Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-04038-6 OPEN

Empowering autonomy in language learning: the sustainable impact of data-driven learning on noun collocation acquisition

Mengyu He [1] & Qin Xie[2,3]✉

This study explores the effectiveness of Data-Driven Learning (DDL) in teaching noun collocations to pre-tertiary learners in Hubei Province, China, using the online corpus tool 'Corpusmate'. Acknowledging the importance and challenge of mastering collocations in learning a foreign language, this study focuses on the effects of DDL on pre-tertiary learners, an area less examined previously due to the complexities associated with using corpus tools. Conducted over two months, the research employed pre-tests, post-tests, and delayed post-tests to measure learners' comprehension and retention of noun collocations. Additionally, a questionnaire was distributed to gather feedback on learners' experiences with the DDL approach and 'Corpusmate'. Results indicated that the experimental group, which received DDL training, showed significant improvements in test scores compared to the control group, which used traditional resources. The experimental group's scores remained high in the delayed post-test, suggesting that the DDL approach facilitated long-term retention of collocational knowledge, although a notable proportion of learners expressed neutral or negative perceptions of the DDL activities. These results highlight the need for further investigation into the attitudes of the participants. Overall, most participants provided positive feedback on the use of 'Corpusmate' in learning noun collocations. These results advocate for the incorporation of corpus consultation into language teaching practices. The study underscores that with appropriate training and tools like 'Corpusmate', the DDL approach can potently aid in the sustained learning of complex language elements, such as collocations, even for younger learners.

[1] School of Foreign Languages, Hubei University of Economics, Wuhan, China. [2] College of Foreign Languages, Minjiang University, Fuzhou, China. [3] Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong, China. ✉email: minjiangxieqin@163.com

Introduction

The acquisition of collocations is a vital element of language learning, directly influencing a learner's proficiency in English as a second (ESL) or foreign language (EFL). Sinclair (1987) and Biber and Conrad (1999) emphasise that the adept use of collocations marks a significant proficiency milestone, affecting both language production and comprehension. Mastery of collocations not only facilitates natural language use but is also crucial for understanding linguistic nuances, as demonstrated by studies like those of Meunier and Granger (2008) and Crossley et al. (2015). However, mastering collocations is challenging for many learners, including those at advanced levels. This challenge can be attributed to the inherent complexity of collocations, the influence of the learner's first language (L1), and certain implicit teaching practices that do not explicitly address collocational use (Boers et al. 2014, Chen, 2019, Gablasova et al. 2017). In particular, Laufer and Waldman (2011), Frankenberg-Garcia (2018), and Pellicer-Sánchez and Boers (2019) found that conventional approaches often inadequately develop learners' understanding of collocations, prompting a search for more effective teaching strategies. This growing body of literature highlights the need for innovative methodologies to better support the learning and application of collocations in ESL and EFL settings.

DDL is an effective approach that uses authentic language data to enhance language learning and supplement traditional teaching methods. DDL leverages corpora—large, digital collections of real language use—to immerse learners in authentic contexts, thereby emphasising the linguistic features under study. This approach, supported by findings from Boulton and Cobb (2017) and further evidenced by Chen and Flowerdew (2018) and Lee et al. (2019), has proven to be an effective supplement to traditional teaching methods. It is particularly beneficial for conveying complex linguistic concepts, offering clear advantages over other instructional methods (Vyatkina, 2016a).

Despite increasing interest in DDL, its adoption in specific educational sectors such as private education, professional learning environments, and pre-tertiary education remains limited (Crosthwaite and Steeples, 2022). This underrepresentation is due to several factors. Prominently, ethical and bureaucratic hurdles hinder access to younger student populations (Brown et al. 2020; Crosthwaite and Steeples, 2022). Additionally, there is a critical need to convince educators, who play a key role in implementing new teaching methods, of DDL's benefits for both teachers and students (Tondeur et al. 2019). Conflicts between DDL approaches and traditional language and literacy teaching methods (Bednarek et al. 2020), the heavy workloads of pre-tertiary educators limiting their capacity for professional development and adoption of digital methods (Park and Son, 2020), and a lack of adequate funding for extensive DDL research in these educational settings (Crosthwaite and Steeples, 2022) also pose significant challenges.

This study explores how DDL can be effectively implemented for teaching noun collocations in a pre-tertiary EFL educational setting. Conducted over two months at a high school in Hubei Province, China, it evaluates the impact of DDL on noun collocation comprehension among Chinese high school students using the online corpus tool 'Corpusmate' (Crothswaite and Baisa, 2023). Noun collocation refers to a natural combination of words where one of the collocation components is a noun, and it is typically paired with other words that frequently occur together in natural language use (He and Ang, 2023). These other words can be verbs, adjectives, or other nouns. In noun collocations, the noun serves as the central element, and the accompanying words are chosen because they are conventionally and semantically compatible with the noun, forming a phrase that is commonly recognized by native speakers as a standard or natural expression. The research utilized pre-tests, post-tests, and delayed post-tests to measure learners' understanding of collocations before, immediately after, and three months following the DDL intervention. A questionnaire was also distributed to gather detailed feedback on learners' experiences with the corpus platform. Grounded in the understanding of language learning as a sustainable and autonomous process, this study examines the intersections of linguistics, technology, and pedagogy. It aims to contribute to the development of more innovative, effective, and sustainable methods for language education.

Literature review

Collocations in the English language. Collocations, defined as combinations of words that frequently appear together in a language, form an integral part of language learning, contributing significantly to fluency and naturalness in language use. The study of collocations has evolved over the years, with seminal works such as Lewis (2000) providing foundational knowledge. Recent research continues to explore the complexity and pedagogical implications of teaching collocations, emphasizing their importance in both written and spoken language (Barfield, 2009).

The ability to accurately use collocations is crucial for language proficiency, as emphasized by Sinclair (1987), Biber and Conrad (1999), and Simpson-Vlach and Ellis (2010). Mastery of collocations is essential for assessing linguistic competence, as highlighted by Meunier and Granger (2008), Crossley et al. (2015), and Nizonkiza and Van de Poel (2019). The skilful use of collocations plays a vital role in both language production and comprehension. Durrant (2019), Rezaee et al. (2015), and Basal (2019) illustrate the importance of collocations in language production, while Gyllstad (2009), Nizonkiza (2014), and Chen et al. (2021) focus on their role in comprehension. Collocations are an integral part of formulaic language and vocabulary, significantly affecting the ability to communicate effectively in English as a second or foreign language, as discussed by Durrant (2019). However, learning collocations such as 'dark night' or 'heavy rain' presents challenges for many learners, including those at advanced levels. Boers et al. (2014) observed that high-frequency verb parts in collocations can be particularly challenging because they do not significantly contribute to the overall meaning and fail to engage learners' attention. The difficulties in mastering collocations are attributed to their inherent complexity, the influence of the learner's first language, and specific teaching methods. Boers et al. (2014) and Gablasova et al. (2017) explored how these factors complicate the learning process. Additionally, researchers like Laufer and Waldman (2011), Frankenberg-Garcia (2018), and Pellicer-Sánchez and Boers (2019) have made significant contributions to understanding these challenges, providing insights into how collocational competence can be developed more effectively.

In response to the challenges associated with learning collocations, experts such as Lewis (2000), Wray (2002), and Nesselhauf (2004) have advocated for the explicit teaching of collocations, emphasising the need to address differences between the learner's first language and the target language. Various teaching strategies have been adopted to facilitate this process. Boers et al. (2014) discuss methods such as dictation, matching exercises, underlining, and inserting verb-noun collocations in classroom settings, which aim to provide learners with frequent and focused exposure to collocations. Moreover, awareness-raising activities that emphasise explicit instruction and encourage learners to reflect on their learning processes have been developed. For instance, Boers et al. (2006), Coxhead (2008), and

Peters (2009) have implemented activities such as phrase noticing, essay writing, before and after interventions and interviews. These activities are designed to make learners more aware of collocations and to foster deeper engagement with the learning materials. However, while these traditional methods have shown some effectiveness in teaching collocations, they also present certain limitations. Traditional approaches often rely heavily on pre-selected materials, which may not fully reflect the authentic and diverse usage of collocations in real language contexts. This can limit learners' exposure to the variety of collocational patterns and reduce their ability to apply these patterns flexibly in different contexts. Furthermore, traditional methods may not adequately cater to individual learner differences, as they often adopt a one-size-fits-all approach without considering learners' specific needs, preferences, or prior knowledge (Nesselhauf, 2004).

In contrast, DDL offers distinct advantages by addressing some of the limitations inherent in traditional methods. DDL involves learners directly interacting with large, authentic corpora to discover and analyse collocational patterns on their own (Johns, 1991; Boulton, 2011). This approach promotes inductive learning, where learners actively engage with the language data, making their own observations and hypotheses about collocation use. Such an approach not only enhances learners' awareness of collocations but also helps them develop a deeper understanding of their usage in various contexts, which is often missing in traditional, more passive learning methods. Moreover, DDL supports a more personalised learning experience. By allowing learners to explore language corpora independently, they can focus on collocational patterns that are most relevant to their learning needs and interests. This individualised approach is particularly beneficial for addressing specific first language (L1) transfer issues, as learners can identify and compare collocations that are prone to interference from their L1. Additionally, the use of corpus tools and software in DDL enables learners to access vast amounts of authentic language data, providing them with a broader and more nuanced exposure to collocations than what traditional textbooks or teacher-selected materials can offer (Boulton and Cobb, 2017). Finally, DDL fosters learner autonomy by encouraging self-directed learning and critical thinking skills. As learners are involved in the discovery process, they are more likely to develop a sense of ownership over their learning, which can lead to increased motivation and engagement (Boulton and Vyatkina, 2021). This empowerment contrasts with the more teacher-centred approaches of traditional methods, where learners may remain passive recipients of knowledge rather than active constructors of their linguistic competence.

Data-driven Learning (DDL) on Collocation Learning. DDL, a novel approach in language education initiated by Johns in 1991, focuses on engaging learners with authentic language examples, primarily through corpora. This approach encourages learners to deduce language patterns inductively and gained recognition through Johns' endorsement of its effectiveness in increasing linguistic awareness and fostering independence in learners (Kwarikunda et al. 2022). At the heart of this approach is the concordancer, a software tool for analysing specific elements in a corpus. Despite its potential, DDL has faced challenges like data overload and the need for well-structured guidance. However, technological advancements, as noted by Boulton and Cobb (2017), have led to the creation of more accessible and intuitive DDL tools, making this approach more viable across different educational contexts.

DDL represents a significant shift in language teaching methodologies, urging learners to explore large text corpora to uncover linguistic patterns. This approach, pioneered by Johns in 1991, laid the foundation for fostering autonomous and sustainable learning experiences. DDL has proven particularly effective in helping learners master collocations, offering a novel and effective avenue in language education. The effectiveness of DDL is reinforced by Boulton and Vyatkina's (2021) extensive review, which examined over 400 studies and highlighted the growing application and interest in DDL for language learning. DDL has received acclaim for its role in enhancing vocabulary and promoting learner autonomy. Cobb's (1997) research demonstrated that learners could make significant strides in recognising language patterns through concordance presentation. Chan and Liou (2005) demonstrated that DDL significantly improves vocabulary acquisition, while Sun and Wang (2003) emphasised its effectiveness in fostering learner independence. Further research by Yeh et al. (2007) on the use of DDL and bilingual collocation concordancers has shown that these tools can effectively teach adjectives and their collocates, providing learners with authentic examples and immediate feedback. Furthering these insights, Liu and Jiang (2009) showed the effectiveness of DDL in offering explicit, context-aware grammar instruction and enhancing overall English proficiency. Rezaee et al. (2015) have investigated the role of scaffolding in improving collocational competence, emphasising the importance of structured support in the learning process. Basal (2019) has compared the effectiveness of online tools and traditional methods in increasing knowledge of adjective-noun collocations, finding that online tools can offer more dynamic and engaging learning experiences.

More recent studies, such as those by Chang and Sun (2022) and Crosthwaite and Steeples (2022) have affirmed DDL's effectiveness in elevating language proficiency, including collocational competence. Chang and Sun's (2022) meta-analysis found a moderate to significant effect of DDL on EFL vocabulary acquisition, implicitly supporting its impact on collocational knowledge. Li (2023) explored how combining DDL with indirect corrective feedback could improve the accuracy of collocation usage in English. Her findings indicated that this combination significantly enhanced collocation accuracy, outperforming dictionaries in rectifying collocation errors.

While the innovative and practical aspects of DDL are recognised, it also presents several challenges. These include the need for technical skills, the time required for corpus-related tasks, and difficulties in accessing computers during class (Chen and Flowerdew, 2018; Geluso and Yamaguchi, 2014; Yoon and Hirvela, 2004; Zare and Delavar, 2022). Besides, students often find it difficult to interpret data from corpora, leading to mixed experiences (Geluso and Yamaguchi, 2014; Yoon and Hirvela, 2004). For example, Yoon and Hirvela (2004) observed that students in a university writing course struggled with the time demands of analysing data and understanding concordance lines using the Collins COBUILD corpus. They concluded that while the approach significantly benefited learners' understanding of complex grammatical structures and usage patterns, it also posed challenges, such as the need for substantial instructor guidance and technical proficiency. Chambers (2007) pointed out that while adult learners may find the self-directed nature of DDL empowering, pre-tertiary students might require more structured guidance and support to benefit from corpus-based activities. This highlights a gap in the design and application of DDL tools and methodologies that consider the unique needs and developmental stages of younger learners. Flowerdew (2015) discussed the use of DDL in specialised fields such as English for Specific Purposes (ESP), where adult learners at the tertiary level benefit from the discipline-specific insights that corpora can provide. These applications often require a degree of linguistic

sophistication and familiarity with subject-specific terminology, skills typically developed at the university level. Given the complexity of the existing DDL tools, there is a pressing need for research and development of DDL tools that are specifically adapted to the cognitive and pedagogical requirements of pre-tertiary education. Implementing DDL effectively depends on having access to suitable corpora and tools, which can pose technical challenges, particularly in resource-limited settings (Zare and Delavar, 2022). These issues highlight concerns about integrating corpus consultation into standard language teaching practices (Chambers, 2019).

Rationale for the present study. Most DDL tools and methodologies are primarily designed for higher education or adult learners, highlighting a significant gap in applications tailored specifically to the cognitive and educational levels of pre-tertiary students, as discussed in the previous section. Moreover, there is often a disconnect between the introduction of new educational technologies and their integration with existing curricula. Research is needed to explore effective ways to incorporate DDL into standard pre-tertiary curriculums, ensuring it becomes an integral part of students' daily learning experiences rather than an extraneous component. While some studies investigate the implementation of DDL, there is typically a lack of thorough assessment regarding its impact on learning outcomes, particularly over the long term. There is a critical need for research aimed at evaluating how DDL can enhance pre-tertiary learners' understanding and use of collocations. Furthermore, DDL tools and methods developed within one cultural or educational context may not seamlessly translate to another, necessitating research into how these approaches can be adapted to various educational settings and cultural backgrounds. This adaptation is especially crucial in ensuring the relevance and effectiveness of teaching collocations to diverse learner populations.

This study aims to address these gaps in DDL research by specifically focusing on the teaching of English noun collocations to pre-tertiary Chinese learners. This focus is particularly pertinent given the substantial differences in lexical structures between Chinese and English, underscoring the need for tailored educational tools and strategies. Chinese primarily uses single, standalone morphemes that are monosyllabic and represented by individual characters, showcasing an isolative structure. In contrast, English employs a complex, fusional morphology involving roots, prefixes, and suffixes that combine to express grammatical relationships (Packard, 2000). Furthermore, English collocations such as "heavy rain" often do not directly translate into Chinese, where it becomes "大雨" (dà yǔ), highlighting differences in idiomatic expressions that can pose translation challenges (Wray, 2002; Zheng, Hu and Xu, 2022).

This research focuses on how DDL impacts these learners' comprehension and use of English noun collocations. Noun collocations play a pivotal role in academic writing by enabling writers to express complex ideas succinctly. The ability to use collocations such as "tackle the issues", "emergency response" and "social context" allows writers to convey precise meanings with fewer words, thereby enhancing readability and comprehension (Biber et al. 1999). This precision is crucial in academic writing, where clarity and brevity are highly valued. For novice writers, mastering noun collocations is particularly challenging due to their lexical density and generic nature. Academic texts often require a high density of information-packed noun phrases, which can be difficult for learners to produce and comprehend (Halliday, 1993). These collocations are not only common in academic discourse but also tend to be more abstract and less transparent in meaning compared to everyday language, making them harder to learn and use correctly (Liu and Lu, 2020).

One significant challenge in mastering noun collocations is the potential for cross-linguistic influence, such as L1 transfer. Learners often apply the collocational patterns of their first language to English, which can lead to errors and non-native-like usage. This issue is compounded by the fact that collocational patterns in English often do not have direct equivalents in other languages, leading to confusion and incorrect usage (He and Ang, 2023). Understanding and using noun collocations appropriately demands explicit instruction and extensive practice. Traditional language instruction methods may not sufficiently address the complexity of collocations, often focusing more on individual vocabulary items rather than on how words combine to form meaningful phrases. This gap necessitates innovative teaching approaches, such as DDL, which leverages real-language data to help learners recognise and practice authentic collocational patterns (Parkinson, 2015).

The study will examine the immediate and prolonged effects of DDL on noun collocation learning, seeking to understand how learners retain this knowledge over time, a concern highlighted by Boulton and Cobb (2017). The study will also evaluate learner confidence, which Lee and Lee (2020) identified as a key determinant of students' willingness to communicate and utilise their language skills. The study will also investigate learners' perceptions of DDL to evaluate its potential integration into mainstream language classes, a topic that Chen and Flowerdew (2018) discussed. Investigating learners' perceptions of DDL offers valuable insights into engagement, motivation, and the effectiveness of DDL across diverse educational contexts, informing adaptations necessary for different learner groups. Feedback from learners not only aids in refining pedagogical approaches and addressing practical challenges in technology and methodology but also contributes to theoretical developments in applied linguistics and educational technology. Such research fills significant gaps in existing literature by integrating the subjective experiences of learners, which have often been overlooked. Ultimately, understanding these perceptions can significantly impact the design and implementation of DDL, leading to more learner-centred approaches and improved educational outcomes.

Notably, previous DDL research primarily utilised complex corpus tools, leading to a demand for more user-friendly options for younger learners. This study, therefore, will assess the effectiveness of 'Corpusmate', a new, simpler corpus platform tailored to this demographic.

Beyond higher education, the study acknowledges the importance of noun collocations in pre-tertiary education, where they serve as foundational blocks for academic writing and advanced literacy skills. Early mastery of these collocations is crucial for students to develop a robust understanding of subject-specific vocabulary and to prepare for the complexities of academic writing in higher education. Thus, instructing and practicing noun collocations at this stage is instrumental for students' academic success and effective communication in their future educational pursuits.

Employing quantitative and qualitative research methods, including standardised language tests and learner feedback, the study aims to explore students' experiences with DDL and the use of the online corpus platform, 'Corpusmate' post-intervention.

The following questions guide the current study:

1. What is the effect of DDL on the learning of noun collocations by pre-tertiary EFL learners after the treatment and after three months?
2. What are these pre-tertiary EFL learners' perceptions of the DDL approach?

Methodology

The participants. This study investigated 70 Chinese learners of English, all aged 16, from a senior high school in Hubei Province, China. These students, who were in their Senior Two year at the time of the study, had previously passed the English language subject with at least 95 out of 150 marks during their Senior One year. Achieving a score of at least 95 out of 150 marks indicates a good foundational understanding of English. This threshold suggests that students who achieve at least 95 marks demonstrate above-average proficiency compared to their peers. This proficiency level is reflective of their ability to grasp intermediate-level English language concepts, perform well in reading comprehension, construct grammatically correct sentences, and understand spoken English at a moderate level (Fang et al. 2008).

The group consisted of 39 males and 31 females, reflecting the school's gender distribution. To ensure consistency in educational backgrounds, these students were selected from two classes that followed the same curriculum and used identical teaching materials. All participants were native Mandarin speakers learning English as a foreign language. A pre-study questionnaire confirmed that none of the participants had prior experience using corpus tools. The students were divided into two groups of 35: a control group and an experimental group. The control group continued their regular English classes and could use any resources except 'Corpusmate', an online corpus platform, during their revision before the collocation test. In contrast, the experimental group was instructed to use 'Corpusmate' during their treatment period.

Corpus tool and Corpora. The study utilised 'Corpusmate', an innovative online corpus platform to facilitate access to a vast collection of English language texts. Developed by Crosthwaite and Baisa (2023), 'Corpusmate' offers a streamlined and simplified experience in language data exploration tailored specifically for younger learners. Its design integrates the most effective features of existing tools into a cohesive digital environment, making it particularly suitable for secondary school students. This user-friendly platform is ideal for fostering independent language learning, providing an intuitive interface for linguistic exploration. The corpus has been compiled from 6 different spoken and written resources: the British Academic Written English (BAWE), TED talks, Simple English Wikipedia, BBC Teach, Elsevier corpus, and BNC 2014 Spoken. Instructions were given to the learners on how to use 'Corpusmate' to explore noun collocations. The interface of 'Corpusmate' is shown in Fig 1.

DDL training. For the two-month training, the learners were provided with selected nouns from the Academic Word List (AWL) to explore noun collocations, i.e. pre-modifiers and nouns that form collocations. AWL, developed by Coxhead (1998) at Victoria University of Wellington, New Zealand, comprises 570 word families. These words were included in AWL for their high frequency across a wide array of academic texts. Significantly, the AWL excludes words found in the top 2000 most common English words, known as the General Service List, thereby tailoring it to academic settings. Designed primarily for academic purposes, the AWL serves as a resource for teachers preparing students for tertiary education and for students independently aiming to acquire vocabulary critical for college and university studies. The words in the AWL are categorised into ten distinct Groups. This categorisation is based on frequency, with the words in the first group being the most common and those in the tenth group being the least common within academic texts. This systematic arrangement assists learners in prioritizing their study of these words in accordance with their frequency of use in academic contexts. In this study, learners were asked to explore 10 nouns from each Group in AWL for each session. The DDL training sessions were conducted twice a week, with each session lasting 80 min.

Steps in conducting DDL training. To conduct DDL training using specific nouns from the AWL developed by Coxhead (1998), we followed several steps to create an engaging and effective educational experience.

Step 1: Selection of Nouns from AWL. We selected nouns from the AWL that align with the learners' proficiency level to ensure that the vocabulary is challenging enough to facilitate learning without being overly difficult or discouraging. This selection process was guided by two English teachers who provided their expertise to ensure the appropriateness of the chosen nouns for the students' academic level. We chose the following noun as example:

Policy

Step 2: Creating Collocations. We asked the learners to use the selected noun to form three types of collocations: verb-noun,

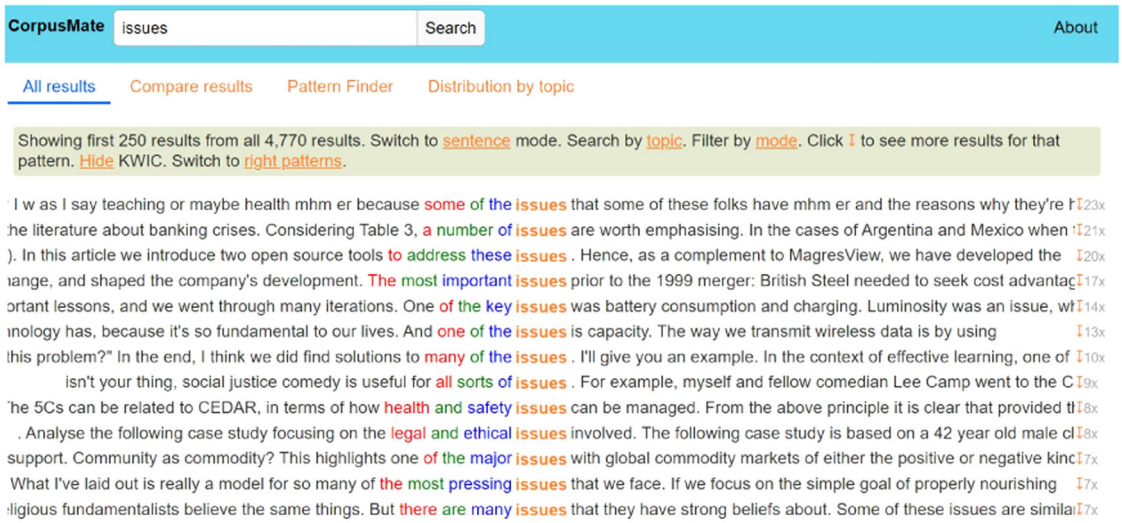

Fig. 1 The interface of 'Corpusmate' by using "issues" as an example.

noun-noun, and adjective-noun. Here are examples for each type using the selected noun:

Verb-Noun Collocations:
Implement policy
Formulate policy
Noun-Noun Collocations:
Policy maker
Policy revision
Adjective-Noun Collocations:
National policy
Effective policy

Step 3: Compilation of authentic examples. We asked learners to extract authentic sentences from 'Corpusmate' that showcase how these collocations are used in different contexts.

Step 4: Analysis and discovery. We asked learners to identify patterns or rules about the use of these nouns in various collocations. This discovery-based approach helps deepen their understanding of the language.

Step 5: Guided practice. After learners have explored these examples, we guided them through structured practice activities. This involved using collocations to complete sentences, and creating their own sentences using these collocations.

Step 6: Feedback and revision. We provided feedback on learners' attempts and encouraged them to revise their sentences or try new combinations based on the feedback.

It is worth noting that during the DDL training phase, students were engaged in forming and analysing a wide variety of collocations, including both standard and non-standard examples. This approach was intentionally chosen to encourage a deeper understanding of collocational patterns and to foster flexibility in language use. The training was not limited to a predetermined set of collocations; instead, it was designed to expose students to a broad spectrum of collocational possibilities, reflecting natural language use.

For the control group, English teachers in the control group were informed about the study's focus on noun collocations and were instructed to integrate this focus into their regular teaching practices. This means that while the teachers continued to deliver the standard curriculum, they also incorporated additional emphasis on noun collocations during their lessons. This ensured that the control group received relevant exposure to collocations, like what was emphasised in the study. The participants were given the same set of nouns extracted from the AWL as the experimental group. To ensure that their learning was aligned with the collocations tested in the assessment, the control group was instructed to focus specifically on identifying and learning collocations involving these given noun heads. They were guided to explore various resources, such as collocation dictionaries, thesauruses, and the internet, to find common nouns, verbs, and adjectives that can precede each noun head provided. This guided exploration aimed to ensure that the control group's learning activities were focused on the same noun collocations emphasised in the test, despite not using the 'Corpusmate' tool or any other corpus-based resources.

While the control group did not use corpus-based tools, the learning process was carefully structured to parallel the objectives of the DDL training. This ensured that both groups were studying the same types of collocations, allowing for a fair comparison of the effectiveness of the DDL approach versus more traditional, resource-based methods. By aligning the content of the control group's instruction with the test items, we aimed to provide a valid measure of the impact of DDL training on collocation acquisition.

Collocation test. To measure the effectiveness of the DDL approach, a 60-item test was developed and administered to all learners before and after the training. The items focused on three prevalent types of noun collocations: verb-noun (e.g., "tackle [the] issues"), noun-noun (e.g., "emergency response"), and adjective-noun (e.g., "social context"), with 20 items dedicated to each category. This distribution ensured equal representation for each type of noun pre-modifier. These collocation types were chosen for their pedagogical significance, as they are commonly used in the language yet pose unique challenges for ESL and EFL learners due to their varied levels of fixedness and potential for cross-linguistic influence, such as L1 transfer (Nesselhauf, 2004; Pérez-Paredes and Sánchez-Tornel, 2014). The test design drew inspiration from previous studies on collocation knowledge assessment (such as Boers et al. 2014; Gyllstad and Schmitt, 2019) and adopted a binary-choice format, requiring learners to choose the correct pair of collocating words. To develop the distractor items for the test, two methods were employed: firstly, incorporating common collocational mistakes found in prior research on ESL and EFL learners (for example, in Lu, 2017), and secondly, integrating input from three native English language instructors from an English language centre in Hubei province, China. The test was administered three times to ensure a comprehensive assessment: before, immediately after, and three months post-training. This approach aimed to gauge both immediate and long-term learning outcomes. Additionally, we randomised the order of items in each session to minimise biases related to question sequence or random guessing.

To ensure the binary-choice test accurately reflected the content and activities covered during the training, a systematic approach was employed in the test design. The test items were developed based on the specific types of collocations and patterns that were emphasised during the training sessions. For example, if students practiced forming noun + noun, adjective + noun, and verb + noun collocations, the test included a proportional number of items from each of these categories. Furthermore, the test items were selected to mirror the level of difficulty and the types of collocational combinations that students encountered during the training. Besides, the test was designed to assess not just the recall of specific collocations but also the students' ability to apply the principles and patterns they learned to new collocations. This was done to evaluate their understanding of collocational usage beyond the examples directly covered in the training. By aligning the test content with the training focus in this manner, we aimed to ensure that the test items were representative of the learning objectives and that students were fairly assessed on their ability to form and recognise collocations as practiced during the training.

The study included both an experimental group, which received DDL training, and a control group, which did not. This design allowed for a comparison of the impact of the DDL training on noun collocation learning between the two groups. The collocation test was carefully designed to align with the DDL training and conventional classroom teaching. The selected nouns from AWL were provided to both groups, ensuring that the test items reflected the specific types of collocations and activities learners engaged with during the training and conventional classroom learning. This alignment allowed for an accurate assessment of the DDL training's effectiveness in comparison with conventional classroom learning. This approach ensured that any improvements observed in the experimental group's test

scores could be attributed to the DDL training, as the test directly assessed the content covered during the training.

By comparing the pre- and post-test results of the experimental group with those of the control group, which did not receive DDL training, the study could isolate the effect of the DDL intervention. This comparative analysis provided a clear measure of the DDL approach's impact on learners' ability to understand and use noun collocations. The control group served as a baseline to determine the natural progression in collocational competence, while the experimental group's results highlighted the added value of the DDL training.

Learner experience and perception with corpus-based tools: questionnaire survey. We used questionnaires to obtain feedback from learners in the experimental group to evaluate the DDL technique's effectiveness. These questionnaires were designed to capture learners' perspectives on several key aspects: the overall utility of the DDL approach, their experience with the online corpora and the software tool 'Corpusmate', and their confidence in independently learning English collocations. This approach aligns with the findings of Chang and Sun (2022), who emphasised the importance of learner feedback in assessing the impact of innovative teaching methods on student motivation and autonomy. The questionnaire was adapted from Crosthwaite and Steeples (2022) and administered directly following the completion of the post-training test. It was divided into two sections, including perceptions of corpus training and perceptions of DDL for improving knowledge or use of collocations, which would provide a comprehensive view of the learners' experiences and opinions.

Data analysis. In addressing Research Question 1 regarding DDL's effects on noun collocation learning, the scores for the collocation tests were calculated by assigning one point for each correct answer. The dataset was reviewed to confirm that it met essential criteria (like normal distribution) for an Analysis of Variance (ANOVA) test. This study used a repeated-measures ANOVA to compare scores from three different testing times: before, immediately after, and three months following the training. This method was chosen as it effectively examines differences in average scores across various experimental conditions over multiple time points (Larson-Hall, 2015). Additionally, it accounts for individual variances within and between groups, allowing for adjustments for any initial differences in knowledge, as measured in the pre-test. In analysing the data, both the p-value (with a standard threshold of 0.05) and the effect sizes (using partial eta-squared) were calculated. For effect sizes, ηp2 values of 0.01 were considered small, 0.06 medium, and 0.14 large (Cohen, 1988). The analysis focused on overall changes in collocation knowledge across the entire test. Cohen's d would be utilised to measure the effect size of the intervention between the experimental and control groups, should there be any differences in the pre-test scores. Cohen's d is a standardised measure of effect size that expresses the mean difference between two groups in terms of standard deviation, allowing for the comparison of effect sizes across different studies and contexts (Cohen, 1988). This test provides a guideline for interpreting the magnitude of effect size. A small effect size indicates a modest difference between groups, while a large effect size indicates a substantial difference.

To address Research Question 2, which aimed to evaluate learners' perceptions of the DDL training and the use of the 'Corpusmate' tool, the questionnaire data were designed and analysed by examining learners' levels of agreement with each statement provided in the survey. This analysis utilised a five-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree," allowing for a nuanced understanding of the participants' attitudes and experiences regarding the DDL approach and the specific functionalities of the 'Corpusmate' tool.

Table 1 Five-point likert scale.

Strongly Disagree
Disagree
Neutral
Agree
Strongly Agree

Likert scale responses. The questionnaire employed a five-point Likert scale for each statement, where participants could indicate their level of agreement. The scale is presented in Table 1.

Questionnaire data collection and analysis. Learners were asked to respond to a series of statements regarding their experiences and perceptions of the DDL training and the 'Corpusmate' tool. These statements were designed to capture various dimensions of their learning experience, including engagement, usefulness, ease of use, and overall satisfaction. After collecting the responses, the questionnaire responses were checked for completeness. Each statement's responses were analysed to assess the levels of agreement or disagreement. This involved calculating the percentage of participants who selected each response option on the Likert scale. For example, 58% of the participants agreed or strongly agreed that the 'Corpusmate' tool was useful in learning noun collocations, while only 11% disagreed or strongly disagreed. Such findings indicate a general consensus on certain aspects of the tool's user-friendliness. The analysis also looked at statements with higher neutrality levels to identify areas where perceptions were less definitive, indicating potential areas for further training or tool improvement. This method allowed for a clear representation of the distribution of responses across the different levels of the Likert scale. The percentages were calculated for each statement to understand the overall sentiment of the learners.

Steps in the analysis

Data tabulation: All responses were first tabulated, categorising each response according to the five-point Likert scale. This step involved organising the data into a structured format suitable for percentage calculation.

Percentage computation: To standardise the data and allow for easier comparison across different statements, the raw counts were converted into percentages. This was done by dividing the number of responses in each category by the total number of respondents for that statement, then multiplying by 100. The formula used for this calculation is:

$$\text{Percentage} = \left(\frac{\text{Number of respones in category}}{\text{Total number of respondents}}\right) \times 100$$

Visualisation and interpretation: The calculated percentages were then visualised using bar charts to provide an easily interpretable overview of the data. This visualisation helped to quickly identify trends and patterns in the learners' responses. For instance, a high percentage of "Agree" and "Strongly Agree" responses would indicate a positive perception of a particular aspect of the DDL training. Once the percentages were calculated and visually represented, the next step was to analyse patterns and trends in

Table 2 Collocation test scores across the groups (mean and SD).

		Pre-test		Post-test		Delayed post-test					
Group	**N**	**M**	**SD**	**M**	**SD**	**M**	**SD**	***df***	**F**	***p***	η_p^2
Experimental	35	63.25	7.72	67.16	8.65	67.46	10.52	2	7.232	0.001	0.190
Control	35	60.32	7.41	61.26	8.01	61.29	7.72	2	4.861	0.013	0.118

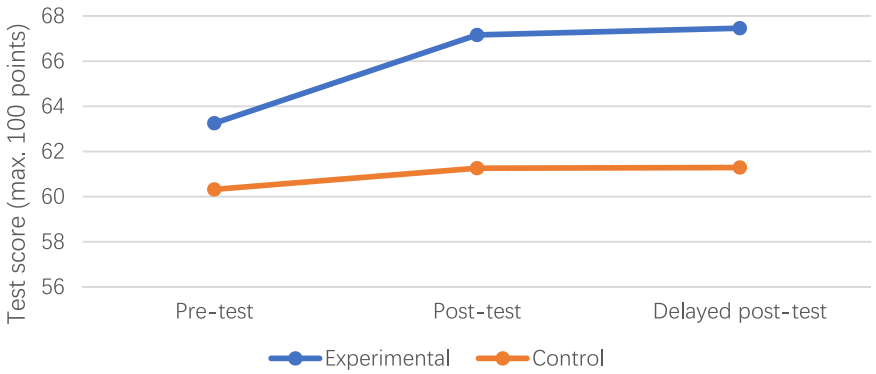

Fig. 2 The mean score on the collocation test for the two groups.

the data. This involved identifying statements with high levels of agreement or disagreement, as well as those with a significant proportion of neutral responses. For example, if a large percentage of respondents consistently selected "Agree" or "Strongly Agree" across multiple statements, this pattern would suggest a generally positive perception of the DDL training and the 'Corpusmate' tool. Conversely, a high percentage of "Neutral" responses could indicate areas where learners were undecided or had mixed feelings, suggesting potential areas for further investigation or improvement.

Results

RQ1: the effects of DDL on the learning of noun collocations by pre-tertiary EFL learners. Table 2 and Fig. 2 illustrate the use of a repeated-measures ANOVA to evaluate the efficacy of DDL training compared to traditional language teaching methods. This assessment was done by comparing the scores of the collocation tests given to both the control and experimental groups at three different times.

The experimental group demonstrated a significant increase in mean scores from the pre-test to the post-test and maintained similar performance in the delayed post-test. Conversely, the control group showed only a slight increase in mean scores, remaining relatively stable across the three collocation tests. For the experimental group, the F-value of 7.232 and a p-value of 0.001 indicate that the differences in scores across the three tests are statistically significant. The eta squared value ($\eta p2 = 0.190$) suggests a medium to large effect size, indicating that the DDL training had a significant impact on the scores. In contrast, for the control group, the F-value of 4.861 and a p-value of 0.013 also denote substantial differences, albeit less pronounced than in the experimental group. The eta squared value ($\eta p2 = 0.118$) implies a small to medium effect size, highlighting the comparatively minor impact of conventional teaching on collocation learning compared with DDL training.

The findings indicate that the treatment given to the experimental group successfully enhanced their performance in the collocation tests. This enhancement is demonstrated by a notable rise in scores from the initial test to the subsequent test, with this improved performance maintained in the later follow-up test. The performance of the control group remained constant, further emphasising the impact of the DDL training. The outcomes demonstrate a substantial increase in the learners' knowledge of noun collocations post-training. The significant rise in mean scores from the pre-test to the post-test robustly supports the efficacy of the DDL approach implemented in this study. The statistical evidence strongly suggests that the learners benefited from the DDL intervention, leading to improved understanding and use of noun collocations in English. This trend persisted three months later, as evidenced by the delayed post-test.

With regard to the pre-test scores, the scores indicate a notable difference in English language proficiency between the control group and the experimental group, which can significantly affect the outcomes and interpretations of the study. The experimental group started with a higher baseline in English proficiency ($M = 63.25$, $SD = 7.72$) compared to the control group ($M = 60.32$, $SD = 7.41$). This initial advantage could imply that the experimental group had either more effective prior learning or inherent capabilities that could influence their ability to benefit from the intervention. Further statistical adjustment, i.e. Cohen's d was utilised to measure the effect size of the intervention between the experimental and control groups, taking into account the initial differences as observed in the pre-test scores. Cohen's d is a measure of effect size that expresses the mean difference between two groups in terms of standard deviation, allowing us to see the relative effect of the intervention considering their starting proficiency levels. The Cohen's d values calculated for both the experimental and control groups are as follows:

Experimental Group: Cohen's $d = 0.517$

Control Group: Cohen's $d = 0.124$

The effect size for the experimental group is moderate (approximately 0.517), indicating a significant effect of the intervention on improving the English proficiency from pre-test to post-test. The effect size for the control group is much smaller (approximately 0.124), indicating a much smaller change in scores from pre-test to post-test, which might be attributed to

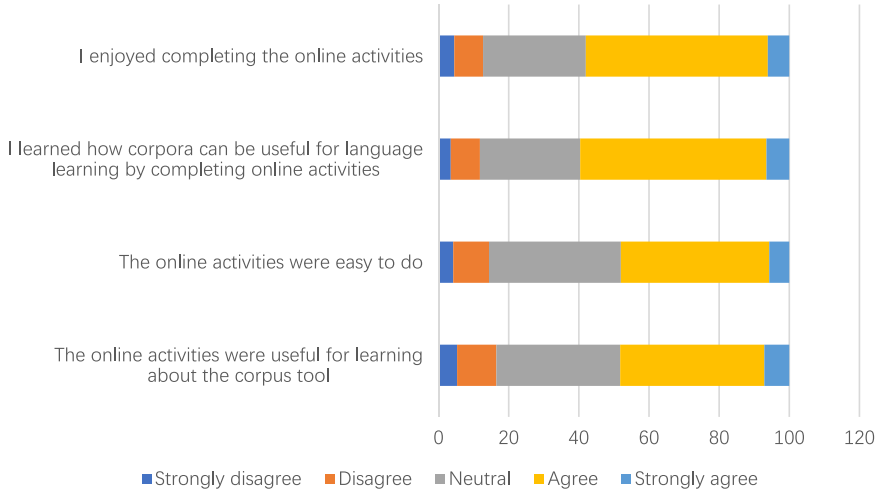

Fig. 3 Learners' survey perceptions of DDL training.

natural fluctuation or minor external influences rather than a structured intervention. The much larger effect size in the experimental group compared to the control group suggests that the DDL intervention had a meaningful impact on the participants' English language proficiency. This demonstrates the effectiveness of the intervention, especially when considering the initial proficiency levels indicated by the pre-test scores. This analysis helps validate the effectiveness of the intervention while accounting for initial proficiency differences between the groups.

RQ 2: Pre-tertiary EFL learners' perceptions

Perceptions of the DDL training. Figure 3 presents learners' perceptions of their DDL training activities, providing a comprehensive view of their experiences and attitudes towards the online activities designed to enhance their understanding of language learning through corpora. A significant proportion of participants expressed enjoyment in completing these online activities, with the majority either agreeing or strongly agreeing with this statement. This high level of enjoyment suggests that the activities were well-received and likely designed in a way that resonates with participants' interests and preferences, making the learning process engaging and enjoyable. A small number of participants remained neutral, indicating that while they did not have strong positive feelings, they also did not find the activities disagreeable. Very few participants disagreed, and there were no strong objections, underscoring the overall positive reception of the online activities.

When asked whether they learned how corpora can be useful for language learning through online activities, most participants again responded positively. The majority agreed or strongly agreed that the activities enhanced their understanding of corpora's educational value. This suggests that the activities were effective in demonstrating the practical applications of corpora for language learning, helping learners see the relevance and benefits of using real-life language examples in their studies. However, some participants were neutral or disagreed, indicating that not all learners found the value of corpora immediately apparent or perhaps did not engage with the material as deeply. This variation highlights the need for further refinement in how the educational value of corpora is communicated and integrated into the learning activities.

Responses regarding the ease of completing the online activities were more varied. While many participants agreed or strongly agreed that the activities were easy to do, a considerable number of responses spanned all five categories. This suggests that while the majority found the activities manageable, there was a notable portion of learners who either felt neutral about the ease or found the tasks challenging. The presence of neutral responses and some disagreement highlights the need for simplifying instructions or providing additional support to ensure all learners can comfortably engage with the activities. Ensuring that activities are accessible and not overly complex is crucial for maintaining high levels of engagement and preventing frustration among learners.

Regarding the usefulness of the online activities for learning about the corpus tool, most participants displayed a favourable view, with many agreeing or strongly agreeing on the tool's utility. This indicates that learners found the activities helpful in understanding and using the corpus tool, which is a positive outcome for the DDL approach. Nevertheless, a modest number of participants were neutral, and a few disagreed, suggesting that while the tool was generally well-received, some learners might not have found it as beneficial or might require further guidance to fully appreciate its capabilities. Some of them might have found the 'Corpusmate' or the activities too complex or challenging to use, leading to frustration or disengagement. Besides, the level of instructional support provided during the DDL training might have varied, influencing learners' ability to effectively use the corpus tool. This feedback points to the importance of continuous evaluation and improvement of teaching methods and educational tools to ensure they meet the diverse needs of all learners.

Overall, the feedback from participants indicates a generally positive attitude towards the DDL activities and the use of the corpus tool. The high levels of agreement on enjoyment and perceived educational value suggest that the activities were engaging and effective for most learners. However, the variability in responses regarding the ease of activities and the usefulness of the corpus tool points to areas for potential improvement. Enhancing the clarity of instructions, providing more support for challenging tasks, and ensuring all learners understand the benefits of using corpora can help maximise the positive impact of DDL activities. These insights are crucial for educators and developers to refine and optimise the design and implementation of DDL tools, ensuring they meet the diverse needs of learners and enhance their language learning experience.

Perceptions of Corpora and 'Corpusmate' for understanding and using noun collocations. Figure 4 presents learners' perceptions regarding the efficacy of corpora and 'Corpusmate' in enhancing their understanding and usage of noun collocations. Firstly, the

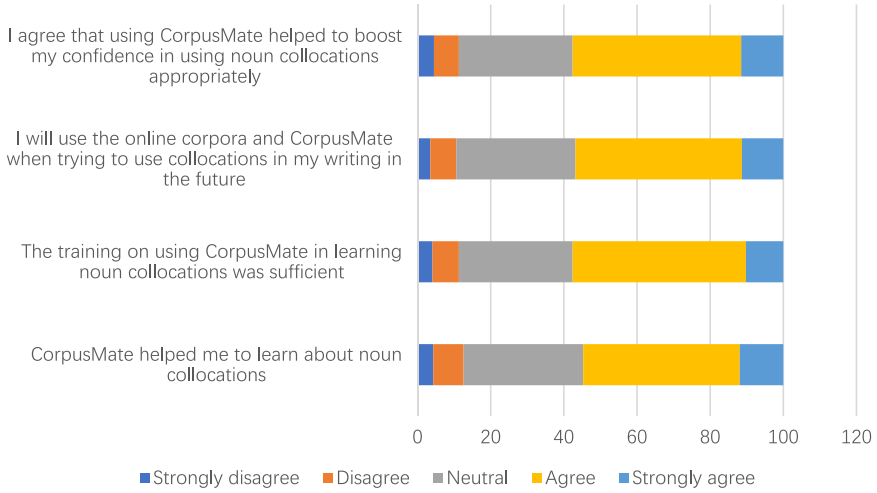

Fig. 4 Learners' survey perceptions of the usefulness of corpora and "Corpusmate" in learning noun collocations.

impact of 'CorpusMate' on boosting confidence in using noun collocations is overwhelmingly positive. A significant majority of participants either agreed or strongly agreed that using 'CorpusMate' helped boost their confidence in appropriately using noun collocations. This suggests that the tool effectively demystifies noun collocations, making learners feel more competent and self-assured in their usage. However, a small fraction of participants expressed neutrality or disagreement, indicating that while the tool was broadly effective, a few learners did not experience a significant confidence boost. This discrepancy might be due to varying initial proficiency levels or personal learning preferences, suggesting a need for more tailored approaches to accommodate all learners.

Secondly, many participants indicated that they would continue to use 'CorpusMate' and online corpora for collocation in their future writing, with most responses falling into agree or strongly agree categories. This shows a strong recognition of the practical value of these tools beyond the immediate learning context, implying that learners see long-term benefits in integrating 'CorpusMate' into their writing practices. Nevertheless, the presence of neutral responses indicates that some learners are uncertain about incorporating these tools into their future writing. This ambivalence might stem from a lack of confidence in independently using the tools or the perceived complexity of integrating them into their writing workflows.

Regarding the sufficiency of training on using 'CorpusMate', most participants agreed or strongly agreed that the training provided was sufficient, indicating that the instructional component was effective for most learners. Proper training is crucial for maximising the utility of educational tools, and these responses suggest that the training met the needs of most users. However, a notable portion of participants remained neutral, and a few disagreed, implying that some learners felt the training could be more comprehensive. This feedback highlights the need for continuous improvement in training materials, perhaps by including more detailed guides, practical examples, or additional support sessions to ensure all learners feel adequately prepared to use 'CorpusMate'.

Lastly, the majority view 'CorpusMate' as useful for learning about noun collocations, with most participants agreeing or strongly agreeing with this statement. This reflects the tool's effectiveness in teaching specific language structures, a critical component for learners aiming to enhance their collocational knowledge. Despite the positive reception, there is a modest group of neutral responses and a few disagreeing, indicating that not all learners found 'CorpusMate' equally beneficial. These mixed feelings could be due to individual differences in learning styles or the need for more personalised approaches within the tool to cater to different learner needs.

Overall, the feedback from learners reveals a generally positive attitude towards 'CorpusMate' and its role in learning noun collocations, with high levels of confidence and intentions to use the tool in future writing. This indicates that 'CorpusMate' is perceived as a valuable educational resource that supports learners in mastering complex linguistic elements. However, the variability in responses regarding the sufficiency of training and the ease of use suggests areas for improvement. To address the concerns of those who felt the training was insufficient, educators and developers could provide more comprehensive training sessions, detailed user guides, and interactive tutorials to ensure all learners can effectively utilise the tool. Offering ongoing support, such as Q and A sessions, discussion forums, or additional help resources, could help learners who remain neutral or uncertain about the tool's value. Regularly collecting and integrating user feedback can help identify specific areas where learners face challenges, allowing for targeted improvements in both the tool and its training materials. Developing adaptive learning pathways within 'CorpusMate' that cater to different proficiency levels and learning styles could enhance its effectiveness and make it more universally beneficial. By addressing these areas, 'CorpusMate' and similar DDL tools can further enhance their educational impact, ensuring they meet the diverse needs of learners and support their ongoing language development effectively.

Discussion

The current study's results reveal a notable improvement in learners' knowledge of noun collocations after DDL training, aligning with existing research that posits collocation acquisition as a gradual process independent of the teaching method (Chen, 2019; Granger, 2019). This study stands out from prior DDL research on collocation learning by not restricting learners to a predetermined set of noun collocations. Instead, learners use nouns derived from the AWL to explore noun collocations freely on the 'Corpusmate'. This approach offers a more authentic context for language learning and use compared to the more common practice of testing with a set range of items from the same domain to evaluate explicit receptive knowledge (e.g., Chan and Liou, 2005). By not being confined to a predetermined set of

collocations, learners can take charge of their learning process, making decisions about which collocations to investigate based on their interests and needs. This autonomy can lead to increased engagement and motivation, as learners feel more empowered and invested in their learning journey. More importantly, the unrestricted exploration of collocations fosters a discovery-based learning environment. Learners are encouraged to investigate and discover language patterns on their own, which can lead to more meaningful and long-lasting learning experiences. Besides, the DDL training in this study allows for a more comprehensive exploration of the language. Learners can see how noun collocations function across different contexts and genres, gaining insights into their versatility and usage. This broad exposure helps learners build a more robust and flexible understanding of collocations, which is essential for achieving fluency and accuracy in language use.

The findings of the present study echo the positive outcomes reported in the literature where DDL is effective. Past studies (e.g., Daskalovska, 2015; Liu and Ma, 2011; Vyatkina, 2016a) have highlighted DDL's impact on enhancing learners' collocational proficiency. The current study adds to this body of research, highlighting the effectiveness of DDL, particularly in learning noun collocations. This DDL training was enhanced by using the AWL and online corpus platforms like 'Corpusmate', which significantly improved learners' comprehension and application of English collocations. A distinctive outcome of our investigation is the sustained improvement in the experimental group's collocation knowledge, which remained stable three months post-training. This contrasts with prior DDL research, where immediate gains often diminished after a short period (Çelik, 2011; Chan and Liou, 2005; Daskalovska, 2015). The prolonged engagement with corpus consultation in this study (2 months) may have contributed to this retention, a feature not typical in earlier studies, which often limited DDL exposure to a few sessions (Boulton and Cobb, 2017). While extended interventions have been suggested to correlate with greater linguistic gains (Boulton and Cobb, 2017; Lee et al. 2019), our results provide further empirical support for this claim, although the exact relationship between intervention duration and knowledge retention warrants more systematic investigation. Furthermore, the control group's modest yet steady progress suggests the benefits of using preferred and familiar language tools, which participants are likely to continue to utilise beyond the study's intervention phase. While DDL methods are effective, traditional study tools retain value in language learning, contributing incrementally to learners' collocational knowledge. In conclusion, the study confirms the efficacy of DDL in collocation learning. It highlights the importance of sustained engagement and the potential benefits of integrating conventional language study tools alongside innovative DDL approaches.

The results of this study present the favourable views learners hold regarding DDL training, with tools like 'CorpusMate' significantly enhancing collocation knowledge—a sentiment shared by previous research from Geluso and Yamaguchi (2014), Rezaee et al. (2015), and Vyatkina (2016b). The findings here further underscore the importance of the user-friendly interface of 'CorpusMate', which plays a crucial role in promoting positive attitudes toward DDL, even among those learners who may have reservations about online learning activities. Most learners will keep using 'CorpusMate' after the training, highlighting the platform's effectiveness in bolstering confidence in using collocations. For pre-tertiary and young learners, in particular, the simplicity of 'CorpusMate' has made engaging with concordance data far less daunting. This accessibility shows the viability of 'CorpusMate' in a high school context, where ease of use is essential to motivate students to adopt new educational technologies.

The DDL approach's alignment with form-focused instruction within communicative language teaching, as outlined by Nation (2001) and Schmidt (2001), is evident in our study. The emphasis on noticing linguistic features through corpus activities complements meaning-focused aspects of language learning, offering a nuanced approach that traditional methods may lack. The integration of DDL brings a unique contribution to the language learning process. When carefully monitored and guided by teachers, these activities enable learners to 'notice' certain features of target words within context, a concept emphasised by Schmidt (2001). This aspect of DDL is especially significant, as traditional methods often do not sufficiently highlight these features, as Willis (2011) noted.

The scores for the experimental group remained high in the delayed post-test, indicating that the learning gains achieved through DDL training were sustained over time. This suggests that the DDL approach not only helps learners acquire collocational knowledge but also aids in long-term retention. The DDL approach involves engaging with authentic language data through corpora, which helps learners see collocations in various real-world contexts. This repeated exposure and practice facilitate deeper cognitive processing, leading to better retention and recall of collocational knowledge over time. Furthermore, DDL promotes autonomous learning by encouraging learners to explore and analyse language patterns independently. This skill development enables learners to continue applying their collocational knowledge beyond the initial training period, contributing to sustainable learning outcomes. By using corpora, learners engage with collocations in meaningful contexts, which enhances their understanding of how collocations function in real language use. This contextual learning aids in transferring knowledge from the classroom to actual language use, supporting long-term retention and application. The interactive and discovery-based nature of DDL can increase learner engagement and motivation. When learners are more engaged, they are likely to invest more effort in learning, which can lead to better long-term outcomes.

Critically examining the study's results in light of past literature reveals a nuanced picture of DDL's role in language instruction. While the long-term retention and continued use of corpus tools like 'Corpusmate' suggest a sustainable impact on language learning, the study also highlights the need for more granular research into the specific factors contributing to these outcomes. The past literature, particularly the work of Chen and Flowerdew (2018), stressed the importance of learner autonomy in successfully applying DDL to academic writing. Our study supports this notion but also suggests that longevity and frequency of engagement with corpus tools may play a significant role in the sustainability of learning gains. The ongoing use of corpus tools outside classroom settings aligns with Chen and Flowerdew's (2018) assertion that learners should extend their corpus interactions beyond teacher-led environments. However, there remains a discrepancy between learners' reported intentions and their actual continued use of these tools, which points to a potential gap between the classroom and real-world application that future studies need to address. Moreover, the current study underscores the importance of considering learner characteristics and the external usage of corpus resources, which, while challenging to control in quasi-experimental designs, are critical for understanding the ecological validity of DDL interventions.

This research supports the value of DDL in fostering long-term language learning benefits. DDL empowers learners to be autonomous, providing them the means for self-guided exploration and study. This approach imparts immediate linguistic abilities and encourages continuous improvement in language skills. Consequently, DDL stands out as a robust and influential teaching method with the potential to transform language

education practices significantly. A key advantage of the DDL method is its direct introduction of learners to corpora, enhancing their skills for independent exploration and analysis. Such independence in learning, as emphasised by Johns (1991), is essential for the sustained effectiveness of language instruction over time. Future research could aim to disentangle the effects of various DDL components by designing studies that allow for the direct investigation of key variables, such as the length and novelty of the intervention or training, as noted in comprehensive reviews by Boulton and Cobb (2017) and Lee et al. (2019). An exploration into how the duration of DDL engagement affects the durability of acquired knowledge would be particularly beneficial from a pedagogical standpoint.

This study has several limitations that should be acknowledged. Firstly, the focus on overall impact rather than specific collocations limits the granularity of the findings, preventing a detailed understanding of which particular collocations exhibited significant improvements. The lack of item-based analysis means that potential outliers or exceptionally memorable items that might have skewed the results were not identified. Additionally, the sample size was relatively small, and the study did not incorporate baseline comparisons with traditional, non-DDL methods or other educational tools, which could provide a more comprehensive context for interpreting the results.

Future research should address these limitations by conducting item-based analyses to identify specific collocations that show significant improvement and to understand the reasons behind these gains. Larger and more diverse sample sizes would enhance the generalisability of the findings. Incorporating baseline comparisons with traditional teaching methods or other educational tools would provide a more robust context for evaluating the effectiveness of DDL. Longitudinal studies are recommended to examine the sustained impact of DDL on collocation learning over a more extended period, providing insights into long-term retention and application of collocational knowledge. Finally, qualitative feedback from learners should be collected to gain deeper insights into their experiences and perceptions, which could inform the refinement of DDL approaches.

Conclusion

The present study underlines the powerful potential of DDL in facilitating a more autonomous and sustainable approach to learning English collocations among Chinese learners. DDL fosters a sense of ownership and independence in the learning process by empowering learners to engage with large bodies of authentic English texts directly. This sense of autonomy is a critical factor in sustainable language education, as it instils in learners the ability and confidence to navigate and continue their language learning journey independently beyond formal instruction.

Furthermore, the findings of the present study have broader implications for autonomous and sustainable language education. The use of online corpora and concordancers as primary resources for language learning can be extended to learning other aspects of the English language and other languages. By transforming pre-tertiary learners from passive recipients of knowledge into active explorers of language patterns, DDL can revolutionise language education, making it more learner-centred, engaging, and sustainable. Given the potential of DDL in language instruction, language teachers must be well-versed in utilising online corpora and concordancers. Hence, professional development programmes could include training on integrating DDL into curriculum planning and teaching strategies. While this study focused on Chinese learners, the findings have broader applicability, pointing to the potential for DDL to enhance language instruction for diverse learner profiles.

In summary, DDL offers a promising approach to English language education, fostering learner autonomy and promoting long-term retention and application of learning. Although this study focused on Chinese learners of English, the principles and strategies explored here have broader applicability in language education. As we move into an increasingly digital age, adopting data-driven, technology-enhanced learning approaches like DDL could become the norm, paving the way for a more autonomous and sustainable future in language education. Despite these encouraging findings, more research is needed to explore the potential of DDL further in diverse learning contexts and for various learner profiles. The field of language education could greatly benefit from such efforts, paving the way for innovative, effective, and sustainable teaching and learning practices.

Data availability

The datasets generated and analysed during the current study are not publicly available due to their inclusion in an ongoing project. However, they can be obtained from the corresponding author upon reasonable request.

Received: 15 January 2024; Accepted: 30 October 2024;
Published online: 11 November 2024

References

Basal A (2019) Learning collocations: Effects of online tools on teaching English adjective-noun collocations. Brit J Educ Technol 50(1):342–356. https://doi.org/10.1111/bjet.12562

Barfield A (2009) Following individual L2 collocation development over time. In: Barfield A, Gyllstad H (eds) Researching collocations in another language. Palgrave Macmillan, New York, pp 208–223. https://doi.org/10.1057/9780230245327_16

Bednarek M, Crosthwaite P, García AI (2020) Corpus linguistics and education in Australia. Aust Rev Appl Linguist 43(2):105–116. https://doi.org/10.1075/aral.00029.edi

Biber D, Conrad S (1999) Lexical bundles in conversation and academic prose. In: Hasselgard H, Oksfjell S (eds) Out of corpora: studies in honor of Stig Johansson. Rodopi, Amsterdam, pp 181–190

Biber D, Johansson S, Leech G, Conrad S, Finegan E (1999) Longman Grammar of Spoken and Written English. Longman, London

Boers F, Eyckmans KJ, Stengers H, Demecheleer M (2006) Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. Lang Teach Res 10:245–261. https://doi.org/10.1191/1362168806lr195oa

Boers F, Demecheleer M, Coxhead A, Webb S (2014) Gauging the effects of exercises on verb-noun collocations. Lang Teach Res 18(1):54–74. https://doi.org/10.1177/1362168813505389

Boulton A (2011) Data-driven learning: The perpetual enigma. In: Gozdź-Roszkowski S (ed) Explorations across languages and corpora. Peter Lang, Frankfurt am Main, pp 563-580

Boulton A, Cobb T (2017) Corpus use in language learning: A meta-analysis. Lang Learn 67(2):348–393. https://doi.org/10.1111/lang.12224

Boulton A, Vyatkina N (2021) Thirty years of data-driven learning: Taking stock and charting new directions over time. Lang Learn Technol 25(3):66–89

Brown C, Spiro J, Quinton S (2020) The role of research ethics committees: Friend or foe in educational research? An exploratory study. Br J Educ Res 46(4):747–769. https://doi.org/10.1002/berj.3654

Çelik S (2011) Developing collocational competence through web based concordance activities. Novitas-ROYAL (Res Youth Lang) 5(2):273–286

Chambers A (2007) Popularising corpus consultation by language learners and teachers. In: Hidalgo E, Quereda L, Santana J (eds) Corpora in the Foreign Language Classroom. Brill, Leiden pp 3-16

Chambers A (2019) Towards the corpus revolution? Bridging the research–practice gap. Lang Teach 52(4):460–475. https://doi.org/10.1017/S0261444819000089

Chan TP, Liou HC (2005) Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. Comput Assist Lang Learn 18:231–251. https://doi.org/10.1080/09588220500185769

Chang J, Sun Y (2022) The role of data-driven learning in EFL vocabulary acquisition: A meta-analysis. J Engl Acad Purp 51:101003

Chen W (2019) Profiling collocations in EFL writing of Chinese tertiary learners. RELC J 50(1):53–70. https://doi.org/10.1177/0033688217716507
Chen HJH, Lai SL, Lee KY, Yang CTY (2021) Developing and evaluating an academic collocations and phrases search engine for academic writers. Comput Assist Lang Learn 2021:1–28. https://doi.org/10.1080/09588221.2021.1937229
Chen M, Flowerdew J (2018) A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. Int J Corpus Linguist 23(3):335–369. https://doi.org/10.1075/ijcl.16130.che
Cobb T (1997) Is there any measurable learning from hands-on concordancing? System 25(3):301–315. https://doi.org/10.1016/S0346-251X(97)00024-9
Cohen J (1988) Statistical power analysis for the behavioral sciences. Routledge, London
Coxhead A (1998) An academic word list (English Language Institute Occasional Publication No. 18). Victoria University of Wellington, New Zealand
Coxhead A (2008) Phraseology and English for academic purposes: challenges and opportunities. In: Meunier F, Granger S (eds) Phraseology in Foreign Language Learning and Teaching. John Benjamins, Amsterdam, Netherlands, pp 149–161. https://doi.org/10.1075/z.138.12cox
Crosthwaite P, Baisa V (2023) Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Corpus Linguist Res 3(3):100066. https://doi.org/10.1016/j.acorp.2023.100066
Crosthwaite P, Steeples B (2022) Data-driven learning with younger learners: exploring corpus-assisted development of the passive voice for science writing with female secondary school students. Comput Assist Lang Learn 37(3):1166–1197. https://doi.org/10.1080/09588221.2022.2068615
Crossley S, Salsbury T, McNamara D (2015) Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. Appl Linguist 36(5):570–590. https://doi.org/10.1093/applin/amt056
Daskalovska N (2015) Corpus-based versus traditional learning of collocations. Comput Assist Lang Learn 28(2):130–144. https://doi.org/10.1080/09588221.2013.803982
Durrant P (2019) Formulaic language for English for academic purposes. In: Siyanova-Chanturia A, Pellicer-Sánchez A (eds) Understanding formulaic language: A second language acquisition perspective. Routledge, London, pp 174–191
Fang X, Yang H, Zhu Z (2008) Creating a unified scale of language ability in China. Mod Foreign Lang 31(4):380–387
Flowerdew L (2015) Corpus-based research and pedagogy in EAP: From lexis to genre. Lang Teach 48(1):99–116. https://doi.org/10.1017/S0261444813000037
Frankenberg-Garcia A (2018) Investigating the collocations available to EAP writers. J Engl Acad Purp 35:93–104. https://doi.org/10.1016/j.jeap.2018.07.003
Gablasova D, Brezina V, McEnery T (2017) Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. Lang Learn 67:155–179. https://doi.org/10.1111/lang.12225
Geluso J, Yamaguchi A (2014) Discovering formulaic language through data-driven learning: Student attitudes and efficacy. ReCALL 26(2):225–242. https://doi.org/10.1017/S0958344014000044
Granger S (2019) Formulaic sequences in learner corpora. In: Siyanova-Chanturia A, Pellicer-Sánchez A (eds) Understanding formulaic language: A second language acquisition perspective. Routledge, London, pp 228–238
Gyllstad H (2009) Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In: Barfield A, Gyllstad H (eds) Researching collocations in another language. Palgrave Macmillan, New York, pp 153–170. https://link.springer.com/chapter/10.1057/9780230245327_12
Gyllstad H, Schmitt N (2019) Testing formulaic language. In: Siyanova-Chanturia A, Pellicer-Sánchez A (eds) Understanding formulaic language: A second language acquisition perspective. Routledge, London, pp 174–191
Halliday MAK (1993) Writing science literacy and discursive power. Falmer Press, London
He M, Ang LH (2023) Profiling a microeconomics noun collocation list: A corpus-based approach. South Afr Linguist Appl Lang Stud 41(2):191–209. https://doi.org/10.2989/16073614.2022.2117708
Johns TF (1991) Should you be persuaded: Two examples of data-driven learning materials. Engl Lang Res J 4:1–16
Kwarikunda D, Schiefele U, Muwonge CM, Ssenyonga J (2022) Profiles of learners based on their cognitive and metacognitive learning strategy use: occurrence and relations with gender, intrinsic motivation, and perceived autonomy support. Humanit Soc Sci Commun 9:337
Larson-Hall J (2015) A guide to doing statistics in second language research using SPSS and R. Routledge, London
Laufer B, Waldman T (2011) Verb-noun collocations in second language writing: A corpus analysis of learners' English. Lang Learn 61(2):647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x
Lee JS, Lee K (2020) Affective factors, virtual intercultural experiences, and L2 willingness to communicate in in-class, out-of-class, and digital settings. Lang Teach Res 24(6):813–833. https://doi.org/10.1177/1362168819831408
Lee H, Warschauer M, Lee JH (2019) The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. Appl Linguist 40(5):721–753. https://doi.org/10.1093/applin/amy012
Lewis M (2000) Teaching collocations: further development in the lexical approach. Language Teaching Publications, Hove
Li XL (2023) Promoting accuracy of collocation use in L2 writing: the role of data-driven learning in indirect corrective feedback. Comput Assist Lang Learn 1–25. https://doi.org/10.1080/09588221.2023.2292554
Liu D, Jiang P (2009) Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. Mod Lang J 93(1):61–78
Liu Y, Lu X (2020) Chinese EFL learners' misconceptions of noun countability and article use. System 90:102222. https://doi.org/10.1016/j.system.2020.102222
Liu S, Ma Z (2011) An empirical study on concordance-based English collocation teaching. Int J Educ Manag Eng 4:46–52. https://doi.org/10.5815/ijeme.2011.04.08
Lu Y (2017) A corpus study of collocation in Chinese learner English. Routledge, London
Meunier F, Granger S (2008) Phraseology in foreign language learning and teaching. John Benjamins, Amsterdam
Nation ISP (2001) Learning vocabulary in another language. Cambridge University Press, Cambridge
Nesselhauf N (2004) Collocations in a learner corpus. John Benjamins, Amsterdam
Nizonkiza D (2014) The relationship between productive knowledge of collocations and academic literacy among tertiary level learners. J Lang Teach 48(1):149–171. https://doi.org/10.4314/jlt.v48i1.8
Nizonkiza D, Van de Poel K (2019) Mind the gap: Towards determining which collocations to teach. SPiL 56:13–31. https://doi.org/10.5842/56-0-775
Packard JL (2000) The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, Cambridge
Park M, Son JB (2020) Pre-service EFL teachers' readiness in computer-assisted language learning and teaching. Asia Pac J Educ 42(2):320–334. https://doi.org/10.1080/02188791.2020.1815649
Parkinson J (2015) Noun-noun collocations in learner writing. J Engl Acad Purp 20:103–113. https://doi.org/10.1016/j.jeap.2015.08.003
Pellicer-Sánchez A, Boers F (2019) Pedagogical approaches to the teaching and learning of formulaic language. In: Siyanova-Chanturia A, Pellicer-Sánchez A (eds) Understanding formulaic language: A second language acquisition perspective. Routledge, London, pp 153–173
Pérez-Paredes P, Sánchez-Tornel M (2014) Adverb use and language proficiency in young learners' writing. Int J Corpus Ling 19(2):178–200. https://doi.org/10.1075/ijcl.19.2.02per
Peters E (2009) Learning collocations through attention-drawing techniques: a qualitative and quantitative analysis. In: Barfield A, Gyllstad H (eds) Researching collocations in another language. Palgrave Macmillan, New York, pp 194–207
Rezaee AA, Marefat H, Saeedakhtar A (2015) Symmetrical and asymmetrical scaffolding of L2 collocations in the context of concordancing. Comput Assist Lang Learn 28(6):532–549. https://doi.org/10.1080/09588221.2014.889712
Schmidt R (2001) Attention. In: Robinson P (ed) Cognition and second language instruction. Cambridge University Press, Cambridge, pp 3–32
Simpson-Vlach R, Ellis NC (2010) An academic formulas list: New methods in phraseology research. Appl Linguist 31:487–512. https://doi.org/10.1093/applin/amp058
Sinclair J (1987) Collocation: a progress report. In: Steele R, Thomas T (eds) Language topics: essays in honor of Michael Halliday II. John Benjamins, Amsterdam, pp 319–331
Sun Y, Wang L (2003) Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. Comput Assist Lang Learn 16(1):83–94. https://doi.org/10.1076/call.16.1.83.15528
Tondeur J, Scherer R, Baran E, Siddiq F, Valtonen T, Sointu E (2019) Teacher educators as gatekeepers: Preparing the next generation of teachers for technology integration in education. Br J Educ Technol 50(3):1189–1209
Vyatkina N (2016a) Data-driven learning for beginners: The case of German verb-preposition collocations. ReCALL 28(2):207–226. https://doi.org/10.1017/S0958344015000269
Vyatkina N (2016b) Data-driven learning of collocations: Learner performance, proficiency, and perceptions. Lang Learn Technol 20(3):159–179
Willis J (2011) Concordances in the classroom without a computer: Assembling and exploiting concordances of common words. In: Tomlinson B (ed) Materials development in language teaching. Cambridge University Press, Cambridge, pp 51–78
Wray A (2002) Formulaic language and the lexicon. Cambridge University Press, Cambridge
Yeh Y, Liou HC, Li YH (2007) Online synonym materials and concordancing for EFL college writing. Comput Assist Lang Learn 20:131–152. https://doi.org/10.1080/09588220701331451
Yoon H, Hirvela A (2004) ESL student attitudes towards corpus use in L2 writing. J Second Lang Writ 13(4):257–283. https://doi.org/10.1016/j.jslw.2004.06.002

Zare J, Delavar A (2022) Enhancing English learning materials with data-driven learning: a mixed-methods study of task motivation. J Multiling Multicult Dev 1-17. https://doi.org/10.1080/01434632.2022.2134881

Zheng H, Hu B, Xu J (2022) The development of formulaic knowledge in super-advanced Chinese language learners: Evidence from processing accuracy, speed, and strategies. Front Psychol 13:796784. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.796784/full

Acknowledgements

We would like to extend our sincere gratitude to the reviewers and the editor for their invaluable comments and suggestions. This study was funded by 2022 Fujian Provincial Education Research Project for Young and Middle-aged Teachers (Grant number: JSZW220026), and the Foundation Project of China Ministry of Education for Humanities and Social Sciences (Grant number: 21YJC760074; 24YJC740007).

Author contributions

MYH: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing-Original Draft, Writing- Review and Editing, Visualization. QX: Methodology, Software, Formal Analysis, Investigation, Data curation, Writing- Review and Editing, Visualisation, Funding Acquisition.

Competing interests

The authors declare no competing interests.

Consent to publish

All student participants agreed that anonymized findings from the study could be published in academic journals, ensuring that no individual participant could be identified. As the study involved non-interventional research through language learning sessions, collocation tests, and questionnaires, all student participants were fully informed that their anonymity would be maintained throughout the study. The research team clearly communicated the purpose of the study, and how the data would be utilized, and assured that there were no foreseeable risks associated with participation. Student participants were also informed of their right to withdraw from the study at any point without any consequences.

Ethical approval

The research received approval from the Scientific Research Ethics Committee at the School of Foreign Languages, Hubei University of Economics. After a thorough review by the Scientific Research Ethics Committee, it was determined that the research design and plan are scientifically sound, fair, and impartial, and do not pose any harm or risk to participants. Participant recruitment adheres to the principles of voluntary participation and informed consent, ensuring the protection of participants' rights and privacy. No conflicts of interest or violations of ethical or legal standards were identified in this project. The researchers confirm that all procedures were conducted in compliance with relevant guidelines and regulations, including the Declaration of Helsinki. Name of the approval body: Scientific Research Ethics Committee, School of Foreign Languages, Hubei University of Economics. Approval Number: 2023-02. Date of Approval: 13 May 2023. Scope of Approval: The approval encompasses all aspects of the research, including participant recruitment, data collection, analysis, and reporting. It applies to all interactions with participants and ensures that ethical principles are upheld throughout the study's duration, ensuring the protection of participant rights, privacy, and informed consent.
Name of the approval body: Scientific Research Ethics Committee, School of Foreign Languages, Hubei University of Economics.
Approval Number: 2023-02.
Date of Approval: 13 May 2023.
Scope of Approval: The approval encompasses all aspects of the research, including participant recruitment, data collection, analysis, and reporting. It applies to all interactions with participants and ensures that ethical principles are upheld throughout the study's duration, ensuring the protection of participant rights, privacy, and informed consent.

Informed consent

Informed consent was obtained from all student participants in written form prior to their participation in the study. Consent was collected on 16 May 2023 by the research team of the School of Foreign Languages, Hubei University of Economics. The student participants are from a senior high school in Hubei Province, China.
Scope of Consent: The written consent covered multiple aspects:
Participation: All student participants agreed to participate in the language learning sessions, collocation tests, and questionnaire survey, acknowledging that their involvement was entirely voluntary.
Data Use: All student participants were informed that the data collected would be used solely for academic research purposes, focusing on language learning effectiveness and that the information would be analysed collectively.

Additional information

Correspondence and requests for materials should be addressed to Qin Xie.