Data-driven Learning Approach for Teaching Verb-Noun Collocations in an English Foreign Language Context

Awatif Katib Alruwaili[1]
Northern Border University, Saudi Arabia

Abstract
This study addresses the learners' actual use of corpora in the classroom for learning verb-noun collocations. The target items were selected through a multi-level approach which consisted of three levels: a corpus-based approach, a phraseological approach and a pedagogical approach. The study relied on data collected from fifty-one participants studying general English at the intermediate level in the foundation year at a university in Saudi Arabia. The study ran for five weeks and included three training sessions for learners on how to use a corpus resource (AntConc) and to read and analyse concordance lines. There were also two test sessions. The participants were tracked via software tracker in both the training and testing sessions. Data were collected through tracking logs and activity sheets. The main finding was that the participants were able to apply the DDL (Data-driven Learning) approach independently in the same way as they had been trained, which indicates that the training was successful. The learners were also able to identify the general verb patterns through the use of concordance lines.

Keywords

1. Introduction
Although the use of corpora in the classroom is no longer a novel methodology, it has not yet reached "maturity" (Boulton and Pérez-Paredes 122). As Boulton and Pérez-Paredes observed, "the focus is switching from corpus linguistics to language pedagogy" (122), therefore, the focal point is on facilitating the learner's use of corpus resources and approaches rather than justifying the use of corpora in the classroom or mentioning the benefits of corpus linguistics. In addition, there is not much evidence or research that corpora can be used as an alternative to textbooks and traditional resources, such as dictionaries (Chambers, "Integrating Corpus" 111). The state of corpora use has not changed significantly since Chambers' article, as has been emphasised by several later researchers,

---

[1] **Dr. Awatif Katib Alruwaili** is Assistant Professor in the Department of Languages and Translation, Faculty of Education and Arts, Northern Border University, Arar, Saudi Arabia. Her research interest is in corpus linguistics and language teaching, with primary focus on corpus-based approaches for teaching collocations. Email: awatif.k.alruwaili@nbu.edu.sa.

including Boulton ("Corpus Consultation for ESP" 261); Hughes (401-2); Boulton and Pérez-Paredes (122-23); Leńko-Szymańska and Boulton (2); and Vyatkina ("Data-driven Learning in Language Pedagogy" 207). Numerous scholars have stressed that there is a lack of research that provides direct evidence of learners' use of corpus resources in EFL contexts, including the Saudi context (see, e.g., Hafner and Candlin 304; Pérez-Paredes et al. 234). This article focuses on the pedagogical aspect of observing learners' use of corpus resources by means of a tracking software in the context of a Saudi classroom with the aim of providing an alternative approach to teaching English to Saudi learners by evaluating the effectiveness of the training on the use of the DDL approach and corpus tools.

2. Theoretical Background

The data-driven learning (DDL) approach can be defined as "the use in the classroom of computer-generated concordances to get students to explore the regularities of patterning in the target language" (Johns and King 1). The approach is based on an inductive approach in teaching (Bernardini 16), where the students analyse the targeted features with or without the teacher's help. Other researchers have studied the effectiveness of DDL applications and attitudes towards them in language learning and teaching (e.g., Hafner and Candlin 303). However, despite an increase in research on corpus consultation, whether direct or indirect, practitioners' awareness is still low in recognising the major benefits that corpus-based resources can provide (Boulton, "Data-driven Learning: On Paper, in Practice" 19; Breyer 154). This is due to many factors, including the nature of the DDL approach and its relation to learners, teachers and resources.

DDL is a communicative, task-based, learner-centred approach that fosters autonomy and learning by discovery. Yet, practitioners and researchers have dealt with it with caution because of the large quantities of concordance lines that may be difficult to manage for learners. Furthermore, DDL is more useful in specific language areas that employ "fuzzy" knowledge that relates to frequent and typical uses rather than rules, such as synonyms, collocations and colligations (Boulton, "Data-Driven Learning: Reasonable Fears" 83). Another challenge is the difficulty of extrapolating rules from corpora data and applying them (Boulton, "Data-driven Learning: Reasonable Fears" 85).

Despite the obstacles in adopting DDL, this approach has been linked to a number of benefits that encourage its use. Firstly, it allows the extracting of rules based on language usage and presenting patterns to explain exceptions. Secondly, it sheds light on the importance of frequency when learning rules and forms, something that is absent from traditional learning methods. Thirdly, a potentially highly beneficial aspect of DDL is language learning through different registers.

2.1 Language Learning Theories Underpinning the Data-driven Learning Approach

A close relationship exists between the DDL approach and second language acquisition theories. Both are based on acquiring language as chunks rather than as morphemes or words in isolation, thus encouraging and stimulating inductive learning strategies. By using corpora as a mediation tool, DDL is, consequently, linked to sociocultural theory. Furthermore, DDL, by definition, urges learners to notice the similarities and regularities of a pattern through reading concordance lines. This is an influential factor in noticing, under the Noticing hypothesis, which is described further in Section 2.1.2.

2.1.1 Sociocultural Theory and the Data-driven Learning Approach

Sociocultural theory views language as "a semiotic tool" (Ellis 516) and learning as a "social phenomenon" (Villamil and Guerrero 23). Language is a means to achieve social goals in the sense that it is a tool to transfer our thoughts and provide a means of communication. Based on this view of language, Sociocultural theory coincides most closely with corpus linguistics theories that focus on meaning rather than form and on communication as the Extended Lexical Unit of Meaning theory. One of the central concepts of Sociocultural theory is mediation, which is defined as the "process through which humans deploy culturally constructed artefacts, concepts, and activities to regulate (i.e., gain voluntary control over and transform) the material world or their own and each other's social and mental activity" (Lantolf and Thorne 79). Therefore, learning does not take place solely in the head of the learner but also through contact with the object (in this case, corpora) to acquire the target form of the language. The learner has to develop or acquire knowledge by employing artefacts and using mediating tools (Ellis 524). The artefacts are "fundamentally social in nature" (Ellis 524), as is the case in this study, where the guided inductive use of corpora can be an artefact for independent use, and learners use mediated tools to develop their knowledge of general verbs patterns. This is illustrated in figure 1.
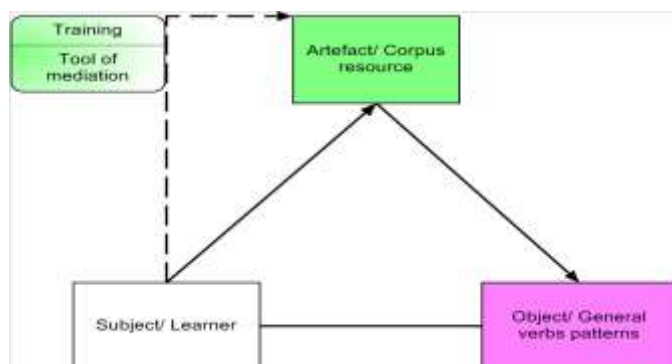
Fig. 1. The Mediating Nature of Learner and Corpus Resource (adapted from Lantolf and Thorne 62)

2.1.2 The Noticing Hypothesis and Data-driven Learning

The Noticing hypothesis is frequently referred to within DDL because of its close relationship with noticing target patterns through concordance lines (Flowerdew 19). According to the Noticing hypothesis, the "learner must attend to and notice linguistic features of the input that they are exposed to if those forms are to become intake for learning" (Schmidt, "Attention" 29). The inductive approach is considered the backbone of DDL, which is dependent solely on the Noticing hypothesis. However, the Noticing hypothesis is rarely linked "explicitly" to corpus-based studies (Flowerdew 20). Three theoretical considerations underlie the Noticing hypothesis: consciousness as intention, consciousness as attention and consciousness as awareness (Schmidt, "The Role of Consciousness", 131-133). In many corpus studies (e.g., Gaskell and Cobb 308; Kennedy and Miceli, "Corpus-assisted Writing" 34), the Noticing hypothesis was employed using the guided inductive approach and deductive approach. The noticing is initiated by the teacher in the inductive approach while in the deductive approach, the noticing is initiated by the learners.

2.2 Empirical DDL Research

Studies that have recorded the active exploitation of corpus tools can be divided into two groups based on the data extraction methods they use: manual logs and software logs. The former was employed in studies such as by Chambers and O'Sullivan (163), O'Sullivan and Chambers (55) and Varley (137) whilst the latter was utilised in studies by Cobb (303), Gaskell and Cobb (308), Chan and Liou (241), Hafner and Candlin (308), and Pérez-Paredes et al (239).

Manual logs can be employed in diverse ways. For instance, there are different forms of logs that learners fill in and various types of information learners collect, such as search words, search results, resources that have been used and/or learners' reflections and conclusions. In this regard, Chambers and O'Sullivan (163) and O'Sullivan and Chambers (55) used a manual log to track the changes (correcting errors) that learners made while using a concordance programme aimed at improving writing skills. Gaskell and Cobb (308) used manual logs and Internet Protocol (IP) tracker to monitor learners' interactions with concordances to ensure that learning took place through the concordances. Therefore, it can be claimed that the results of their research have an extra degree of validity as the authors relied on the actual results of the interaction process during consultation with the concordance. Varley exclusively employed reflective logs to draw learners' attention to the strategies and types of information corpora and concordances provide (137). The logs were similar to reflex reports, in which

learners describe their attitudes towards using corpus tools in learning lexico-grammatical features.

Studies that use software trackers (software logs) are also diverse. Chan and Liou used a tracker programme to record learners' actions during the concordance consultation, such as searched words, number of times each word was searched and answers to each online unit (241). Higher scores were indicated for the items taught through concordances than by the traditional methods. Therefore, it can be assumed that the concordance helped in improving collocation knowledge based on the results of the test and the tracker records.

A significant contribution to corpus linguistics research was Cobb's use of protocol files that recorded the interactions between learners and the PET.200 (a CALL programme) (303). The files presented not only the number of interactions with the programme but also the time spent on every task. The study revealed that the treatment group that used the concordance lines spent more time than the control group reading the concordance lines during the consultation, whereas the actual time spent on doing the tasks was the same in both groups.

Many studies (Hafner and Candlin 308; Yoon and Hirvela 277; Chambers and O'Sullivan 162; O'Sullivan and Chambers 54; Varley 146; Geluso and Yamaguchi 240) indicate that training learners in the use of the DDL approach as well as corpus tools is essential to fostering positive attitudes towards corpus tools and the corpus approach. At the same time, it seems that learners' proficiency levels do not significantly affect their attitudes. What is missing from the reviewed studies is the hands-on training and descriptions of how learners actually use the corpus tool and how they cope with the DDL approach, with exceptions of Varley (137), and Hafner and Candlin (312).

The reviewed studies shed light on two main arguments. First, learners' proficiency level is not essential for the use of corpus-based resources in the classroom, although it may present some difficulties at the beginning. Second, gradual and guided training is essential for the successful implementation of corpus-based resources in the classroom. Studies conducted by Vyatkina ("Data-driven Learning for Beginners" and "Data-driven Learning of Collocations") demonstrate that a gradual and guided use of the DDL approach is more effective than the unguided. In addition to the scarcity of research in the areas outlined above, most research on learners' uses of corpora in the classroom do not provide sufficient information on pre-task training with corpus tools. Thus, this article targets these gaps.

2.3 Research Questions
The study explores the following research questions:

1. Do intermediate Saudi students know how to use the DDL approach independently in the language classroom for learning general verbs? If so, How?

2. Is there a link between the participant's use of the DDL approach and the successful completion of tasks?

3. Materials and Methods
3.1 Participants
The participants were 51 female Arabic-speaking students in their foundation year at the Northern Border University (NBU) in Saudi Arabia. All completed three training sessions and two testing sessions. According to the NBU placement test, the participants (N=51) were intermediate level EFL students. The participants were between the ages of 18 and 25 and studied English in school for approximately seven years. The study ran for five weeks.

The experiment was divided into two sections: training sessions and testing sessions. Each session was held in a computer laboratory for an hour. The training sessions focused on how to use the corpus tool, AntConc (specifically, it is a concordancer). These three training sessions included a short introduction to corpus linguistics and how it is used in language analysis. During the training sessions, the participants were trained to use AntConc for investigating general verbs patterns following Sinclair's model (*Reading Concordances*).

During the testing sessions, the participants were asked to perform tasks with no guidance, following the same procedure as in the training sessions but with different items. Data were collected from the testing sessions in two ways: (a) participants completing activities (tasks) by answering set questions, and (b) the information generated by the tracker (tracking logs).

3.2 Materials
3.2.1 Reading Concordances Procedure
Sinclair's reading concordance model, adopted here, is both comprehensive and detailed. I chose this framework because it encourages three major cognitive processes in learning vocabulary: noticing, retrieval and creative use. Nation (98-100) argued that these three cognitive processes are significant conditions for improving vocabulary learning as well as in evaluating vocabulary-teaching tasks. Sinclair's framework informed the design for the teaching materials used in this experiment, and the activities were divided according to the framework which consists of seven steps: *initiate, interpret, consolidate, report, recycle, result* and *repeat* (Sinclair, *Reading Concordances* xvi-xvii).

This framework seems to be designed for use by researchers more than learners as it is time-consuming and thus may demotivate learners. Thus, I modified the design slightly for classroom use, retaining the following components: *initiate, interpret, consolidate* and *report*. The *recycle* and *repeat* steps were excluded because I used a pre-prepared set of data that did not allow for additional investigation. The *result* step was incorporated with the *report* step to

save time. The steps throughout the session were labelled as follows: *observe* (initiate), *investigate* (interpret and consolidate) and *report*.

This study included seven verbs which are: *have*, *do*, *make*, *take*, *give*, *get* and *put*. The verb *put* was used only as an example in the introductory session. Each session included three verbs and three collocates for each verb. The testing sessions were duplicates of the training sessions, except for the collocates of each verb. There was no intervention or guidance from the teacher. Table 1 shows the general verbs and their collocates used for each session.

Table 1
General Verb Patterns Featured in Training and Testing Sessions

Session	1st GV	2nd GV	3rd GV
1st training session	**Make** (sense, use, contact)	**Have** (fun, nightmares, dealings)	**Give** (birth, notice, effect)
2nd training session	**Take** (place, advantages, care)	**Do** (justice, homage, battle)	**Get** (permission, access, dark)
1st testing session	**Make** (amends, comparisons, jokes)	**Have** (hysterics, priority, similarities)	**Give** (chase, offence, interviews)
2nd testing session	**Take** (hold, effect, office)	**Do** (penance, homework, deals)	**Get** (revenge, ideas, leave)

The materials in this study were divided along the following lines:

The introductory session (technical stage): procedures and steps in using corpus analysis tools (first training session).

Training sessions (linguistic stage): language deduction and analysis, the product to be achieved by the learners (teaching materials, including activities and tasks). The main goal of this stage was to teach the learners to use a concordancer to identify general verb patterns and encourage learners to focus on patterns and units of meaning (sessions two and three).

Testing sessions (assessment stage): the completion of the tasks (in sessions four and five).

3.2.2 Activities
The activities were based on encouraging three cognitive processes (noticing, retrieval and creative use) that the use of the concordance lines could stimulate (Nation 116). The activities were divided into three themes:

A. Generating the concordance lines
B. Searching or investigating the keywords
C. Understanding the meanings of the keywords from the context (through reading the concordance lines for each collocation)

The tasks were divided into three sections: (1) noticing the pattern which includes activities that encourage participants to notice the repeated pattern of the verb and its collocates; (2) form recognition which includes activities that encourage the student to notice the form of the collocations through reading the concordance lines; (3) meaning recognition which encourages participants to determine the meaning of each collocation by reading the concordance lines. Students are then asked to produce the meaning of the collocation by explaining the differences between each verb pattern. They then write the explicit meaning of the collocation (either in English or in Arabic).

3.3 Data Collection
Data collection was conducted using a mixed-method design consisting of tracking logs and activity sheets (tasks). The All-In-One Keylogger software was installed on all the computers in the English Lab. At the end of each session, I logged in as an administrator and extracted the data files (screenshots and textual logs) for each participant and saved them on a hard-disk according to a number assigned to each participant.

4. Findings
4.1 The Learning Process of the DDL Approach in General Verb Patterns
To examine whether the participants had learnt how to use the DDL approach independently following the two guided sessions, the participants' tracking logs from the training and testing sessions were examined and compared in terms of two main variables: (1) participants' successful completion of the steps required to complete the tasks (henceforth *process*) and (2) the time participants spent completing the different stages of the process during each session (henceforth *time spent*). The comparison of the process was based on the steps taken to investigate each verb in each session. A comparison of the training session and testing session logs ascertained if the training was successful and if participants were able to work independently during the testing sessions.

4.1.1 The Use of the *Process*
The *process* scores from the first and second training sessions and the second testing session showed similar mean values (12.50), while the second testing session demonstrated a high standard deviation (1.08), indicating greater variance. The first testing session reflected the lowest mean (11.70) and a high standard deviation (1.62). The general performance of the entire process was better in the

training sessions than in the testing sessions, with a difference of 1.4 in mean values, as shown in table 2.

Table 2
Means (M) and Standard Deviations (SD) of the Process During Each Stage and for Each Session

Steps	Training 1		Training 2		Testing 1		Testing 2	
	M	SD	M	SD	M	SD	M	SD
	(Max=3)							
Type	3.00	0.00	2.98	0.14	2.84	0.37	3.00	0.00
Start	3.00	0.00	2.98	0.14	2.84	0.37	3.00	0.00
Sort	2.96	0.20	2.98	0.14	2.78	0.42	2.98	0.14
R1	2.75	0.59	2.94	0.31	2.27	1.00	2.53	1.03
Total of the four steps per session	11.71	0.79	11.88	0.73	10.75	2.15	11.51	1.17
Total session score + the score from the download corpus file	12.7	0.67	12.90	0.62	11.70	1.62	12.50	1.08
Total per session type	M (25.6) SD (1.29)				M (24.2) SD (2.70)			

As each of the four steps must be followed during each session and for each verb, the means for all steps should be very similar. A difference occurred during step 4 (R1, one word to the right of the target word.); this difference was observable in both session types and was accompanied by a high standard deviation among participants.

The Friedman test was conducted to examine the differences in *process* scores across the four sessions. The results indicated a statistically significant difference in the processes followed across the four sessions: first training, second training, first testing, second testing, x^2(3, n=51) = 31.42, p<.001. The median values showed a decrease in the steps followed for the first testing session (Mdn=12), while the other sessions displayed similar values (Mdn=13).

After establishing the presence of a statistically significant difference somewhere among the four sessions, a Wilcoxon test was used as a follow-up (pairwise comparison of Bonferroni adjusted p-values). This test showed that there was a statistically significant difference between the scores of the first testing session and the training sessions. The first testing session had the lowest score, therefore signalling the fewest number of steps completed, as shown in figure 4.
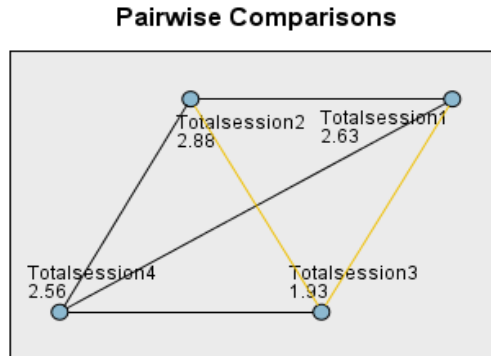
Pairwise Comparisons

Fig. 2. Pairwise Comparison of the Processes Followed During the Four Sessions

The significance of the difference between the first testing session and the first training session was $p=.039$, while it was $p=.01$ between the first testing session and the second training session. Other comparisons failed to show significant differences.

The above analysis showed the overall trends and the general differences across the testing and training sessions. Table 3 displays the percentage of participants who successfully completed each step for each verb in each session. A detailed discussion of the sessions follows.

Table 3
The Percentage of Participant Completion of All Steps for Each Verb (All Sessions)

	1st training session			2nd training session			1st testing session			2nd testing session	
Steps	Make	Have	Give	Take	Do	Get*	Make	Have*	Give	Take	Do
Type in GV	100%	100%	100%	100%	100%	98%	98%	86%	100%	100%	100%
Start	100%	100%	100%	100%	100%	98%	98%	86%	100%	100%	100%
Sort	100%	100%	96%	100%	100%	98%	94%	84%	100%	98%	100%
Use R1	100%	90%	84%	100%	98%	96%	72.5%	68.6%	86%	82%	84%

First Training Session. The participants investigated three general verbs (*make, have, give*) under the guidance of the researcher and followed the required steps. Some participants did not use the R1 feature with the second and third verbs. The steps followed by the participants during this process are outlined in table 3.

Participants' application of the sort and R1 features clearly diminished over the course of the session, perhaps because they felt it unimportant or forgot how to execute it.

Second Training Session. Although all sessions were designed similarly, a wildcard (* is a symbol that can be used to search for words of unspecified prefix or suffix, e.g., *talk** would yield concordance lines including *talked, talking,* etc.) was included in the second training session (the third verb, *get*) to determine whether participants could properly utilise a wildcard during their investigations. However, during the second training session, one participant did not use the wildcard. Thus, she did not discover all the patterns related to *get* (if the wildcard * was not used, only some of the concordance lines would appear; see the screenshot taken of Participant 43's work in figure 5). The majority of the participants followed the required steps.
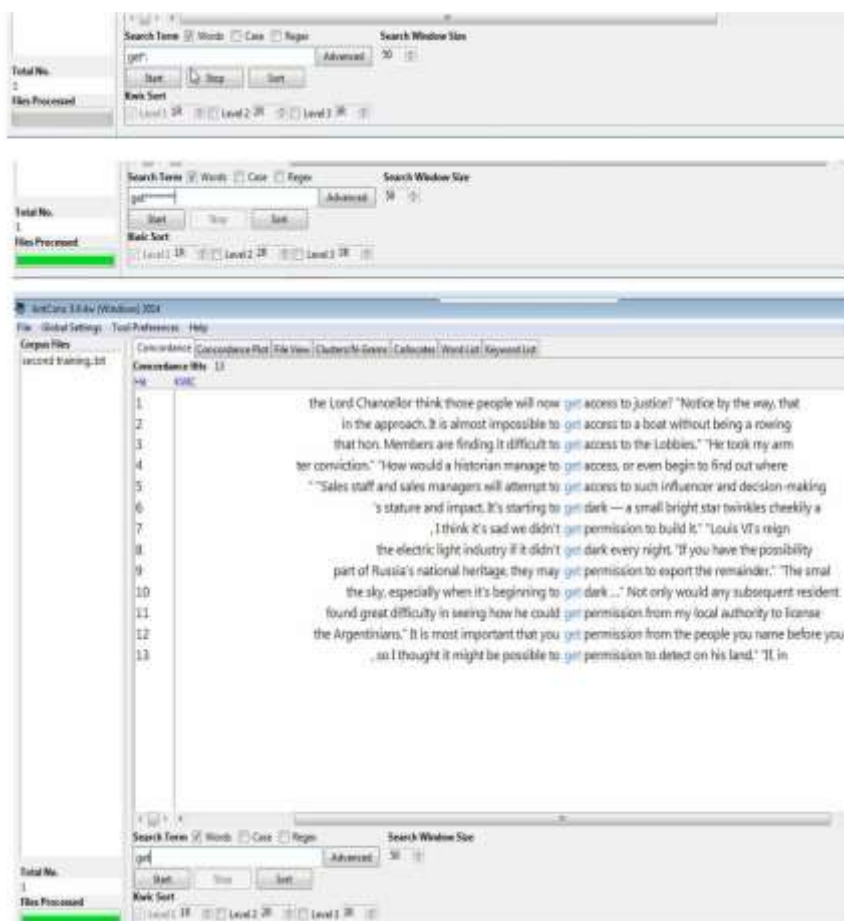

Fig. 3. The Screenshot Taken of Participant 43's Work

First Testing Session. During this session, the participants received no guidance from the researcher and engaged in the activities independently based on the instruction they received during the training sessions. The participants had to use a wildcard for the second verb, *have*, as specified in the handout provided for the session. The participants did not appear to perform well with the second verb, which can be explained by their apparent difficulty with the wildcard (*). Some participants did not utilise the R1 feature, perhaps because they were able to determine the pattern without it, which rendered the step unnecessary, as shown in table 3.

Second Testing Session. The participants' performance, as measured by the use of the required steps, improved during this session. As noted in the first testing session, some participants did not use the R1 feature during this session. As illustrated in table 3, the participants' performances were similar across all three verbs.

4.1.2 Technical difficulties

In exploring the participants' use of the DDL approach, it is essential to discover any technical difficulties participants may have experienced during the sessions. Two technical hurdles were noticed: the downloading of the data file in each session and the use of the wildcard (*) during the second training session and first testing session. Tables 4 and 5 illustrate the participants' attempts to utilise these technical functionalities that might have affected the overall process. All participants successfully downloaded the data file, though some required several attempts; however, one participant failed to use the wildcard during the second training session, and seven failed to use it during the first testing session.

Table 4
Attempts to Download the Data File

Session	Attempts to download the data file				
	1st attempt	2nd attempt	3rd attempt	4th attempt	5th attempt
1st training	86.3%	13.7%	0	0	0
2nd training	76.5%	19.6%	0	2%	2%
1st testing	54.9%	39%	2%	2%	2%
2nd testing	74.5%	19.6%	5.9%	0	0

Table 5
Attempts to Use the Wildcard *

Session	Attempts made to use *				
	1st attempt	2nd attempt	3rd attempt	4th attempt	Failed to use
2nd training	90%	3.9%	5.9%	0	2%
1st testing	74.5%	5.9%	15.7%	3.9%	13.7%

4.1.3 Time Spent on the Process

As previously stated, the second step required to answer the first research question involved an analysis of the time spent on each session. Significant differences in the *time spent* by participants on the various sessions could indicate that using the concordancer became easier, grew more difficult or maintained the same difficulty level.

The mean values showed a decrease in *time spent* from the first training session (M=57.8) to the second training session (M=41.6). Moreover, *time spent* decreased between the first testing (M=41.5) and second testing sessions (M=38.9). Of all the sessions, the first training session demonstrated the highest *time spent*. While both training sessions were instructor-facilitated with the same guidance provided during each, the participants spent more time on the first training session.

A one-way repeated measure (ANOVA) was conducted to compare the time devoted to each session. A significant effect was found for the time spent on each session, Wilks' Lambda=.22, f(3,48)=56.2, p<.001, multivariate partial eta squared=.78. The effect sizes were large (.78) according to Cohen's (284-87) criteria for effect (.01=small, .06=moderate and .14=large).

There was a significant difference between the *time spent* on the first and second training sessions (M=16.28, *p*=.000), the first training and the first testing sessions (M= 16.3, *p*=.000) and the first training and the second testing sessions (M=18.88, *p*=.000). There was no significant difference, however, in the *time spent* between the second training, first testing and second testing sessions.

Inferential tests were conducted to discover any differences between the testing sessions in terms of the process and the *time spent* on each step. Paired-sample *T*-tests were also conducted to uncover any significant differences in the time spent during each testing session. The results indicated that there was no significant difference in *time spent* between the testing sessions. In fact, the mean value of *time spent* on the first testing session was similar to that of the *time spent* on the second testing session. Thus, the participants were shown to have spent a similar amount of time on both testing sessions.

4.2 DDL Use and the Successful Completion of the Activities
This section attempts to answer the second research question by uncovering a relationship between the use of DDL and the completion of the activities. Only the testing sessions were examined to investigate the participants' performances, wherein they worked independently. The participants' performances (scores on activities) during the two testing sessions were first checked to determine if there was a significant difference between them.

4.2.1 Participants' Performances
The Wilcoxon Signed Rank Test was conducted to determine any difference in activity scores between the two testing sessions. The results indicated a significant difference between the scores of the testing sessions: $z= -4.092$, $p<.05$, with a moderate effect size ($r= 0.37$) according to Cohen's guidelines. Table 6 outlines the means and standard deviations of the activity scores in both testing sessions.

Table 6
Means and Standard Deviations of the Activity Scores in Both Testing Sessions

Activity scores	Mean	SD
1st testing session	44.49	3.83
2nd testing session	46.77	1.63

The mean of the ranks in favour of the second testing session was 21.31 (SD=1.6), while the mean of the ranks in favour of the first testing session was 14.00 (SD=3.8). The scores for the second testing session were higher than those obtained during the first testing session (see table 6). Further, the process scores for the second testing session were higher than those obtained during the first testing session. This result indicated that the participants' ability to follow the same DDL approach processes improved during the second testing session.

The participants' responses to the activity questions were further examined according to the type of activity. As the activities had been divided into three cognitive processes, the participants' performance on each one was examined to determine how participant engagement with the concordance lines could enhance those processes, as described in the following paragraphs.

a. *Noticing the Pattern:* The noticing activity sought to draw participants' attention to the repeated verb patterns and lead them to identify those patterns. The participants succeeded in noticing the patterns.
b. *Form Recognition:* The aim of this activity is to identify the verb patterns by attempting to generalise those patterns. The participants successfully recorded the repeated patterns, except for *have* (74.5% recorded only three

repeated patterns), which can be explained by their failure to use the wildcard (*) (see table 5). The participants were able to generalise the general verb patterns based on the concordance lines. Therefore, the participants noticed the patterns that were repeated from previous activities; they then discovered that the repeated patterns could be used to form the hypothesis that the verb was always followed by a noun. Most participants recognised the form of the repeated words that followed the general verbs (noun).

  c. *Meaning Recognition:* This activity determined whether the concordance helped the participants to recognise the meaning of verb patterns. The results indicated that the participants were able to identify the meaning of most verb patterns; moreover, their performance improved during the second testing session. However, care must be taken while discussing these findings, as the two sessions had different sets of general verbs and, therefore, the results cannot be generalised. What can be argued is that the concordancer did help participants to recognise the meaning of the general verb patterns. Previous studies have yielded similar results, such as those found in Chan and Liou (241).

5. Discussion

This study found that learners used the DDL approach to complete activities independently and spent less time in each subsequent session. The results are consistent with other studies arguing that guided use, efficient training and teacher-guided activities result in better use and greater benefits in learning the targeted items (Chambers 120; Cheng et al. 183; O'Sullivan and Chambers 61; Yoon and Hirvela 264; Smart 197).

5.1 Independent Use of the DDL Approach

The main finding suggests that the participants were able to use the corpus tool and apply the DDL approach independently after the training. The performance in the second testing session was better than in the first testing session. Less time was spent on both testing sessions than on the training sessions, which was anticipated due to the guidance and instructions the learners received in the training session. However, even though the participants received the same instructions and guidance in both training sessions, the amount of time they spent in the second training session decreased, possibly suggesting a developed familiarity with the DDL approach and tool. The same thing happened in the testing sessions: the amount of time decreased from the first testing session to the second, but with no significant difference.

  In general, the performance of the entire process was better in the training sessions than in the testing sessions, with the first testing session having the lowest score, thus signalling the fewest number of steps completed. This might be explained by the participants' lack of familiarity with using the corpus independently. However, we must be cautious about this result as it could also be

related to the participants' difficulty in using the wildcard for the second verb (*have*) in that session. The improvement in performance in the second testing session suggests that the participants took the lead in using the DDL approach independently and might have been more confident in doing so. This finding supports Kennedy and Miceli's ("Corpus-assisted Creative Writing" 29) suggestion that time and practice are necessary for improving the use of corpus applications in a classroom.

This study provides enough data on the classroom use of the corpus to enable the view of participants as individuals within an otherwise monolithic class, in contrast to Yoon who suggested that "many corpus studies have regarded learners as a monolithic group rather than as idiosyncratic individuals" (32). This closer examination of the process for each session and the time spent on each session for the participants is rich information that is often missing in the literature. Even though many scholars (e.g., Horst et al. 107; Chambers 121; Hafner and Candlin 304) acknowledge the importance of tracking learners' use of corpus-based resources, studies that have used tracking tools either did not collect the search strings to track details of learners' interactions (Gaskell and Cobb), or did not include the tracking details of learners' interactions with the resource (Hafner and Candlin).

Another strength of using direct-observation methodology is to recognise any problematic aspects that may be encountered during the use of corpus-based resources that learners fail to report. As previously noted, with direct guidance, only one participant failed to use the wildcard during the second training session. Without direct guidance, seven participants failed to use the wildcard during the first testing session.

This study has overcome the obstacle highlighted by Pérez-Paredes et al.: "researchers should be cautious as to the extent to which actual interaction with the corpus has taken place at all" (246). This study addressed this issue by providing a richly detailed review of the interactions with the concordance use.

5.2 The Participants' Use of the DDL Approach and the Successful Completion of Tasks

The learners' overall performance improved significantly after they received proper training and gained confidence in using the corpus. This finding is in line with a study conducted by Yoon and Hirvela in which the intermediate group received more training in using the corpus than the advanced group, which led to more positive attitudes (277). O'Sullivan and Chambers reached the same conclusion, indicating that the need for training and guidance with the use of "an appropriate corpus" can lead to more active use of the corpus (65).

In the study, participants not only received training on the technical use of the corpus but were also given guidelines on how to read and analyse the concordance lines, following Sinclair's (*Reading Concordances*) model. This

enhanced their performance in the answering activities. As seen in previous research, when participants receive technical guidelines without training on how to read and analyse the concordance lines, or if they undergo a short training period, the result is either insignificant performance or less positive attitudes (e.g., Pérez-Paredes et al. 244; Gaskell and Cobb 311-17).

Even so, it can be argued that the concordancer helped the participants to recognise the meaning of verb patterns. Previous studies have reported similar results: Chan and Liou revealed that de-lexicalised verbs are more effectively learnt through web-based concordancing (241). Cobb also found that learners can acquire knowledge of new vocabulary through concordances (311). Using convergent tasks is another factor that may affect the participants' performance. Bernardini recommended that training on the use of the corpus should begin by using convergent tasks in which the guided learners can reach the same conclusions or outcomes (26); once the learners familiarise themselves with the interface, they can move on to divergent or independent tasks. In the current study, the testing sessions were designed in the same way as the training sessions to guarantee a smoother transition from the dependent to the independent conditions.

The type of guidance provided in this study is not frequently used in the literature, namely, guidance when the learners come into contact with the resource. Pérez-Paredes et al. emphasise that this type of guidance is "less frequently found" (236). Written guidelines on how to conduct analyses are also recommended when the learners have little or no previous training (Gotz and Mukherjee 59), as was provided in the testing sessions. Many researchers have supported the use of this exemplar guidance in corpus consultation through teacher-guided activities (Bernardini; Chambers; Cheng et al.; Yoon and Hirvela). Pérez-Paredes et al. offered guidance before learners engaged with the tasks, but not while performing the task (238), as was done in the present study. Based on each type of task, the participants' performance was quite good, as they were able to notice the repeated patterns and recognised the forms (form-meaning activities). In the meaning-recognition activity, they were able to assert the meaning of most of the verb patterns. Sinclair noted that corpus data can enhance language teaching and that the form-meaning link can be taught through this data to minimise the learning load ("A Way with Common Words," 160). Thus, learners can use the corpus to develop their creative use of language or "creative processing" (Nation 110), which occurs when they come across familiar target items but with different meanings in different contexts.

In the present study, the activities took language-learning theories into consideration, including the Noticing hypothesis. The inductive approach followed in the activities was entirely dependent upon the Noticing hypothesis with guidance from the teacher when the learners were asked to examine verb patterns. This type of teacher-directed noticing is called *pedagogical mediation*, first

introduced by Johns ("Should You Be Persuaded"), and subsequently used by McCarthy and Widdowson. Flowerdew stated that the majority of DDL research adopts the "guided-inductive approach" in which the teacher initiates the search, and the students start to learn by noticing (21). Boulton stated that the DDL approach goes well with the constructivist paradigm for language learning in that learners explore the data to detect patterns and they are not taught the overt rules ("Data-driven Learning: On Paper, in Practice" 35). The concordance context considers rich linguistic input, which helps when learning novel words, as was demonstrated in Cobb's study. Due to the absence of a pre-test or post-test to measure the acquisition levels, the present study cannot claim that the concordancer has helped or improved general verb acquisition. However, we can argue that through the set of activities, the participants were encouraged to engage in an in-depth analysis of general verb patterns. In other words, the participants were studying the items independently in a rich semantic, syntactic and collocational information context. The results indicate that the participants' ability to follow the same DDL processes improved during the second testing session.

6. Conclusion

In conclusion, the study is significant in that it provided learners with effective guidance and direct steps to avoid the difficulties faced in previous research, such as the lack of knowledge and skills in using corpora and concordancers, difficulties concerned with working on corpus data (Cheng et al. 182; Lavid 251), and general difficulties in using a corpus for the first time (Chambers; Chambers and O'Sullivan; Cheng et al.; Yoon; Yoon and Hirvela). It also provides tracking information for using a concordancer, which has been talked about more than "tested with real learners" (Gaskell and Cobb 317).

The tracking methodology used in this study provided data based on direct observations (textual and screenshot data) of the participants' use of the corpus, which could have increased the validity of the results. Pérez-Paredes et al. claimed the validity of their experiment because they used the direct observation methodology as opposed to using indirect methodologies (249). According to Pérez-Paredes et al., using indirect observation methodologies can result in misleading data because the data provided may be affected by differences between what learners think they did versus what they actually do (234). Another issue is that data of the indirect observation methodologies can be affected by the learners' attitudes and perceptions when they report on what they are doing by narrowing the scope of the data to what the learners think is beneficial or difficult. These issues have been confirmed in previous studies (Chapelle and Mizuno 42; Fischer 429).

This study was conducted in response to the repeatedly stressed need to fill a gap in the knowledge regarding how learners cope with using corpora in a

classroom as well as tracking learners' direct use of corpora for language learning. This study also followed Yoon's suggestion to look "at the students' individual experiences in the analysis of corpus use" (32), which is rarely done as most studies consider learners as a monolithic group. In general, corpus use in the classroom can be affected by many factors, such as the learners' language proficiency and their familiarity with the new approach. Thus, utilising direct observation methodologies are essential to evaluate how each learner copes with the use of corpora in the classroom.

Works Cited

Anthony, Lawrence. AntConc (Version 3.4.3) [Computer Software]. Waseda University, 2014. http://www.laurenceanthony.net/software/antconc/ January 2020.

Bernardini, Silvia. "Corpora in the Classroom: An Overview and Some Reflections on Future Developments." *How to Use Corpora in Language Teaching*. Ed. John Sinclair. Amsterdam: John Benjamins, 2004. 15-38.

Boulton, Alex and Pascual Pérez-Paredes. "Researching Uses of Corpora for Language Teaching and Learning." *Editorial to ReCALL* 26.2 (2014): 121-27.

Boulton, Alex. "Data-Driven Learning: Reasonable Fears and Rational Reassurance. CALL in Second Language Acquisition: New Approaches for Teaching and Testing." *Indian Journal of Applied Linguistics* 35.1 (2009): 81-106.

---. "Data-driven Learning: On Paper, in Practice." *Corpus Linguistics in Language Teaching*. Ed. Tony Harris and Maria Moreno Jaen. Bern: Peter Lang, 2010.17-52.

---. "Corpus consultation for ESP: A review of empirical research." *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Ed. Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet. Amsterdam: John Benjamins, 2012. 261-91.

Breyer, Yvonne. "Learning and Teaching with Corpora: Reflections by Student Teachers." *Computer Assisted Language Learning* 22.2 (2009): 153-72.

Chambers, Angela. "Integrating Corpus Consultation in Language Studies." *Language Learning and Technology* 9.2 (2005): 111-25.

Chambers, Angela and Ide O'Sullivan. "Corpus Consultation and Advanced Learners' Writing Skills in French." *ReCALL* 16.01 (2004): 158-72.

Chan, Tun P. and Hsien C Liou. "Effects of Web-Based Concordancing Instruction on EFL Students' Learning of Verb-Noun Collocations." *Computer Assisted Language Learning* 18.3 (2005): 231-51.

Chapelle, Carol A. and Suesue Mizuno. "Students' Strategies with Learner-Controlled CALL." *CALICO Journal* 7 (1989): 25-47.

Cheng, Winnie, Martin Warren and Xu Xun-feng. "The Language Learner as Language Researcher: Putting Corpus Linguistics on the Timetable." *System* 3 (2003): 173-86.

Cobb, Tom. "Is There any Measurable Learning from Hands-on Concordancing?" *System* 25.3 (1997): 301-15.

Cohen, Jacob W. *Statistical Power Analysis for the Behavioural Science*. 2nd ed. USA: Lawrence Erlbaum Associates, 1998.

Ellis, Rod. *The Study of Second Language Acquisition*. Oxford: Oxford UP, 2008.

Fischer, Robert. "How do We Know what Students are Actually Doing?: Monitoring Students' Behaviour in CALL." *Computer Assisted Language Learning* 20 (2007): 409-42.

Flowerdew, Lynne. "Data-driven Learning and Language Learning Theories: Whither the Twain Shall Meet." *Multiple Affordances of Language Corpora for Data-Driven Learning*. Ed. Agnieszka Leńko-Szymańska and Alex Boulton. Amsterdam: John Benjamins, 2015. 15-36.

Gaskell, Delian and Tom Cobb. "Can Learners Use Concordance Feedback for Writing Errors?" *System* 32.3 (2004): 301-19.

Geluso, Joe and Atsumi Yamaguchi . "Discovering Formulaic Language Through Data-Driven Learning: Student Attitudes and Efficacy." *ReCALL* 26.2 (2014): 225-42.

Hafner, Christoph A. and Christopher N. Candlin. "Corpus Tools as an Affordance to Learning in Professional Legal Education." *Journal of English for Academic Purposes* 6.4 (2007): 303-18.

Horst, Marlise, Tom Cobb and Ioana Nicolae. "Expanding Academic Vocabulary with an Interactive Online Database." *Language Learning and Technology* 9 (2005): 90-110.

Hughes, Rebecca. "What Can a Corpus Tell Us About Grammar Teaching Materials?" *The Routledge Handbook of Corpus Linguistics*. Ed. Anne O'Keeffe and Michael McCarthy. Abingdon, Oxon: Routledge, 2012. 401-12.

Johns, Tim and Philip King. "Classroom Concordancing." *English Language Research Journal* 4 (1991): 1-31.

Johns, Tim. "Should You be Persuaded: Two Examples of Data-driven Learning." *ELR Journal* 4 (1991): 1-16.

Kennedy, Claire and Tiziana Miceli. "An Evaluation of Intermediate Students' Approaches to Corpus Investigation." *Language Learning and Technology* 5.3 (2001): 77-90.

---. "Corpus-assisted Creative Writing: Introducing Intermediate Italian Learners to a Corpus as a Reference Resource." *Language Learning and Technology* 14.1 (2010): 28-44.

Lantolf, James and Steven Thorne, eds. *Sociocultural Theory and the Genesis of Second Language Development*. Oxford: Oxford UP, 2006.

Lavid, Julia. "Contrastive Patterns of Mental Transitivity in English and Spanish: A Student Centred Corpus-based Study." *Corpora in the Foreign Language Classroom.* Ed. Encarnacion Hidalgo Tenorio et al. Amsterdam: Rodopi, 2007. 237-52.

Leńko-Szymańska, Agnieszka and Alex Boulton. "Data-driven Learning in Language Pedagogy." *Multiple Affordances of Language Corpora for Data-driven Learning.* Ed. Agnieszka Leńko-Szymańska and Anne Boulton. Amsterdam: John Benjamins, 2015. 1-14.

McCarthy, Michael. *Spoken Language and Applied Linguistics.* Cambridge: Cambridge UP, 1998.

Nation, Paul I.S. *Learning Vocabulary in Another Language.* 2nd ed. Cambridge: Cambridge UP, 2013.

O'Sullivan, Íde and Angela Chambers. "Learners' Writing Skills in French: Corpus Consultation and Learner Evaluation." *Journal of Second Language Writing* 15.1 (2006): 49-68.

Perez-Paredes, Pascual, et al. "Tracking Learners' Actual Uses of Corpora: Guided vs Non-guided Corpus Consultation." *Computer Assisted Language Learning* 24.3 (2011): 233-53.

Schmidt, Richard. "The Role of Consciousness in Second Language Learning." *Applied Linguistics* 11 (1990): 129-58.

---. "Attention, Awareness and Individual Differences in Language Learning." *Perspectives on Individual Characteristics and Foreign Language Education.* Ed. Chan Wai Meng et al. Germany: Hubert and Co., 2012. 27-50.

Sinclair, John. "A Way with Common Words." *Out of Corpora: Studies in Honour of Stig Johansson.* Ed. Hilde Hasselgard and Signe Oksefjell. Amsterdam: Rodopi, 1999. 157-79.

---. *Reading Concordances: An Introduction.* London: Longman, 2003.

Smart, Jonathan. "The Role of Guided Induction in Paper-based Data-driven Learning." *ReCALL* 26.2 (2014): 184-201.

Varley, Steve. "I'll Just Look That Up in the Concordancer: Integrating Corpus Consultation into the Language Learning Environment." *Computer Assisted Language Learning* 22.2 (2009): 133-52.

Villamil, Olga and Maria De Guerrero. "Socio-cultural Theory: A Framework for Understanding the Socio-cognitive Dimensions of Peer Feedback." *Feedback in Second Language Writing: Contexts and Issues.* Ed. Ken Hyland and Fiona Hyland. Cambridge: Cambridge UP, 2006. 23-41.

Vyatkina, Nina. "Data-driven Learning for Beginners: The Case of German Verb-Preposition Collocations." *ReCALL* 28.2 (2016): 207-26.

---. "Data-driven Learning of Collocations: Learner Performance, Proficiency and Perceptions." *Language Learning and Technology* 20.3 (2016): 159-79.

Widdowson, Henry G. "On the Limitations of Linguistics Applied." *Applied Linguistics* 21.1 (2000): 3-25.

Yoon, Hyunsook. "More Than a Linguistic Reference: The Influence of Corpus Technology on L2 Academic Writing." *Language Learning and Technology* 12 (2008): 31-49.

Yoon, Hyunsook and Alan Hirvela. "ESL Student Attitudes Toward Corpus Use in L2 Writing." *Journal of Second Language Writing* 13.4 (2004): 257-83.