Using Data-Driven Learning for Second Language Academic Writing Acquisition

A.J. Holmberg

Northern Arizona University

Article Info

Keywords:
Keyword 1; data-driven learning
Keyword 2; Corpus
Keyword 3; L2 Writing

Article History:
Received : 26/09/2023
Revised : 25/11/2023
Accepted : 28/11/2023
Available Online: 30/01/2024

Abstract

Corpus linguistics is a rapidly growing subfield in linguistics. Correspondingly there is more corpus data available than ever before. Existing corpora can be used to assist teachers in helping their students learn how language is used in a natural setting. While there are several studies on data-driven learning in the field, very few studies have reviewed the overall effectiveness of this teaching method concerning different language contexts and structures. To fill this gap in the literature, this synthesis paper aimed to describe the different ways corpus data is used for data-driven learning in a second-language academic setting. Additionally, this paper evaluated the effectiveness of using data-driven learning techniques for various English language contexts. Using the Linguistics and Language Behavior Abstracts (LLBA) database with the keywords: 'data-driven learning,' 'corpus,' and 'L2 writing', this synthesis paper identified eight empirical studies that examined the use of data-driven learning in the classroom and its effectiveness for second language learning. Two more studies were found using the reference lists of the original eight empirical studies.

Introduction

It has been a little over three decades since Johns (1991) advocated for the use of machines as an instrument for language learning. Exposing students to authentic language data and helping them to truly 'discover' a language was the basis of data-driven learning (DDL). A true definition of DDL has not been concretely established, and it means different things to different people in the linguistics and language teaching field. In essence, all descriptions of DDL involve using corpus data and tools for pedagogical reasons and purposes (Boulton & Tyne, 2015). Boulton and Tyne (2015) also mention that one of the most significant alleged advantages of using DDL techniques is that learners will become more sensitive to authentic language uses than they would in a conventional language learning environment.

Since the coining of the term DDL, the field of corpus linguistics has significantly expanded, and the amount of authentic language data available to teachers and students has grown exponentially. Methods of using corpus data for L1 and L2 writing have been discussed at great length (Flowerdew, 2010). Uses of corpus data have ranged from being used simply as a tool for teachers to plan what to teach in their classrooms to providing students with the knowledge to use corpora for themselves to correct their errors (Crosthwaite, 2017). There has also been a significant amount of research into how to use DDL in a language classroom, with specific attention paid to collocation tools and concordancing (Yoon, 2011), which documented how to teach students to use these corpus tools to learn linguistic structures or correct their errors. This is an important trend because corpus data for DDL purposes is only helpful if teachers know how to use DDL and what structures and contexts it is most effective for in an L2 academic writing environment.

This synthesis paper aims to describe how DDL is used in a second language (L2) writing context and to evaluate the effectiveness of these methods and teaching techniques. There have been many meta-analyses and reviews of DDL published in previous decades, but in an approach using technology, so much can change quickly, making this synthesis paper necessary to keep up to date with effective DDL teaching methods and activities. The results of this research will influence the discussion on how DDL can be used in an academic L2 writing environment and how it can be used effectively. Therefore, this paper seeks to answer the following research questions: (1) What are the ways DDL is used in academic L2 writing? and (2) How effective are the different uses of DDL in teaching academic L2 writing?

Method

The Linguistics and Language Behavior Abstracts (LLBA) database was used to find peer-reviewed articles on DDL methods and usage. The search terms ‘L2 writing’, ‘corpus,’ and ‘data-driven learning’ were used to locate relevant studies, yielding 13 initial results. This pool of studies was then screened using the following inclusion criteria:

1) The paper involves empirical research with quantitative data

2) The paper includes the usage of non-learner corpora to facilitate learner growth

Eight studies fit the inclusion criteria and made up the initial synthesis pool. Two more relevant studies that match the inclusion criteria were found in the reference sections of the

original eight studies. A total of 10 studies were included in this synthesis paper. Despite the small number, the breadth of these 10 studies covers a wide range of uses for DDL.

Findings and Discussion

Usages of Data-Driven Learning

DDL is a broad approach to foreign language learning used in language classrooms in various ways. For this synthesis paper, DDL is considered to be the usage of corpus data or tools for pedagogical purposes (Boulton & Tyne, 2015). This broad definition is used in this paper because a more precise and comprehensive description of DDL has yet to be widely agreed upon. Due to this, it is necessary to review how DDL is used in a language classroom before its effectiveness can be considered. In this synthesis paper, the 10 empirical studies identified in the literature search process show three general ways language instructors use DDL in academic classrooms: general collocations, colligations, and revision tasks with coded error feedback.

General collocations (e.g., important, beautiful) are one of the most common ways DDL is used in a language learning classroom based on the studies found in this paper. Larsen-Walker's (2017) study used DDL to improve students' usage of linking adverbials (LA) (e.g., then, also). It involved 24 advanced English for Academic Purposes (EAP) students, 12 in a control group and 12 in a treatment group. The control group was taught LAs in a traditional present, practice, and production (PPP) instructional sequence. The treatment group was introduced to the MICUSP corpus and taught to find LAs in context using the keyword search (KWIC) function. The instructor used corpus data to model inductive reasoning and allowed students to implicitly learn the LA structures while being there to help scaffold during the process.

Yeh et al. (2007) used corpus data and collocation tools to teach 19 first-year college students in Taiwan how to find and use synonyms for common adjectives like 'big,' 'important,' 'hard,' and 'beautiful.' The study taught the students how to read concordance lines so that students were able to explore and recognize patterns in the adjective use of semantically similar words.

On the other hand, Koosha and Jafarpour (2006) used similar DDL methods to aid students in learning collocations of prepositions. The study had 200 Iranian participants, with a control group being taught using traditional grammar books and a treatment group following a DDL lesson structure. A pretest was given to the participants to determine their collocational knowledge of prepositions. Following the pretest, participants were given 30

hours of instruction on collocations of prepositions using Brown Corpus data. The lessons were based on concordance lines focusing on the keywords in context. Students were then given a posttest to assess their knowledge of collocations.

In a final study, Chan and Liou (2005) employed a DDL approach to teaching verb-noun collocations to EFL students. The study consisted of 32 EFL students being taught how to use a Chinese-English concordancer to find and analyze English verb-noun collocations over five web-based units. The units covered gap-fill exercises and how to read concordance lines from corpus data. The DDL method was evaluated in two ways, a qualitative survey of how the students perceived the usage of corpus data and a quantitative analysis of the students' improvement from the pre- to the posttest.

All four of these studies (Chan & Liou, 2005; Koosha & Jafarpour, 2006; Larsen-Walker, 2017; Yeh et al., 2007) used DDL in the same manner, which was to use corpora data to help students learn various collocations. The amount of DDL instruction time varied dramatically across the studies, but they all included some form of instruction based on concordance lines from corpus data. Instruction tended to involve some early gap-fill exercises so students' attention could be drawn to target collocations. All the studies followed one of the central tenets of DDL, which is to allow students to learn inductively by noticing patterns in the corpus data. This emphasizes students recognizing language usage patterns and not relying on rote memorization like other language teaching methods. Not all of these studies had a control group, but they all included a pre-and posttest in determining whether DDL proved effective. Using DDL to help students learn collocations is a common approach, but it is not the only way to employ DDL in a language classroom.

Another primary usage of DDL in an L2 writing classroom was using corpus data to help with colligations, a type of collocation based on specific grammatical patterns (nouns collocating with other nouns, adverbs collocating with verbs, etc.). Three studies gathered for this synthesis paper used these colligations as the basis of their DDL instruction.

Yilmaz (2017) created a corpus to teach 30 Turkish EFL learners to properly use abstract nouns in different forms. A pretest was used to test students' initial knowledge of how to use a list of 10 abstract nouns. The pretests were then marked for errors without explanation. Participants were then split into a control group that used dictionaries to learn about abstract noun colligations and a treatment group that received corpus instruction similar to the studies described above. After instruction, both groups took a posttest on the list of 10 abstract nouns. In another study, Huang (2014) followed nearly the same

procedures with 40 Chinese university students majoring in English. This study also examined abstract nouns (albeit only a list of five abstract nouns) using pre, immediate post, and delayed posttests. Huang (2014) also created a topic-specific corpus to match participants' proficiency levels. The control group used dictionaries, while the treatment group used corpus materials. The main difference between the two studies was that Huang (2014) did not utilize electronic corpus data and instead opted for paper-based DDL usage.

The third study (Yunus & Awab, 2014) differed from the other two regarding the linguistic feature being taught. Yunus and Awab (2014) also used paper-based DDL corpus methods but used them to teach colligational patterns of prepositions to 20 Malaysian law students. The study had a control group using dictionaries and a treatment group using printouts of concordance lines showing the colligational patterns involving prepositions.

All three studies followed very similar procedures in using DDL in an L2 writing situation, the only difference being the linguistic features taught and participants' L1. This DDL usage of teaching colligations is similar to implementing the general collocation usage discussed previously. These two usages could be considered the same if not for the focus on the grammatical versus non-grammatical functions being taught.

In three other studies, Tono et al. (2014) and Crosthwaite (2017, 2020) used DDL to help learners with revision tasks with coded error feedback. Tono et al. (2014) examined two writing samples from 93 undergraduate students attending two universities in Tokyo. The first writing sample was a 15-minute essay with no dictionaries or revisions. The researchers gave coded feedback focusing on three main types of errors: grammatical form errors, word omission errors, and word addition errors. Three weeks later, students received instructions on using the keyword-in-context (KWIC) tool for the British National Corpus (BNC). The students were asked to revise their errors from the first writing sample using corpus data and tools.

Similarly, Crosthwaite (2017) examined 32 students from various areas in China. The study involved having students submit a 400–600-word sample of their work and then having the teacher provide feedback on the samples, which came in the form of highlights in MS Word. The feedback was coded into different types of errors, including lexical errors, phrasing errors, grammatical errors, and errors of word omission/addition. The participants were then taught how to use corpora to correct errors in their writing. This included teaching how to find collocations in the BNC. Participants were then tasked with correcting their

original samples using the BNC corpus to help with their error corrections. The third study, by Crosthwaite (2020), followed a similar pattern but was held as an online program instead.

All these studies used DDL to help students correct errors in their written work. Each study coded students' errors and then used the BNC as data for error correction. One of the critical features of DDL used in these studies was the focus on implicit learning with corpus tools, allowing the students to find language patterns that would help them correct their errors on their own. These studies show that revision tasks with coded error feedback are an established way DDL is currently used in language classrooms. Crosthwaite's (2020) study went a step further by using DDL for understanding registers and genres, which other studies in this synthesis paper did not include.

Effectiveness of Data-Driven Learning Techniques

As the previous section described, it has been established that various DDL techniques can be used in a language classroom. Still, there is no reason to implement these techniques if they are less effective than traditional L2 writing teaching techniques. Now that the uses of DDL have been established, it is essential to delve into the effectiveness of these techniques and their ability to promote L2 writing acquisition skills. There is also importance in distinguishing the effectiveness of DDL across students of different levels of language proficiency.

Four studies used corpus data and DDL techniques to teach certain collocations (Chan & Liou, 2005; Koosha & Jafarpour, 2006; Larsen-Walker, 2017; Yeh et al., 2007). These studies reported that when DDL was used to help students learn collocations. Participants were able to use them more accurately and frequently than conventional instruction. This remained consistent regardless of collocation types and participants' L1s.

Despite the differences in sample size, L1, and linguistic feature focus, these studies revealed that DDL was more effective than conventional instruction. The main variation in the results comes from how much more effective these DDL tactics were compared with other forms of instruction. Larsen-Walker's (2017) results showed that the treatment group used LAs correctly at a higher rate (91.4%) than the control group (87.7%), although the difference was not statistically significant. These results indicated DDL being slightly more effective but likely not effective enough to justify the time spent on corpus usage instruction. The other studies showed much more correct usage and retention improvement when using DDL corpus collocation techniques. Yeh et al. (2007) and Chan and Liou (2005) saw a nearly 200% increase from the pre- to the posttest. These findings indicate that the participants were

twice as successful in using synonyms and verb-noun collocations when they learned them using DDL. These improvements remained steady even after a delayed posttest.

These studies show that using DDL methods and corpora for teaching various kinds of collocations can be more effective than conventional methods in terms of advanced, university-level language students. However, it is important to note that some collocations appear more conducive to DDL methods than others. Still, all areas did report an increase in correct usage after DDL, which is important to note even if they are not all statistically significant. Another thing that varied significantly among the studies was the number of hours of instruction the treatment groups received in using corpus tools (e.g., concordancing) and finding collocations in the data. The studies had a range of instruction time spanning from one-hour to 15 hours. Critically, the number of instructional hours did not impact learning outcomes in these studies.

In terms of using DDL for revision tasks, the researchers (Crosthwaite, 2017, 2020; Tono et al., 2014) concluded that using corpora for error correction was more suitable for certain error types than others. Crosthwaite (2017) determined that students used corpora to revise word choice, collocation, and phrasing errors accurately, but they were unlikely to use corpus data for grammatical errors. These results are corroborated by Tono et al. (2014), which showed that collocation additions and omissions were commonly corrected with corpus data, but grammatical errors were rarely corrected effectively. The researchers suggested this because it is difficult for learners to decide which words to search using corpus tools to correct the error. In that way, Tono et al. (2014) conceded that using DDL for coded feedback on grammatical mistakes has significant limitations. While these two studies used quantitative data to evaluate the effectiveness of using DDL for error correction, Crosthwaite (2020) added some qualitative data for the same type of DDL instruction. The 6-point Likert scale (6=strongly agree, 1=strongly disagree) post-course questionnaire that the researchers used indicated that students agreed that using corpus tools helped improve their writing skills (5.0), vocabulary (5.1), phrases (5.0), and grammar (4.4). It is important to note that participant evaluations match the results of the other two studies in the usage of vocabulary and phrases. However, participant views on using corpus tools for correcting grammatical errors did not match the data from the other two studies, which showed that participants often did not use corpus tools to correct grammatical mistakes accurately. Another critical observation is that many participants responded negatively about the time needed to learn how to use corpora tools and how difficult corpora can be to use. Despite these difficulties,

many participants were positive about autonomously using corpus tools to correct their writing in the future. This qualitative data shows that students share many of the positives of DDL instruction that the quantitative data provided, but also that participant views may also be overconfident in using corpora to correct grammar mistakes.

In terms of collocations and word choice, students that used corpus tools saw quite a significant success rate in both studies. Crosthwaite's (2017) analysis was more comprehensive in correcting errors by looking at seven different error types compared to the three in Tono et al. (2014). In Crosthwaite's (2017) results, participants had an average 82.3% success rate across all error types, even the grammatical errors that were less likely to be corrected with corpus tools. Even though the purpose of the studies was not primarily to compare corpus users' error corrections with non-corpus users' errors, there were still essential takeaways from this comparison in the studies. For the error types that saw high corpus usage in correction, like word choice and collocation errors, the corpus users corrected these errors at 1.7 and 2.5 times the rate of their non-corpus user counterparts. Thus, showing that using DDL techniques in error correction can be significantly more effective than conventional instruction in some areas. However, grammatical errors were corrected much more successfully by the non-corpus users in comparison to their counterparts (Crosthwaite, 2017). Across the three studies, it can be seen that grammatical errors are a struggle for students to correct using corpus data but using DDL is very effective for other types of errors like collocation and word choice errors. Ultimately, the results from these three studies show that using DDL for error correction is effective, but only for certain types of errors, and that student opinions on DDL instruction and corpus usage generally match the assessment data.

This synthesis paper examined the different ways that DDL is used in an L2 writing environment and the effectiveness of those uses of DDL. Of the 10 empirical studies that were synthesized, all of them were in a university setting with participants ranging from a lower-intermediate level to advanced language proficiency. Koosha and Jafarpour (2006) also had an additional group of slightly lower language proficiency participants. There were three main ways that DDL was employed in the classroom. Four studies used DDL to teach general collocations (Chan & Liou, 2005; Koosha & Jafarpour, 2006; Larsen-Walker, 2017; Yeh et al., 2007). All of these studies, to varying levels, found that DDL methods to teach the collocations of a variety of parts of speech were more effective than conventional L2 writing instruction.

Besides general collocations, other studies used DDL to teach colligations (Huang, 2014; Kamariah & Awab, 2014; Yilmaz, 2017). The usage of DDL and the results of these studies coincided with those of the general collocation studies: DDL was, to varying degrees, more successful in immediate learning and longer-term retention of the target linguistic features than traditional instruction.

Other than collocations and colligations, several studies used DDL for revision tasks using coded error feedback (Crosthwaite, 2017, 2020; Tono et al., 2014). The outcomes of these studies showed that DDL was effective in teaching students to correct their errors in most areas besides grammatical-type errors. Additionally, students perceived DDL as more helpful than traditional instruction in helping with all types of errors. Still, participants did comment on the difficulty of implementation and learning how to use corpora and corpus tools.

From the findings in this synthesis paper, a few important conclusions should be pointed out. Overall, it is apparent that DDL provides plenty of advantages over traditional instruction in the form of error correction and collocation knowledge. However, it is important to note a few limitations of DDL from a practical implementation standpoint. Incorporating corpora use in a classroom requires significant resources, time, and student cooperation to be used effectively. Some of these concerns can be counteracted by using paper-based DDL, which does not require technology access or technologically savvy students. However, the limitation of time cannot be sufficiently counteracted. A large amount of time is required for various steps in the DDL implementation: instructors must become familiar with using corpora and corpus tools; instructors must spend time finding a corpus that will suit their students' proficiency and purpose; class time must be spent instructing students in using corpus data and tools (to varying degrees of success); and students must be given adequate time to familiarize themselves with the data in order to learn inductively

Conclusions

Initially, the implementation of DDL techniques seems like an obvious choice because of the advantages over traditional instruction based on the findings of the studies in this paper. Still, the limitations of DDL make the overall value of DDL implementation more nuanced. DDL is not a complete solution to L2 writing instruction because, while it is more effective than a traditional PPP approach in many areas (collocations, certain revision tasks, etc.), it lags in effectiveness in other areas (correcting grammatical errors). It also takes

specific instructor knowledge, student motivation, technology knowledge, and vast amounts of time for material development, corpus instruction, and its implicit learning focus. These factors could be why most studies are conducted with higher-level university-aged participants. In order to make proper recommendations for DDL usage, more research is needed with participants of varying age ranges, proficiency levels, and initial general technology levels.

This synthesis paper finds the usage of DDL beneficial over conventional L2 writing instruction in many areas, like collocations and certain types of revision tasks. However, DDL was less effective in helping L2 learners correct grammatical errors in revision tasks (Crosthwaite, 2017, 2020; Tono et al., 2014). Through this paper, it can be seen that DDL is a positive tool for the L2 writing environment. However, it has numerous limitations: lower effectiveness on grammar error revisions, corpus instruction time requirements, and instructor knowledge of corpora and corpus tools. Still, DDL can be used as an addition to an established L2 writing curriculum to help provide students with more tools to improve their writing. As technology continues to march forward, DDL, with more research, could become increasingly prevalent in L2 writing contexts as its limitations are possibly resolved, and corpus tools continue to strengthen its benefits. The studies discussed in this paper mainly worked with university students. Still, as corpora become more accessible, it will be essential to observe how DDL can be applied to contexts other than university students.

REFERENCES

Boulton, A., & Tyne, H. (2013). Corpus linguistics and data-driven learning: A critical overview. *Bulletin Suisse de Linguistique Appliquée*, *97*, 97–118.

Chan, T., & Liou, H. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning, 18*(3), 231–250.

Crosthwaite, P. (2020). Taking DDL online. *Australian Review of Applied Linguistics, 43*(2), 169–195.

Crosthwaite, P. (2017). Retesting the limits of data-driven learning: Feedback and error correction. *Computer-Assisted Language Learning, 30*(6), 447–473.

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 444–457). Routledge.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL, 26*(2), 163–183.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), Classroom Concordancing. *English Language Research Journal* (pp. 1-16). Sciedu Press.

Kamariah Y., & Awab, S. (2014). The impact of data-driven learning instruction on Malaysian law undergraduates' colligational competence. *Kajian Malaysia, 32(*1), 79–109.

Koosha, M., & Jafarpour, A. A. (2006). Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL Journal, 8*(4), 192-209.

Larsen-Walker, M. (2017). Can Data Driven Learning address L2 writers' habitual errors with English linking adverbials? *System, 69*, 26-37.

Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL, 26*, 147–162.

Yeh, Y., Liou, H., & Li, Y. (2007). Online synonym materials and concordancing for EFL college writing. *CALL, 20*(2), 131-152.

Yilmaz, M. (2017). The effect of data-driven learning on efl students' acquisition of lexico-grammatical patterns in efl writing. *Eurasian Journal of Applied Linguistics, 3*(2), 75-88.

Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes, 10*(3), 130–139.