recent years is the use of digital corpus-based applications (such as SKELL, COCA, and Just The Word) through a Data-Driven Learning (DDL) approach. This article aims to conceptually examine how digital corpus-based applications can contribute to English collocation mastery and to formulate its implications for ICT integration in English language education. The review was conducted through library research on primary literature published within the last ten years concerning collocation, corpus linguistics, and data-driven learning. The review shows that the use of digital corpus tools is consistently reported to contribute positively to learners' collocation mastery, provided it is accompanied by adequate corpus-literacy training and purposeful engagement rather than brief, unguided exposure. This article*

recommends integrating digital corpus-literacy training into English language education curricula, particularly within teacher-training programs.

Keywords: collocation; digital corpus; ICT; data-driven learning; English language education

INTRODUCTION

Mastering collocations—conventional and recurrent word combinations such as "make a decision" or "heavy rain"—remains one of the most persistent challenges for learners of English as a foreign language (EFL), even at advanced proficiency levels (Nesselhauf, 2003). Unlike grammatical rules, which can often be explained through explicit formulas, collocational knowledge is inherently probabilistic: it depends on how words habitually co-occur in authentic language use rather than on logical or semantic necessity (Sinclair, 1991). Consequently, EFL learners frequently rely on direct translation from their first language, producing expressions that, although grammatically correct, sound unnatural to native speakers.

At Universitas Al-Khairiyah (UNIVAL), students of the Tadris Bahasa Inggris (English Language Education) study program are expected to produce accurate and idiomatic English, both in academic writing and in their future roles as English teachers. Preliminary classroom observation, however, indicates that many students continue to produce collocational errors such as "do a mistake" instead of "make a mistake," suggesting that conventional instruction—largely dependent on textbooks and teacher explanation—has not been sufficient to build strong collocational competence.

One promising response to this problem is the integration of Information and Communication Technology (ICT), particularly corpus-based digital tools, into language instruction. Digital corpora and free online concordancers such as SKELL, COCA, and Just The Word allow learners to search authentic language data and observe how words naturally combine, an approach known as Data-Driven Learning (DDL) (Godwin-Jones, 2017). A growing body of research has shown that DDL and corpus consultation can meaningfully improve learners' collocational knowledge (Boulton & Cobb, 2017; Du, Afzaal, & Al Fadda, 2022; He & Xie, 2024; Wu, 2021). Nevertheless, empirical evidence from the Indonesian tertiary EFL context—particularly from teacher-training programs such as TBI UNIVAL—remains limited, leaving open the question of whether such tools are equally beneficial for pre-service English teachers who must model correct language use for their future students.

This article, therefore, aims to conceptually examine (1) why collocation remains a persistent difficulty for EFL learners from a linguistic-theoretical standpoint, (2) how digital corpus-based applications and the Data-Driven Learning approach have been shown to address this difficulty, and (3) what pedagogical implications this body of research holds for the integration of ICT into English Language Education curricula, particularly within teacher-training programs such as TBI UNIVAL. It is expected that this conceptual synthesis will provide a coherent theoretical basis for future empirical research and for curricular decisions regarding the place of corpus-based ICT tools in English teacher education.

THEORY

Collocation refers to the tendency of certain words to co-occur more frequently than chance would predict, forming semi-fixed combinations that native speakers produce

automatically (Sinclair, 1991). Sinclair's idiom principal contrasts with the open-choice principle, arguing that fluent language production relies heavily on pre-fabricated chunks rather than on word-by-word grammatical construction. From a pedagogical perspective, Nesselhauf (2003) demonstrated that even advanced learners struggle to reproduce native-like collocations, frequently substituting a collocate with a near-synonym that is semantically plausible but statistically rare in authentic usage (e.g., "strong tea" is idiomatic, while "powerful tea" is not). This pattern suggests that collocational competence cannot be developed through grammar instruction alone; it requires repeated, meaningful exposure to authentic co-occurrence patterns.

Corpus linguistics offers a methodological solution to this pedagogical challenge. A corpus—a large, electronically stored collection of authentic texts—allows learners and teachers to observe language “in the wild” through concordance lines that display a search word together with its immediate linguistic environment (Godwin-Jones, 2017). When learners are guided to explore corpus data themselves rather than simply receiving rules from a teacher, the approach is referred to as Data-Driven Learning (DDL). Boulton and Cobb's (2017) meta-analysis of DDL studies found a generally positive effect on vocabulary and collocation learning, although the magnitude of the effect varies depending on learners' proficiency level, prior training in corpus use, and the type of corpus tool employed.

Recent advances in ICT have made corpus consultation considerably more accessible to language learners who lack advanced technical skills. User-friendly online concordancers such as SKELL, COCA, and Just The Word allow learners to type a word and instantly view its most frequent collocates and example sentences, without requiring knowledge of corpus query syntax. Empirical studies confirm the pedagogical value of such tools: Wu (2021) found that learners who used DDL to discover collocations in their own writing produced more accurate and varied collocational patterns than a control group relying on dictionaries alone; Satake (2022) reported similar gains in collocation accuracy following short, structured corpus-consultation training; and He and Xie (2024) demonstrated that sustained, autonomous engagement with a corpus tool over time predicted durable gains in noun-collocation knowledge among pre-tertiary learners. In the context of academic and professional writing, Chen and Flowerdew (2018) further showed that graduate students who incorporated corpus consultation into their writing process gained not only in collocational accuracy but also in confidence and independence as writers.

Complementing these findings, Sun and Park's (2023) systematic review of 32 empirical studies concluded that corpus-based collocation instruction is broadly effective across proficiency levels and educational contexts, although its success depends on adequate learner training and sustained engagement rather than one-off exposure. Du, Afzaal, and Al Fadda (2022) additionally revealed, through corpus-driven analysis of learner writing, that verb-noun collocation errors tend to persist even at higher proficiency levels, reinforcing the need for continuous, technology-supported practice. Moon and Oh (2018) similarly found that data-driven correction techniques were effective in helping learners “unlearn” persistent, over-generalized collocational errors in secondary EFL classrooms, while Vyatkina (2016) showed that even beginning-level learners could benefit from carefully scaffolded corpus tasks when the tool and materials were adapted to their proficiency.

Taken together, these studies suggest a coherent theoretical proposition: the more frequently and purposefully learners engage with digital corpus-based tools, the stronger

their collocational competence is likely to become. This proposition forms the conceptual foundation of the present study, which examines the relationship between digital corpus tool use and collocation mastery among pre-service English teachers at UNIVAL.

METHOD

This article employs a qualitative conceptual-review method based on library research (studi kepustakaan). This approach was chosen because the aim of the study is not to test a hypothesis empirically, but to synthesize, compare, and critically discuss existing theoretical and empirical literature in order to construct a coherent conceptual understanding of how digital corpus-based applications relate to English collocation mastery.

The literature search was conducted through academic databases and indexing services, including Google Scholar, ScienceDirect, Frontiers, and journal repositories such as Language Learning & Technology and ReCALL, using keyword combinations such as “collocation,” “corpus linguistics,” “data-driven learning,” “digital corpus tools,” and “ICT in English language teaching.” Sources were selected based on the following inclusion criteria: (1) published in peer-reviewed journals or academic books, (2) directly addressing collocation, corpus-based instruction, or data-driven learning in the context of English language teaching, and (3) predominantly published within the last ten years (2016–2025), with a small number of foundational, frequently cited classic works retained for historical and theoretical grounding.

The selected sources were analyzed using thematic synthesis, in which recurring concepts, findings, and arguments across studies were identified, grouped into coherent themes, and critically compared to reveal points of convergence, divergence, and gaps in the existing literature. These themes subsequently form the structure of the Pembahasan (Discussion) section of this article.

RESULTS AND DISCUSSION

Why Collocation Remains a Persistent Challenge

A recurring theme across the reviewed literature is that collocational competence develops differently from grammatical competence. Sinclair (1991) argued that fluent native-like production relies heavily on the idiom principle—the retrieval of pre-fabricated word combinations—rather than on the open-choice, word-by-word construction that grammar instruction typically emphasizes. Building on this view, Nesselhauf (2003) showed empirically that even advanced learners frequently substitute a collocate with a semantically plausible but statistically atypical near-synonym (e.g., “*powerful tea” instead of “strong tea”), precisely because such choices cannot be ruled out by grammatical logic alone. Du, Afzaal, and Al Fadda (2022), analyzing large-scale learner corpus data, further demonstrated that verb-noun collocation errors persist across proficiency levels, indicating that the problem is not simply a matter of insufficient vocabulary size but of insufficient exposure to authentic patterns of co-occurrence. Taken together, this strand of literature converges on a single conceptual point: because collocation is fundamentally usage-based rather than rule-based, it calls for a pedagogy centered on exposure to authentic language data—precisely the affordance that corpus-based digital tools are designed to provide.

Digital Corpus-Based Applications and Data-Driven Learning

The second theme emerging from the literature concerns how digital corpus tools operationalize this usage-based exposure through Data-Driven Learning (DDL). Godwin-Jones (2017) describes DDL as an approach in which learners consult concordance lines—authentic sentences containing a target word—to inductively discover patterns of use, rather than being told the rules directly. Several empirical studies reviewed for this article report converging evidence of DDL's effectiveness specifically for collocation learning. Wu (2021) found that learners who used corpus consultation to check and revise collocations in their own writing produced more accurate and more varied collocational patterns than peers who relied solely on dictionaries. Satake (2022) reported comparable gains in collocation accuracy following brief, structured corpus-consultation training, while He and Xie (2024) showed that sustained, autonomous engagement with a corpus tool over an extended period predicted more durable collocation gains than short-term exposure alone. In the domain of academic writing specifically, Chen and Flowerdew (2018) documented that graduate students who incorporated corpus consultation into their writing process gained not only in collocational accuracy but also in confidence and independence as writers. Boulton and Cobb's (2017) meta-analysis situates these individual findings within a broader pattern: across a large number of DDL studies, the overall effect on vocabulary and collocation learning is positive and moderate in size, though it varies according to learners' proficiency and the amount of training they receive in using the tool.

At the same time, the literature is not uniformly optimistic. Sun and Park's (2023) systematic review of 32 empirical studies concluded that the benefits of corpus-based instruction are conditional rather than automatic: learners who consult a corpus tool without understanding how to interpret concordance lines or word-sketch data tend to derive limited benefit, and gains are strongest when corpus use is scaffolded and sustained rather than incidental. Moon and Oh (2018) illustrate one form this scaffolding can take, showing that structured, data-driven correction tasks were effective in helping secondary EFL learners “unlearn” persistent, over-generalized collocational errors. Vyatkina (2016) extends this picture to lower-proficiency learners, demonstrating that even beginners can benefit from corpus-based tasks provided the tool and materials are carefully adapted to their level. Table 1 summarizes the focus and key contribution of the primary empirical studies reviewed in this article.

Table 1
Summary of Reviewed Empirical Studies on Corpus-Based Collocation Learning

Study	Focus	Key Contribution
Wu (2021)	L2 writing	Corpus-based discovery of collocations improved accuracy and variety compared to dictionary use.
Satake (2022)	Corpus training	Brief, structured corpus-consultation training improved collocation accuracy.
He & Xie (2024)	Pre-tertiary learners	Sustained, autonomous corpus engagement predicted durable noun-collocation gains.
Chen & Flowerdew (2018)	Academic writing	Corpus consultation increased collocational accuracy and writer confidence among graduate students.
Moon & Oh (2018)	Secondary EFL	Data-driven correction tasks helped learners unlearn persistent collocational errors.
Vyatkina (2016)	Beginners	Scaffolded corpus tasks benefited even lower-proficiency learners.
Sun & Park (2023)	Systematic review (32 studies)	Corpus-based instruction is broadly effective, but conditional on adequate learner training.

Sumber: disintesis oleh penulis dari literatur primer yang dikaji (2016–2024).

Pedagogical Implications for ICT Integration in English Teacher Education

The third theme concerns what this body of research implies for curriculum and pedagogy, particularly within programs that prepare future English teachers, such as TBI UNIVAL. As prospective educators, students are expected not only to master collocations for their own communicative competence, but eventually to model and teach these patterns to their own pupils. The literature reviewed above suggests that simply making corpus tools available is insufficient; the conditional nature of DDL's benefits, as emphasized by Sun and Park (2023), implies that explicit training in corpus-tool literacy should be embedded within relevant courses—such as Vocabulary, Structure, or ICT-integrated ELT courses—rather than left to students' independent initiative. Scaffolding techniques such as guided concordance-reading worksheets or collaborative corpus-search tasks, in line with Vyatkina's (2016) findings for lower-proficiency learners and Moon and Oh's (2018) structured correction tasks, appear especially promising for learners unfamiliar with corpus interfaces. Embedding such training early in a teacher-education curriculum also carries a second-order benefit: it equips future teachers not merely to use corpus tools themselves, but to design corpus-informed collocation activities for their own future classrooms, extending the pedagogical value of this technology beyond a single generation of learners.

Before proceeding to the conclusion, it is worth emphasizing that although digital corpus tools cannot replace explicit instruction or meaningful communicative practice, they offer a uniquely humanistic pedagogical opportunity: they invite learners to become active investigators of language rather than passive recipients of rules, fostering the autonomy, curiosity, and self-directed spirit that lies at the heart of meaningful, lifelong language education.

CONCLUSION

This conceptual review set out to examine why collocation remains difficult for EFL learners, how digital corpus-based applications and the Data-Driven Learning approach have been shown to address this difficulty, and what implications this literature holds for ICT integration in English teacher education. The reviewed literature converges on the conclusion that collocational competence is fundamentally usage-based rather than rule-based, and that digital corpus tools—by providing learners with direct, searchable access to authentic co-occurrence patterns—offer a pedagogically coherent response to this challenge. At the same time, the literature consistently cautions that the benefits of corpus-based learning are conditional: they depend on adequate learner training and sustained, purposeful engagement rather than incidental or one-off exposure to a corpus tool.

These conclusions affirm the pedagogical value of integrating digital corpus tools into English Language Education programs, particularly for pre-service English teachers, who must eventually model and teach authentic collocational patterns to their own future students. Guided, scaffolded exposure to corpus-based learning is therefore not only a matter of improving personal language accuracy, but also of preparing confident, technologically literate educators capable of designing corpus-informed activities in their own classrooms.

Based on this synthesis, several recommendations are offered. First, English Language Education programs such as TBI UNIVAL are encouraged to embed structured corpus-literacy training—covering accessible tools such as SKELL, COCA, and Just The

Word—within relevant courses such as Vocabulary, Structure, or ICT in ELT, rather than assuming students will adopt such tools independently. Second, lecturers may design collaborative, guided corpus-search tasks to help less confident students move from passive to active, exploratory engagement with authentic language data. Finally, given that the conclusions of this article are derived from a synthesis of existing literature, future empirical research—employing experimental, quasi-experimental, or longitudinal designs with learners in the Indonesian EFL teacher-training context—is recommended to test and extend these conceptual propositions with field-based classroom data.

REFERENCES

- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393.
- Chen, M., & Flowerdew, J. (2018). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes*, 50, 97–112.
- Du, X., Afzaal, M., & Al Fadda, H. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology*, 13, 752134.
- Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, 21(3), 9–27.
- He, M., & Xie, Q. (2024). Empowering autonomy in language learning: The sustainable impact of data-driven learning on noun collocation acquisition. *Humanities and Social Sciences Communications*, 11, Article 1516.
- Moon, S., & Oh, S.-Y. (2018). Unlearning overgenerated be through data-driven learning in the secondary EFL classroom. *ReCALL*, 30(1), 48–67.
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223–242.
- Satake, Y. (2022). The effects of corpus use on learning L2 collocations. *JALT CALL Journal*, 18(1), 34–53.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Sun, W., & Park, E. (2023). EFL learners' collocation acquisition and learning in corpus-based instruction: A systematic review. *Sustainability*, 15(17), Article 13242.
- Vyatkina, N. (2016). Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL*, 28(2), 207–226.
- Wu, Y. A. (2021). Discovering collocations via data-driven learning in L2 writing. *Language Learning & Technology*, 25(2), 192–214.