Asian Journal of English Language Teaching, *34,* 5-22

© 2025 The Chinese University of Hong Kong

Is Data-Driven Learning with AI-Generated Example Sentences Effective for Elementary-Level Learners? A Case Study of the Synonyms "collect" and "gather"

Ryosuke NAKAHARA *ryosuke_nakahara@yamate-gakuin.ed.jp
Yamate Gakuin Junior and Senior High School, Japan

Despite its potential, data-driven learning (DDL) remains underutilized in pre-tertiary education, partly due to the lack of tools that provide ample and level-appropriate example sentences. This classroom-based case study investigates whether generative artificial intelligence (GenAI) can generate suitable example sentences to help elementary-level learners distinguish between the synonyms "gather" and "collect", and whether the learners can discern their differences through hands-off DDL with the AI-generated example sentences using a pre- and post-test design. Analyses based on the English Vocabulary Profile (EVP) and Coh-Metrix revealed that GenAI can generate sentences that align with the CEFR levels specified in the prompt (i.e., A1–A2). Furthermore, a Bayesian Wilcoxon signed-rank test yielded a Bayes factor (BF_{10}) of 31.564 and an effect size (δ) of –0.68, indicating that the hypothesis that post-test scores are higher than pre-test scores is approximately 32 times more plausible, representing a medium effect size. Although this is a small-scale, classroom-based case study, the findings suggest that GenAI can function as a user-friendly concordancer, offering accessible tools for younger learners and potentially addressing some of the key barriers to implementing DDL in schools.

Introduction

More than three decades have passed since Johns (1991) proposed a language learning method called data-driven learning (DDL), in which learners themselves discover linguistic patterns within linguistic data, such as corpora. While many researchers view DDL as a promising learning approach, it has not yet become mainstream in educational settings, particularly in primary and secondary schools. Indeed, a meta-analysis by Boulton and Cobb (2017) revealed that most DDL research has been conducted with university students, and only a limited number of studies have focused on school-aged learners. As Pérez-Paredes (2020) argues, one reason for the limited implementation of DDL in schools is the scarcity of tools specifically designed for beginner or elementary-level learners, along with the unfamiliarity of both teachers and learners with corpora. While the number of pedagogical DDL tools has been increasing recently (e.g., Nishigaki & Akasegawa, 2023), these tools still offer a smaller quantity of example sentences compared to authentic corpora.

Over the last 30 years, technology has advanced significantly. Among these advancements, the release of a generative artificial intelligence (GenAI) tool called ChatGPT by OpenAI in 2022 astonished people worldwide. DDL researchers were no exception, as ChatGPT appeared to challenge the value of their years of hard work (Lin, 2023). However, as understanding of GenAI deepens, some DDL researchers (e.g., Crosthwaite & Baisa, 2023; Lin, 2023; Mizumoto, 2023) have begun to advocate for the coexistence of DDL and GenAI, adopting the mindset of "If you can't beat them, join them" (Lin, 2023, p.1). One way to achieve this coexistence is to utilize GenAI as a concordancer (Flowerdew, 2024; Lin, 2023). Unlike traditional corpora, GenAI raises several concerns when used in DDL, such as hallucinated output, opaque training data, and limited replicability (Crosthwaite & Baisa, 2023). However, its inherently fluid and generative nature may enable it to produce level-appropriate example sentences on demand. These shortcomings, therefore, do not negate the pedagogical potential of GenAI; rather, they underscore the importance of its principled and carefully managed integration into DDL practices. In particular, given the continued lack of accessible DDL tools for beginner and elementary-level learners, GenAI has the potential to function as a user-friendly concordancer capable of generating a large number of example sentences of moderate difficulty. Assuming this is the case, DDL with AI-generated example sentences may be able to address the factors limiting its practice in primary and secondary schools, such as the lack of example sentences in DDL tools and the unfamiliarity of corpora among teachers and learners. However, to date, it remains unclear whether GenAI can generate example sentences suitable for beginner or elementary-level learners to use in DDL activities. Furthermore, it is essential to verify whether DDL with AI-generated example sentences is effective. To explore these issues, the present study conducts a small-scale, classroom-based case study with 31 Japanese junior-high students. By examining both the linguistic suitability of AI-generated sentences and subsequent learning outcomes, I aim to provide preliminary—but practically relevant—evidence on the feasibility of GenAI-mediated DDL in secondary education.

Literature Review

Data-Driven Learning

The widespread use of computers has enabled linguists to process vast amounts of linguistic data, leading to numerous studies using corpora. In this context, Johns (1991) proposed a new approach to corpus use, where learners themselves, not linguists, engage directly with linguistic data to identify linguistic patterns as if they were linguists. This innovative method, known as data-driven learning (DDL), has since attracted significant attention from researchers and has been shown to be effective in various areas. For example, it has shown to be effective for error correction in writing (e.g., Yoon & Jo, 2014), the acquisition of verb-preposition collocations (e.g., Saeedakhtar, Bagerin, & Abdi, 2020), and the finding of the relationship between verb semantics and syntactic patterns (e.g., Nakahara & Sugiura, 2025). Moreover, some meta-analyses have further demonstrated the effectiveness of DDL (e.g., Boulton & Cobb, 2017; Boulton & Vyatkina, 2021; Mizumoto & Chujo, 2015). For example, Boulton and Cobb (2017) synthesized findings from 64 studies and reported medium to large learning effects.

However, some studies have highlighted the challenges faced by learners, particularly those at beginner and elementary-levels, when directly using corpora (e.g., Nishigaki &

Akasegawa, 2023). This has led to the development of various types of DDL incorporating pedagogical scaffolding. For instance, some studies (e.g., Saeedakhtar et al., 2020; Yoon & Jo, 2014) have suggested hands-off DDL, where learners use handouts containing concordance lines selected in advance by teachers. Additionally, guided DDL approaches (e.g., Smart, 2014), in which teachers facilitate DDL activities, provide structured support for effective engagement with concordance lines.

Other studies have focused on developing DDL tools based on pedagogical corpora. For instance, Nishigaki and Akasegawa (2023) developed *hDDL*, a DDL tool specifically designed for secondary students, emphasizing the importance of building "a bridge to authentic corpus use" (p. 135). One of the key features of *hDDL* is that it exposes learners to complete English sentences that are level-appropriate. Additionally, it functions as a parallel corpus, offering Japanese translations of the sentences simultaneously to support learners' understanding. Although there are pros (e.g., ease of use for those at beginner and elementary-levels) and cons (e.g., lack of authenticity) to the practice of pedagogically aided DDL, the common underlying principle across all types of DDL is that learners generalize linguistic patterns through exposure to numerous example sentences (Nakahara & Sugiura, 2025). In other words, DDL is a learning method that embodies the principles of a usage-based model (e.g., Tomasello 2003), which sees domain-general cognitive abilities (e.g., pattern-finding skills) as the key to language acquisition. Since this view requires a vast amount of input, DDL is particularly well suited to learners in EFL environments where sufficient input is often unavailable (O'Keeffe, 2023). In such an environment, DDL compensates for the lack of input and help learners to "discover language facts and to modify or replace previous knowledge with new knowledge" (Nishigaki & Akasegawa, 2023, p.132).

GenAI and DDL

GenAI tools such as OpenAI's ChatGPT have had a significant impact across various areas, including English education. Some GenAI platforms, including ChatGPT, employ Large Language Models (LLMs), which are described as "employing deep learning techniques on extensive web-collected text data (i.e., corpus)" (Mizumoto, 2023, p. 2). This indicates that the development of LLMs is grounded in vast amounts of linguistic data.

Given the accessibility of user-friendly AI tools like ChatGPT, it would not be surprising if some were to claim that DDL is no longer necessary. However, a number of researchers now advocate for the coexistence of GenAI and DDL (e.g., Crosthwaite & Baisa, 2023; Flowerdew, 2024; Lin, 2023; Mizumoto, 2023). One promising approach to integrating GenAI with DDL is to use GenAI as a concordancer. Lin (2023) highlights the advantages of employing GenAI as an accessible tool for concordance-based learning. While GenAI is capable of directly answering linguistic questions—for example, by clearly explaining the difference between synonyms—Lin (2023) argues that active and deeper engagement with language data is more likely to lead to successful learning.

As Johns (1991) observed, one of the questions learners most frequently ask is, "What is the difference between …?" In a recent class, I was asked exactly this question with respect to the synonyms *collect* and *gather*. DDL is well suited to helping students explore such lexical distinctions. However, pedagogical DDL tools (e.g., *hDDL*) sometimes contain too few example sentences at least as of now. When I searched *hDDL* in January 2025, only two

examples were available for each word. Given the labor required to compile level-controlled example sentences, it is unsurprising that certain items are under-represented, and *hDDL*'s real strength seems to lie in highlighting more general patterns rather than in providing exhaustive coverage of every lexical item. By contrast, GenAI can produce numerous, level-appropriate example sentences in a single step, potentially filling this gap.

However, GenAI's capabilities are not without limitations, and its effectiveness varies across tasks. For instance, Mizumoto and Teng (2025) report that, in a classification task requiring nuanced contextual judgment, even the best-performing LLM achieved only moderate agreement with human coders ($\kappa = 0.68$). This cautionary evidence, though drawn from a different task, underscores the need to critically evaluate GenAI across specific task types, including those relevant to DDL. That said, whether GenAI can generate level-appropriate example sentences, and whether DDL using such AI-generated examples is effective, remains insufficiently explored. Uchida (2025) examined the first issue by analyzing ChatGPT-4o output with the CEFR-J based vocabulary level analyzer (CVLA; https://cvla.langedu.jp/). He found that only 5 % of the AI-generated passages matched the CEFR level specified in the prompt. Uchida also noted a systematic bias: when prompted for low-level text (A1–A2), ChatGPT tended to generate texts that were even easier than the target level. In a DDL context, however, this downward bias is not necessarily detrimental. Sentences that are slightly easier than the stated level reduce cognitive load and may make it easier for learners to notice linguistic patterns. What matters is that most of the words in each sentence are already familiar to the learners. For this reason, it seems more appropriate to verify GenAI output with the English Vocabulary Profile (EVP). EVP is a CEFR-based word list that "[focuses] on what learners actually know rather than prescribing what they should know" (Capel, 2012, p. 1). In other words, EVP does not merely label words by level; it identifies which specific words learners at a given level are empirically known to know. Moreover, as Nishigaki and Akasegawa (2023) argue, elementary-level learners are hindered not only by advanced vocabulary but also by long sentences and complex grammar. Indeed, Crossley and McNamara (2014) demonstrated that simplified texts enhance comprehension among L2 learners. Thus, these aspects, too, must be verified in GenAI output. On this basis, the present study poses following two research questions:

RQ1. Can GenAI generate example sentences that are suitable for elementary-level learners?
RQ2. Can learners discern the differences between synonyms through hands-off DDL with the AI-generated example sentences?

Method

Participants

Thirty-one ninth-grade Japanese EFL students participated in this study. The participants had a mean age of 14.84 years and had been studying English for an average of 4.81 years. As the participants were minors, a written explanation was provided to both the students and their guardians. This document detailed how their data would be handled and clarified that the results of this study would not affect their academic performance. It also disclosed that the study utilized example sentences generated by GenAI.

All participants had taken the EIKEN Test in Practical English Proficiency (EIKEN) as part of their regular coursework, and their EIKEN grades were therefore used as a measure of English proficiency. EIKEN is a widely recognized English language proficiency test in Japan, consisting of seven levels (Grades 5, 4, 3, Pre-2, 2, Pre-1, and 1). These grades correspond to CEFR, ranging from A1 for Grade 3 to C1 for Grade 1. The number of participants at each EIKEN grade in this study is shown in Table 1.

Table 1. Proficiency of Learners

Grade	CEFR Level	Number
4	-	1
3	A1	4
Pre-2	A1-A2	18
2	A2-B1	7
Pre-1	B1-B2	1

Note. EIKEN Foundation does not provide an official CEFR equivalence for Grade 4.

Based on EIKEN grades, although only one student with Grade Pre-1 is at the CEFR B1 to B2 level, most of the participants were at the CEFR A1 to A2-B1 level.

Materials

Target Items and Handout

The synonyms *gather* and *collect* were selected as the target items based on a student's question raised in a previous class. As Söğüt (2023) states, near-synonyms are often confusing for L2 learners, and indeed, several students seemed to struggle to distinguish between these two words. Therefore, these items were deemed suitable targets for the DDL activity. Two handouts were prepared for the task. One is an *example sentence handout* containing printed example sentences. The example sentences were created using the ChatGPT-o1 model. Uchida (2025) employed the ChatGPT-4o model but pointed out potential issues with reproducibility due to the model's continual updates. In response, the present study used o1 model, the most recently released publicly available model at the time of data collection. As o1 model has been reported to offer improved reasoning capabilities, it was considered an appropriate choice for exploring the applicability of newer models in generating pedagogically suitable example sentences. The prompt and its English translation were as follows:

Original Prompt

Data-Driven Learning で中学３年生が、類義語である gather と collect の違いを学ぶための教材を作成したい。CEFR で A1~A2 レベルの適切な語彙難易度の例文を対訳付きで出力してください。 例文はそれぞれ１５文ずつ、KWIC 形式で出力してください。

English Translation

To create teaching materials for ninth graders to learn the differences between the synonyms *gather* and *collect* through Data-Driven Learning, please provide 15

example sentences for each word in KWIC format at the CEFR A1-A2 vocabulary level, along with Japanese translations.

Example sentences were generated in a single pass; no prompt adjustment or iterative regeneration was performed. The AI-generated example sentences were transcribed into a handout with a well-organized format to make them easy for learners to reference, but no changes were made to the content of the sentences. Refer to Appendix 1 for the handout.

The other handout is a *DDL guiding handout* designed to help learners identify the differences between *gather* and *collect*. For example, the handout included instructions such as: "Focus on what comes after the verb and use a rectangle to highlight the object if there is one," or "Let's extract the highlighted objects and think about what kinds of words are used as objects."

Tests

Paper-based pre- and post-tests were also prepared, each consisting of 20 fill-in-the-blank questions requiring students to choose either *gather* or *collect* to complete sentences corresponding to the Japanese translation. An example question from the test is as follows:

彼女は地元のチャリティーイベントのために寄付金を集めることを望んでいます。
She wants to _____ donations for the local charity event.

While the pre- and post-tests contained the same questions, the order of the questions was randomized. Refer to Appendix 2 for the complete set of tests.

Procedure

After completing the pre-test, which took approximately five minutes, the participants began the DDL activity. In the first part of the activity, they spent about 20 minutes independently analyzing the differences between *gather* and *collect* using the *example sentence handout* and the *DDL guiding handout*. In the second part, they shared their findings and opinions in groups, and the activity lasted no longer than 15 minutes. Finally, the participants took the post-test, which also took approximately five minutes.

It should be noted that this study is best described as a case study that arose from an actual classroom incident. When a student asked about the difference between *gather* and *collect*, I planned a DDL activity with the pedagogical tool (i.e., *hDDL*). Because only a few examples for these particular verbs were available in the platform, additional sentences were needed, and GenAI was used to supplement the activity. Accordingly, no control group comparing GenAI-mediated DDL with traditional DDL was included.

Analysis

First, lexical difficulty was assessed using the EVP via Text Inspector (https://textinspector.com/), a web-based text analysis tool that provides EVP-based

vocabulary tagging. After tagging the sentences, the proportion of words at each vocabulary level was calculated from the EVP tagging results using R version 4.3.3. The calculation was performed using word types, rather than tokens, to ensure that each unique word was only counted once within the dataset.

In addition, the syntactic difficulty of the AI-generated sentences was assessed automatically using Coh-Metrix. Coh-Metrix is a computational tool that provides multilevel analyses of texts, including measures of word concreteness and syntactic simplicity (Graesser, McNamara & Kulikowich, 2011). Of the many indices generated by Coh-Metrix, five measures shown in Table 2 were selected for analysis. Because the dataset consisted of 30 isolated example sentences rather than continuous paragraphs, paragraph-level measures (e.g., cohesion indices) were not included in the analysis.

Table 2. The Adopted Coh-Metrix Indices

Indices	Description
DESSL	Sentence length: mean number of words
DESSLd	Sentence length: standard deviation
PCSYNz	Text easability: syntactic simplicity (z-score)
PCSYNp	Text easability: syntactic simplicity (percentile)
SYNNP	Mean number of modifiers per noun phrase

Note. PCSYNz and *PCSYNp* are based on a principal component analysis of multiple syntactic features (e.g., sentence length and the use of familiar, simple syntactic structures) and indicate the degree to which sentences are easy to process.

Next, to evaluate the effectiveness of DDL using AI-generated example sentences, a Wilcoxon signed-rank test was conducted to compare pre-test and post-test scores. To avoid a ceiling effect and ensure the validity of the analysis, participants who scored 80% (i.e., 16 out of 20) or higher on the pre-test were excluded, as they were deemed to already possess sufficient knowledge of the target items. Consequently, data from 24 participants were included in the analysis. Given the limited sample size, a Bayesian approach was adopted in this study. As Miočević, Levy, and Schoot (2020) state, “Bayesian statistics are not based on large samples (i.e., the central limit theorem) and hence may produce reasonable results even with small to moderate sample size” (p. 3). Whereas a traditional frequentist paired-samples t-test simply makes a binary judgment regarding the significance of the difference between two tests based on a *p*-value, which is highly sensitive to sample size, a Bayesian paired-samples model (i) quantifies the strength of evidence for a learning effect via a Bayes factor and (ii) provides a full posterior distribution along with a 95% credible interval for the effect size, offering richer information particularly valuable with a modest sample size of 24. The analysis was conducted using JASP 0.19.3.

Finally, since a decrease in post-test scores was observed in some learners, it was hypothesized that their prior knowledge of the target items might have influenced the changes in their scores. To investigate this possibility, pre- and post-test scores were plotted on a scatterplot to visually examine the relationship between initial performance and subsequent score changes.

Results

Lexical Difficulties

Figure 1 displays the EVP-based CEFR-level proportions for example sentences generated by GenAI. Table 3 presents the corresponding vocabulary-type counts and their proportions.

As shown in Figure 1 and Table 3, the example sentences mainly consist of CEFR A1 or A2 level words. However, some B1, B2, and unlisted vocabulary items were also identified. The vocabulary items that appeared in the example sentences and are either beyond the A2 level or unlisted are listed in Table 4.

Figure 1. CEFR Levels of Vocabulary in Example Sentences Based on EVP

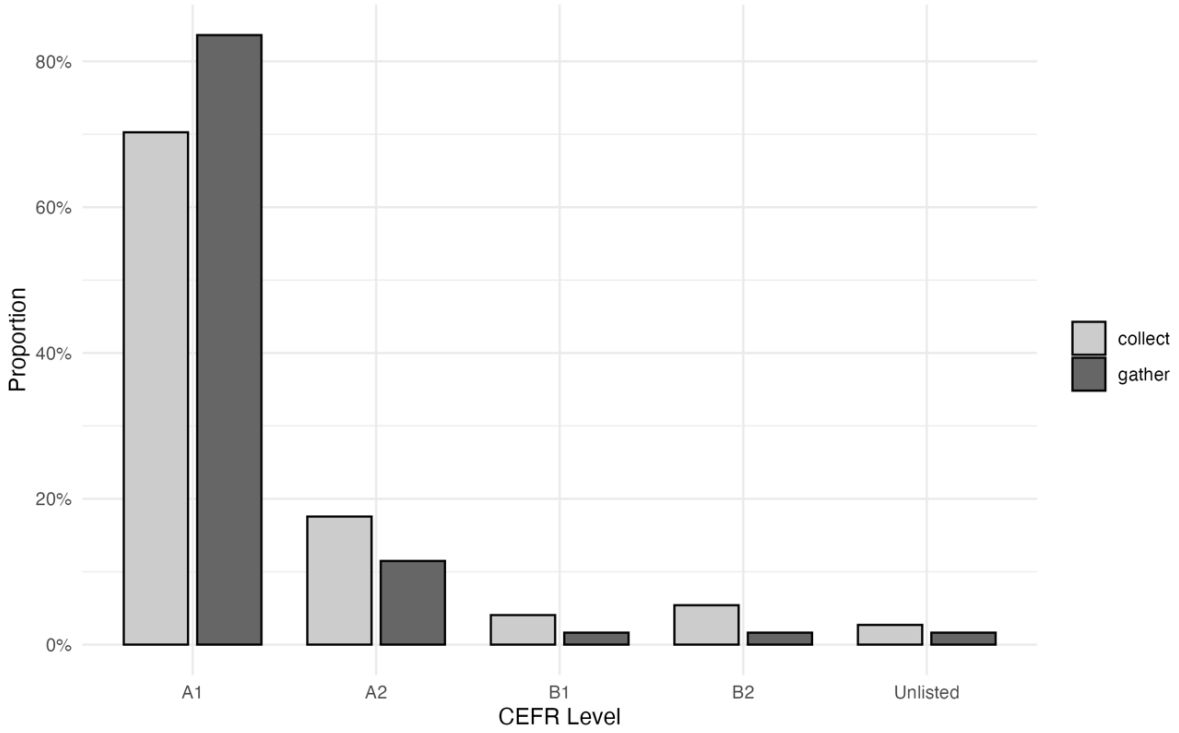


Table 3. Vocabulary Type Counts and Proportions by CEFR Level Based on EVP

	CEFR	Count	Ratio
gather	A1	51	0.84
	A2	7	0.11
	B1	1	0.02
	B2	1	0.02
	Unlisted	1	0.02
collect	A1	78	0.78
	A2	13	0.13
	B1	3	0.03
	B2	4	0.04
	Unlisted	2	0.02

Note. Ratios may not sum to 1.00 due to rounding.

Table 4. Vocabulary Items Beyond the A2 Level or Unlisted

	B1	B2	Unlisted
gather	thought	shell	campfire
collect	bucket	data	rainwater
	coin	donate	sticker
	event	shell	
		trading	

Syntactic Difficulties

According to the syntactic difficulty analysis conducted using Coh-Metrix, the mean number of words per sentence (i.e., DESSL) in the example sentences generated by GenAI was 7.6, with a standard deviation (i.e., DESSLd) of 1.04. The z-score for syntactic simplicity (i.e., PCSYNz) was 2.28, corresponding to the 98.87th percentile (i.e., PCSYNp), indicating that the syntactic structure of the example sentences was simpler than that of over 98% of texts in the TASA (Touchstone Applied Science Associates) corpus, which serves as the norming base for the Coh-Metrix easability scores. The mean number of modifiers per noun phrase (i.e., SYNNP) was 0.54. A summary of these results is presented in Table 5.

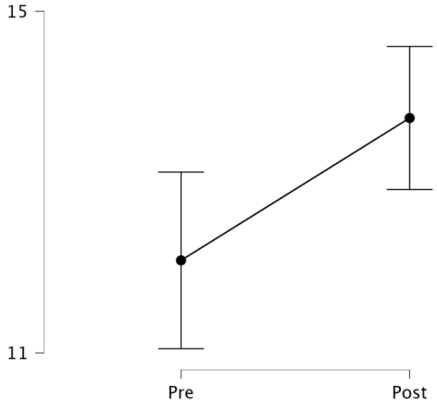
Table 5. Summary of Syntactic Difficulties

Indices	Score
DESSL	7.6
DESSLd	1.04
PCSYNz	2.28
PCSYNp	98.87
SYNNP	0.54

The Effectiveness of DDL Using AI-generated Example Sentences

The change in mean score is visually depicted in Figure 2 and the descriptive statistics for the pre- and post-tests are presented in Table 6.

Figure 2. Change in Mean Pre- and Post-Test Scores



Note. Error bars represent 95 % credible intervals.

Table 6. Descriptive Statistics for Pre- and Post-Tests

	Pre	Post
Valid	24	24
Median	12.50	14.00
Mean	12.08	13.75
Std. Deviation	2.45	1.98
Skewness	-0.45	0.20
Std. Error of Skewness	0.47	0.47
Kurtosis	-1.19	-0.52
Std. Error of Kurtosis	0.99	0.92
Shapiro-Wilk	0.89	0.94
p-value of Shapiro-Wilk	0.01	0.16
Minimum	8	11
Maximum	15	18

Because the pre-test scores violated the normality assumption (Shapiro–Wilk $W = .89$, $p = .01$), a Bayesian Wilcoxon signed-rank test with the default Cauchy (0,0.707) prior was applied. MCMC convergence was assessed via R-hat, which equaled 1.004. As values ≤ 1.1 indicate convergence, the sampling chains have effectively converged and the resulting posterior distribution can be considered reliable. A Bayes factor (BF_{10}) was 31.564, which indicates that the hypothesis "post-test scores are higher than pre-test scores" is about 32 times more plausible than the null hypothesis of no difference. According to Goss-Sampson (2020), BF_{10} values between 10 and 30 constitute "very strong evidence" (p. 40). The posterior median effect size was $\delta = -0.68$ (95 % credible interval [–1.14, –0.22]), indicating a medium-sized improvement in post-test scores relative to pre-test.

In short, the Bayesian analysis provides very strong evidence for a medium-sized improvement in post-test scores following DDL with AI-generated example sentences. The results are summarized in Figure 3 and Table 7.

Figure 3. Prior and Posterior Distribution of the Effect Size

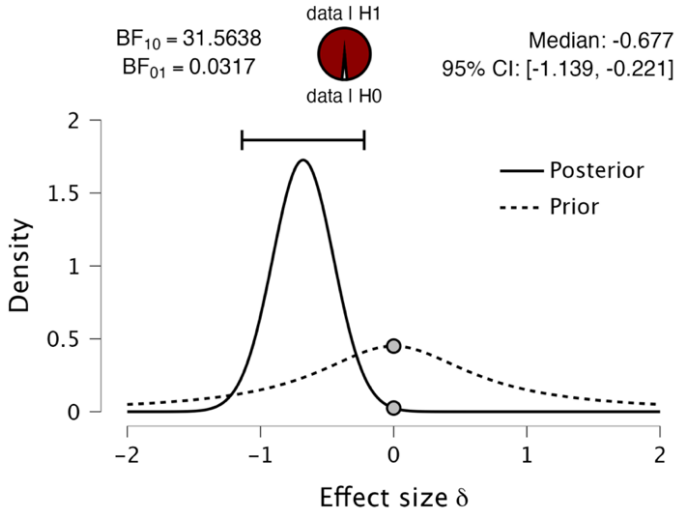


Table 7. Bayesian Wilcoxon Signed-Rank Test

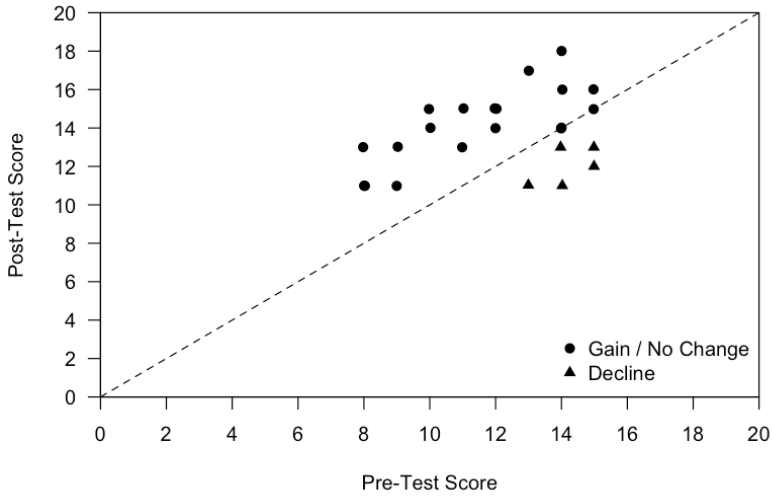
Measure 1		Measure 2	BF_{10}	W	R-hat
Pre	-	Post	31.564	35.500	1.004

Note. Result based on data augmentation algorithm with 5 chains of 1000 iterations.

Relationship between Pre- and Post-test scores

Finally, to visualize how individual scores changed following the DDL activity, participants' pre- and post-test scores are plotted in Figure 4. The figure shows that most participants improved, as indicated by data points located above the diagonal line representing identical pre- and post-test scores. Notably, however, several participants with high initial scores (13 or above) exhibited a decrease in their post-test scores.

Figure 4. Scatter Plot of Individual Pre- and Post-Test Scores



Discussion

Based on the EVP analysis, about 90 % of the words in the example sentences for both *gather* and *collect* fall within the CEFR A1–A2 bands. While some words exceed the specified levels indicated in the prompt (i.e., A1–A2), they do not appear overly challenging for learners. For instance, words such as *bucket*, *coin*, *data*, and *sticker* are likely familiar to the participants, as they are loanwords commonly used in everyday Japanese. Thus, from a perspective of the lexical difficulties, it can be concluded that GenAI is capable of generating example sentences suitable for elementary-level learners. Uchida (2025), using CEFR-J levels, reported that GenAI output often fails to match the level requested in the prompt; the findings do not replicate that pattern. The present data instead suggest that GenAI tends to favor vocabulary that learners at a given level are already likely to know, rather than the vocabulary they ought to know according to prescriptive CEFR descriptors. The divergence may also reflect differences in the underlying models: the present study employed the ChatGPT-o1 model, whereas Uchida used the ChatGPT-4o model. Model architecture or prompt handling could therefore account for the contrasting outcomes. To better understand the cause of the

discrepancy, future research should use the same prompts while varying the language model, and apply a consistent level-diagnostic tool (such as EVP or CEFR-J).

In addition, the results demonstrated that GenAI can generate level-appropriate example sentences from the perspective of the syntactic difficulty. The mean length of the generated sentences was 7.6 words. Uchida, Arase, and Kajiwara (2024) classified 7,511 English sentences into CEFR levels from A1 to C2, reporting mean sentence lengths of 8.40 words for A1 and 10.34 words for A2 sentences. Given these findings, the mean sentence length of 7.6 words in the current study is shorter than the CEFR A1 level. This suggests that the AI-generated example sentences are likely to be simple enough to be accessible for elementary-level learners. Furthermore, the PCSYN value of 2.28 (98.9th percentile) indicates that the syntactic simplicity of the AI-generated sentences is greater than that of 98% of texts in the Coh-Metrix reference corpus. This means the sentences are not only shorter but also use simpler, more familiar syntactic structures. In addition, the SYNNP value of 0.54, representing the mean number of modifiers per noun phrase, further supports the syntactic simplicity of these example sentences.

Considering the Bayesian Wilcoxon signed-rank test, it appears that the participants successfully learned the distinction between the synonyms *collect* and *gather* through the DDL activity. Despite the wide 95 % credible interval, an inevitable consequence of the modest sample size, the posterior median effect ($\delta \approx -0.68$) falls in the medium range. Boulton and Cobb's (2017) meta-analysis found medium-to-large effects for traditional DDL, so the present result suggests that DDL supplemented with GenAI can approach the efficacy of conventional corpus-based DDL. However, this finding should be interpreted with caution given the small sample size, and further research is needed to replicate these results.

It was observed that there was a tendency for some learners with pre-test scores higher than 13 to achieve lower scores in the post-test. Conversely, learners with pre-test scores below 12 did not show any decrease in their scores. Although this pattern may merely reflect a ceiling effect, it may also serve as evidence that DDL effectively operates within the framework of the usage-based model, which provides the theoretical foundation for DDL. As Boulton and Cobb (2017) state, proponents of the usage-based model (e.g., Tomasello, 2003) view language knowledge as a dynamic system that constantly evolves based on input. Assuming this to be the case, learners without prior knowledge of the distinction between *gather* and *collect* (i.e., those with pre-test scores below 12) have started constructing linguistic patterns, resulting in improved scores. Meanwhile, learners with prior knowledge (i.e., those with pre-test scores above 13) appear to restructure their understanding, leading to temporary backsliding due to input that contradicted their existing knowledge (McLaughlin, 1990). In other words, despite differences in score fluctuation tendencies, it can be assumed that both learners with and without prior knowledge gained some form of new understanding. This aligns with what Nishigaki and Akasegawa (2023) describe as the "learner's mission" in DDL: "to discover language facts and to modify or replace previous knowledge with new knowledge on his or her own" (p. 132). To validate this hypothesis, future longitudinal studies are necessary to verify whether the learners' scores exhibit a U-shaped developmental trajectory.

Pedagogical Implications

It is worth reiterating that this study originates from a learner-generated question: "What is the difference between *collect* and *gather*?" This point directly leads to pedagogical implications regarding when and for what linguistic items AI-based DDL can be utilized effectively. Semantic and usage differences among near-synonyms are often difficult to fully capture through dictionary descriptions alone, and DDL has been recommended for teaching such vocabulary (e.g., Sibel, 2023). However, in practice, DDL tools often lack sufficient examples in simplified English suitable for beginner learners. In particular, when focusing on lexical distinctions such as near-synonym usage rather than more general grammatical patterns, it is not hard to imagine the difficulty in collecting an adequate number of examples, given the effort required. In this context, AI-generated example sentences have the major advantage of being able to immediately supplement appropriate-level input.

That said, when using simplified example sentences with reduced lexical and syntactic complexity as in this study, there is a risk that learners are exposed to only a limited range of language patterns. As a compensatory strategy, it would be effective to position DDL with AI-generated example sentences as a "bridge to authentic corpus use," as proposed by Nishigaki and Akasegawa (2023, p. 135) in their development of hDDL. In other words, simple AI-generated sentences can be used at the early stages to reduce cognitive load, and learners can gradually transition to hands-on corpus exploration as their proficiency increases. This staged approach ensures progressive exposure to more diverse syntactic and lexical patterns.

Furthermore, when implementing DDL using AI-generated example sentences in educational settings, several factors warrant careful consideration. One key factor is the choice of model. The findings of this study suggest that selecting a model with strong reasoning capabilities can yield example sentences that better align with learners' proficiency levels. In addition, the design of prompts may play a critical role in ensuring output quality. According to the Prompt Engineering Guide by DAIR.AI (https://www.promptingguide.ai/), effective prompts typically comprise the following four components:

• **Instruction** - a specific task or instruction you want the model to perform
• **Context** - external information or additional context that can steer the model to better responses
• **Input Data** - the input or question that we are interested to find a response for
• **Output Indicator** - the type or format of the output

In the context of DDL, these components may correspond to:

• **Instruction** – a request indicating how many example sentences to generate
• **Context** – information such as participants' school grade or their CEFR level
• **Input Data** – target synonyms or grammatical structures
• **Output Indicator** – requirements such as KWIC format or bilingual output

Notably, the prompts employed in this study incorporated all four of these components.

Finally, in light of the possibility of U-shaped development observed in this study, DDL should not be positioned as a one-off learning activity. Rather, it should be positioned as one component within a larger learning framework, ensuring learners have ongoing opportunities for input.

Conclusion

Even though the present study is a small-scale, classroom-based case study, it revealed two key

findings. First, GenAI can generate example sentences that are level-appropriate for elementary learners to reference during DDL, in terms of both lexical and syntactic difficulty. Second, DDL using these AI-generated example sentences appears to be effective. These results suggest that GenAI has the potential to serve as a user-friendly concordancer. To date, DDL has not become a mainstream practice in language education, especially in secondary school settings. However, this study highlights that integrating GenAI into classroom activities can open up new possibilities for the active use of DDL among pre-tertiary learners.

Despite its promise, this study has several limitations that temper the generalizability of the findings. First, the study employed a single-group classroom case-study design. This modest sample size (n = 24) yielded a wide 95 % credible interval, and the lack of a matched control group meant that causal inferences remained tentative. Future research should adopt a randomized control-group design with a larger cohort to tighten these intervals and strengthen causal claims. Second, the intervention targeted only one synonym pair (*collect* vs. *gather*); the suitability of GenAI-generated examples may differ for other lexical items. Replication across a wider range of linguistic targets is therefore essential.

Third, the DDL activity relied on teacher-prepared handouts. Although a fine-grained analysis of students' annotations was beyond the scope of this study, the design of the handouts may have influenced the learning outcomes. For example, some tasks prompted learners to inspect the objects following each verb, yet nearly all participants articulated the gather–collect distinction primarily in semantic terms—e.g., *gather* means "to bring scattered items together," whereas *collect* means "to accumulate items over time, often as a hobby." A handout designed to more explicitly direct attention to grammatical and usage patterns, such as prototypical objects, might have fostered a deeper, form-focused understanding of the synonym pair. Moreover, replacing these teacher-prepared materials with a more interactive, hands-on format where learners query GenAI directly could increase engagement and, in turn, yield greater learning gains. Subsequent studies should empirically test this possibility, ideally incorporating a delayed post-test to gauge long-term retention. Finally, this study did not examine learners' perceptions of the DDL activity. Incorporating not only quantitative but also qualitative data—such as whether learners found the AI-generated example sentences appropriately easy—could yield deeper insights into the effectiveness of DDL with GenAI.

Although this study has several limitations, it nevertheless underscores the potential for DDL and GenAI to coexist. Both DDL and GenAI have their respective strengths and weaknesses, but their integration may pave the way for making DDL more accessible and actively utilized in educational settings, including primary and secondary schools. It is hoped that this research serves as a step toward broader adoption of the promising learning method in schools.

Note

During the preparation of this research paper, the author utilized ChatGPT for proofreading and to refine linguistic expressions. Following the use of this tool, the author thoroughly reviewed and revised the content. All responsibility for the final content rests solely with the author.

References

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*, 348–393. doi: 10.1111/lang.12224

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3), 66–89. doi: 10125/73450

Capel, A. (2010). A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, *1*, e3. doi:10.1017/S2041536210000048

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, *3*, e1. doi:10.1017/S2041536212000013

Crossley, S. A., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language, 26*(1), 92–113. doi: 10125/66686

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics, 3*(3). doi:10.1016/j.acorp.2023.100066

DAIR.AI. (2025, April). *Prompt engineering guide*. https://www.promptingguide.ai/

Flowerdew, J. (2024). Data-driven learning: From Collins Cobuild Dictionary to ChatGPT. *Language Teaching*, 1–18. doi:10.1017/S0261444824000144

Goss-Sampson, M. (2020). *Bayesian Inference in JASP: A guide for students*. doi: 10.17605/OSF.IO/CKNXM

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. Educational Researcher, 40(5), 223-234. doi：10.3102/0013189X11413260

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4*, 1–16. https://lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf

Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics, 3*(3). doi: 10.1016/j.acorp.2023.100065

McLaughlin, B. (1990). Restructuring. *Applied Linguistics, 11*(2), 113–128. doi: 10.1093/applin/11.2.113

Miočević, M., Levy, R., & van de Schoot, R. (2020). Introduction to Bayesian statistics. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 1–12). Routledge.

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics, 3*(3). doi: 10.1016/j.acorp.2023.100074

Mizumoto, A., & Chujo, K. (2015). A Meta-analysis of Data-driven Learning Approach in the Japanese EFL Classroom. *English Corpus Studies,* 22, 1-18.

Mizumoto, A., & Teng, M. F. (2025). Large language models fall short in classifying learners' open-ended responses. *Research Methods in Applied Linguistics, 4*(2), Article 100210. doi:10.1016/j.rmal.2025.100210

Nakahara, R., & Sugiura, M. (2025). Generalizing linguistic patterns through data-driven

learning: A study of the dative alternation in Japanese learners of English. *Language Teaching Research*, *0*(0). doi:10.1177/13621688241305787

Nishigaki, C., & Akasegawa, S. (2023). Development and revision of DDL tools for secondary school students: What we can do to nurture autonomous corpus users? *English Corpus Studies, 30*, 131–149.

O'Keeffe, A. (2023). A theoretical rationale for the importance of patterning in language acquisition and the implications for data-driven learning. *Nordic Journal of English Studies, 22*, 16–41. doi: 10.35360/njes.793

Pérez-Paredes, P. (2020). The pedagogic advantage of teenage corpora for secondary school learners. In P. Crosthwaite (Ed.), *Data driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb–preposition collocations. *System, 91*, Article 102268. doi: 10.1016/j.system.2020.102268

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL*, *26*(2), 184–201. doi:10.1017/S0958344014000081

Söğüt, S. (2024). L2 Learners' Perspectives on data-driven learning for identifying properties of near-synonymous words: A convergent mixed-methods study. *GIST – Education and Learning Research Journal*, *27*. doi: 10.26817/16925777.1777

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Uchida, S. (2025). Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction*, *14*(1), 2078. doi: 10.29140/vli.v14n1.2078

Uchida, S., Arase, Y., & Kajiwara, T. (2024). Profiling English sentences based on CEFR levels. *International Journal of Applied Linguistics, 175*(1), 103–126. doi: 10.1075/itl.22018.uch

Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology, 18*, 96–117. doi: 10125/44356

Ryosuke NAKAHARA is an English teacher at Yamate Gakuin Junior and Senior High School in Yokohama, Japan. His current research interests include data-driven learning and computer-assisted language learning, particularly through the use of GenAI and game-based classroom response systems. ryosuke_nakahara@yamate-gakuin.ed.jp

Appendix

Appendix 1
The example sentence handout

似た意味の単語の違いを発見しよう！ 例文シート

3年（　）組（　）番（　　）

gather

1	We can	gather	apples from the tree.	私たちは木からリンゴを集めることができます。
2	They	gather	together every Sunday.	彼らは毎週日曜日に一緒に集まります。
3	She likes to	gather	flowers in the garden.	彼女は庭で花を摘むのが好きです。
4	Let's	gather	in the library after class.	授業の後で図書館に集まりましょう。
5	Can you	gather	some water for me?	私のために少し水を汲んできてもらえますか？
6	I want to	gather	my friends for a party.	私はパーティーのために友達を集めたいです。
7	Birds	gather	on the roof in the morning.	鳥たちは朝、屋根の上に集まります。
8	Students	gather	their books at the end of class.	生徒たちは授業の終わりに本をまとめます。
9	People	gather	at the bus stop every day.	人々は毎日バス停に集まります。
10	We	gather	information from the internet.	私たちはインターネットから情報を集めます。
11	Please	gather	your toys before dinner.	夕食の前におもちゃを片付けてください。
12	They	gather	shells on the beach for fun.	彼らは遊びとしてビーチで貝殻を集めます。
13	I	gather	my thoughts in the morning.	私は朝に考えを整理します。
14	Children	gather	around the teacher to listen.	子どもたちは先生の周りに集まり、話を聞きます。
15	We	gather	around the campfire at night.	私たちは夜、キャンプファイヤーの周りに集まります。

collect

1	I like to	collect	stamps from different countries.	私はいろいろな国の切手を集めるのが好きです。
2	She will	collect	shells at the beach.	彼女はビーチで貝殻を集める予定です。
3	They	collect	old coins for a hobby.	彼らは趣味で古いコインを集めています。
4	He loves to	collect	comics in his free time.	彼は暇なときに漫画を集めるのが大好きです。
5	We can	collect	rainwater in that bucket.	私たちはそのバケツで雨水を集めることができます。
6	My sister wants to	collect	dolls.	私の妹は人形を集めたがっています。
7	You should	collect	all the papers and files.	あなたはすべての書類やファイルを集めるべきです。
8	Let's	collect	our things and go home.	持ち物をまとめて帰りましょう。
9	I plan to	collect	more photos for my album.	私はアルバムのためにもっと写真を集めるつもりです。
10	She tries to	collect	postcards when she travels.	彼女は旅行のときにポストカードを集めようとします。
11	He asks his friends to	collect	stickers with him.	彼は友達に、一緒にステッカーを集めようと誘います。
12	My brother and I	collect	trading cards together.	兄（弟）と私は一緒にトレーディングカードを集めています。
13	We need to	collect	flowers for the school event.	学校の行事のために花を集める必要があります。
14	You can	collect	data on your computer.	コンピューターでデータを収集できます。
15	They	collect	books to donate to the library.	彼らは図書館に寄付するために本を集めます。

Appendix 2
The complete set of the test

似た意味の単語の違いがわかるかな？

3年（　）組（　）番（　　）

次の各文の下線部分に **(A) gather** または **(B) collect** のどちらか適切なものを選び記号で答えなさい。

(1) ファンたちはいつも選手を待つためにスタジアムのゲートに集まる。
The fans always _____ at the stadium gates to wait for the players.

(2) 彼女は地元のチャリティーイベントのために寄付金を集めることを望んでいます。
She wants to _____ donations for the local charity event.

(3) この紙が飛ばないようにするために、石をいくつか集めてもらえますか？
Could you _____ some stones to hold down this paper?

(4) 私たちはそのイベントのために、多くのボランティアを集める必要があります。
We need to _____ many volunteers for the event.

(5) 正午までに、すべてのテスト用紙を集める必要があります。
I need to _____ all the test papers by noon.

(6) 話をする前に、少し時間を取って考えを整理するようにしましょう。
Take a moment to _____ your thoughts before you speak.

(7) 各生徒から図書費を集めるのを手伝ってもらえますか？
Could you help me _____ the library fees from each student?

(8) 研究のために、このテーマについてさらに多くのデータを集める必要があります。
We need to _____ more data on this topic for our research.

(9) ショーを始める前に、みなさんの注意を集める必要があります。
We want to _____ everyone's attention before we start the show.

(10) 彼らはいろいろな都市を訪れるたびに、ポストカードを集めることを望んでいます。
They want to _____ postcards from different cities they visit.

(11) テーブル用に、庭から花を集めるつもりです。
I'm going to _____ some flowers from the garden for the table.

(12) 教室を出る前に、持ち物をまとめるようにしてください。
Please _____ your belongings before leaving the classroom.

(13) 私たちは貧しい人々のために古い服を集めています。
We _____ old clothes for the poor.

(14) 私たちはイベントのために十分なスポンサーを集めようとしました。
We tried to _____ enough sponsors for the event.

(15) 話し合いをするために、イスを円形に集めましょう。
Let's _____ the chairs in a circle for the discussion.

(16) 彼は趣味として古い切手を集めることに決めました
He decided to _____ old stamps as a hobby.

(17) 私たちは公園を綺麗に保つためにペットボトルや空き缶を集めます。
We _____ bottles and cans to keep the park clean.

(18) 私たちが見つけた証拠を寄せ集めて、それを分析しましょう。
Let's _____ the evidence which we found and analyze it.

(19) 彼らは古い公園を守るために署名を集める予定です。
They plan to _____ signatures to save the old park.

(20) 司書は明日、延滞している本を回収します。
The librarian will _____ overdue books tomorrow.

解答欄

(1)	(2)	(3)	(4)	(5)
(6)	(7)	(8)	(9)	(10)
(11)	(12)	(13)	(14)	(15)
(16)	(17)	(18)	(19)	(20)