International Journal of English Language & Translation Studies
ISSN: 2308-5460

Using Data Driven Learning to Reduce Collocational Errors of Iranian Intermediate EFL Learners in Paragraph Writing

[PP: 111-117]

Sa'eed Esmailnia
English Language Department, Chabahar *Maritime* University
Iran
Hoshang Khoshsima
(Corresponding Author)
English Language Department, Chabahar Maritime University
Iran
Yousef Bakhshizadeh
English Language Department, Chabahar Maritime University
Iran
Hadi Hamidi
Department of English Language, School of Health Management and Information Sciences
Iran

ABSTRACT

Data Driven Learning (DDL) is a significant topic in the field of second or foreign language acquisition, which is in need of more rigorous research. In this regard, the present study was an attempt to investigate how DDL affects students' ability to use collocations correctly within their paragraph writing. The participants were Iranian intermediate EFL students at a private English language institute across educational levels and fields of study. Fifty-two participants, under the instructor's guidance, had access to a website with links to the Wikipedia Corpus Site (WCS) - a list of keywords, collocations, and instructional materials. After finding key words in the context in Wiki-Corpus, students worked on keyword combinations. Students' pretest paragraphs produced ample documentation of misuse of collocations, echoing previous research. Pretest and posttest paragraphs provided data for measuring improvement in accuracy of students' use of collocations in writing. Posttest paragraph results indicated a decrease in errors and better use of collocations in the treatment group. Comparing the treatment and control groups, the researcher found that there were significantly fewer errors in the use of collocations for the experimental group compared to that of the control group in the posttest. The findings proved that DDL had a positive mediating role in decreasing collocational errors in students' writing.

Keywords: *Data Driven Learning, Collocation, Paragraph Writing, Collocational Errors, EFL*

ARTICLE INFO	The paper received on	Reviewed on	Accepted after revisions on
	25/10/2019	**26/11/2019**	**20/01/2020**

Suggested citation:
Cite this article as: Esmailnia, S., Khoshsima, H., Bakhshizadeh, Y. & Hamidi, H. (2019). Using Data Driven Learning to Reduce Collocational Errors of Iranian Intermediate EFL Learners in Paragraph Writing. *International Journal of English Language & Translation Studies*. *7(4). 111-117.*

1. Introduction

Although electronic technology has proved to be a useful tool in motivating the students and helping them in the process of language learning, teachers and students still hesitate to use it in classroom setting (Maftoon, Hamidi, & Sarem, 2012). Despite its slow adoption in academic settings, e-learning provides several benefits to individuals and organizations (Abdekhoda, Dehnad, Javad, Mirsaeed, & Gavgani, 2016). English as a Foreign Language (EFL) students can benefit from technology enhanced language learning, especially when working on the writing skill. One way to enhance students' writing is to encourage them to use language corpora in which they can access word combinations.

It was about three decades' ago that Johns (1991) with his semantic work of integrating corpus consultation into the classroom introduced an approach named, Data Driven Learning (DDL) of having students interact with one or more corpora to ameliorate their second language writing or other related skills. Since then, other scholars have used DDL in different segments of language teaching (Mishan, 2004; Smart, 2014). Learners develop a mental blueprint for the structure of a language through repeated exposures to it. From the flood of input, learners acquire a mental reservoir of meaningful linguistic

signs, such as words and constructions, and then crystallize this into specific syntactic patterns (Ellis, 2006). It is necessary to have students pay attention to the correct forms and to how those correct forms are different from what they write. DDL, involving the direct or indirect application of corpus technology, has received considerable attention among researchers and teachers over the past two decades.

DDL is inductive and discovery-based, or teacher directed and deductive. An inductive approach positions the instructor as a guide who first acquaints students with the materials and provides the target forms and then monitors students' use (Larsen-Walker, 2017). Celik (2011) and Chang (2014) showed that DDL can improved discipline-specific academic writing. Also Kosha and Jafarpor (2006) used DDL in their research to help Iranian students to use prepositions and articles correctly. They investigated the effectiveness of DDL in improving the use of prepositions in the writing of Iranian learners of English.

DDL has its own advantages in the following aspects: Firstly, DDL is based on naturally occurring language in corpora, which can provide authentic input for learners. Secondly, DDL promotes learners' active involvement in the learning process, which usually requires learners to discover or explore language rules by themselves based on their observation and analysis of the concordance output. Thirdly, unlike the rule-based language learning tending to separate grammar and lexis, DDL fosters a more lexicon-grammatical approach by allowing students to use a concordance to retrieve frequently occurred lexical or grammatical patterns for a search item (Flowerdew, 2015). Due to the above advantages, DDL is suggested as effective in promoting foreign language or second language (L2) learning. The earlier studies about DDL confirmed the positive sides of DDL in various aspects of language learning, such as promoting learner autonomy, increasing language awareness, enhancing noticing skills, extending learners' cognitive abilities, etc. Therefore, the researcher intended to apply DDL in writing, and the aim of this research was to fundamentally revolve around DDL in order to enhance students' ability in second language learning and remove habitual collocational errors while writing paragraphs. Hence, the present study attempted to answer the following research question:

RQ: Does data driven learning (DDL) have any significant role in reducing collocational errors by using collocations in writing tasks?

H0: Data driven learning (DDL) does not have any significant role in reducing collocational errors by using collocations in writing tasks.

2. Review of the Related Literature

2.1 Data-Driven Learning and Collocation

Collocation is considered the "co-occurrence of words at a certain distance, and a distinction is usually made between co-occurrences that are frequent (or more precisely, more frequent than could be expected if words combined randomly in a language) and those that are not" (Herbst, 1996, p.380). Consequently, it is essential to make students aware of chunks, giving students opportunities to identify, organize and record these. Hill (1999, as cited in Richard & Rodgers, 2003) explains that most learners with "good vocabulary knowledge" have problems with fluency because their "collocational competence" is very limited, and that, especially from intermediate level, we should aim at increasing their collocational competence with the vocabulary they have already got. Data-driven learning (DDL) highlights learning from a great quantity of linguistic resources or language examples (Schmitt, 2002). DDL setting gives contextualization for the target language to be acquired, so that learners are encouraged to work as linguistic researchers, hypothesizing and testing lexical or grammatical usage patterns (Johns, 1991). DDL has received much attention over the past few years owing to the prevalence of electronic corpora. Proponents of corpus-driven language pedagogy suggest that a key advantage to this approach is the genuine nature of native speaker corpus data in contrast to "concocted" textbook examples (McCarthy, 1988).

2.2 Empirical Studies

In recent years, a number of studies have been conducted on DDL and its role in second language teaching. The aim of these studies has been to find how DDL can be used to improve the quality of language teaching. In one study conducted by Hadley (2002) and the title "An Introduction to Data-Driven Learning" showed rationale for allowing DDL more prominence in the EFL classroom. The results of this study will encourage others to experiment with data-

driven learning in their classrooms, either as a main emphasis or alongside a standard classroom text. In 2007, a study conducted by Viola, Graf, Kinshuk, and Leo argued that learning styles are incorporated more and more in e-education, mostly in order to provide adaptively with considering the learning styles of students. Results showed the effectiveness of data-driven methods for patterns extraction even when unexpected dependencies were found.

Boulton (2009) argues that the potential for corpora in language learning has absorbed a significant amount of attention in recent years, including in the form of data-driven learning (DDL). This paper describes a simple experiment to see how lower-level learners cope with corpus data with no prior training. The language focus here is on linking adverbials in English, which are renowned to be difficult to teach using traditional methods. In 2012, an article was written on the usage of DDL by Chia, Wangb, Houc, Jin. Their article presented a data-driven optimal terminal iterative learning control (TILC) approach for linear and nonlinear discrete-time systems. Maccio and Cervellera (2012) also used DDL technique in control policies in complex system in automatic controller. He presented an approach based on local learning, relying on Nadaraya–Watson models (NWMs), introduced for the problem of deriving an automatic controller able to exploit data collected during the operation of some complex plant or system by a reference teacher (e.g., a human operator). The result of their study showed, DDL can be efficient even to teach varieties of movements, applications, no matter how much they are complicated.

One research conducted by Talai and Fotovatnia (2012) and the title "Data-driven Learning: A Student-centered Technique for Language Learning" showed that the use of student-centered methods in language teaching can be an alternative to be exerted by teachers in language classrooms. Their research showed that, such methods can create an atmosphere of excitement for the students because they themselves are supposed to discover word meanings, grammatical patterns, parts of speech and other aspects of language through receiving small tips from the teacher. In one study Lin and Lee (2015) aims to investigate the experience of six early-career teachers who team-taught grammar to EFL college students using data-driven learning (DDL) for the first time. The results show that the teachers found DDL an innovative and interesting approach to teaching grammar, approved of DDL's capacity to provide more incentives for students to engage in discussion, and endorsed its effectiveness in transforming relatively passive students into active learners.

Another research conducted by Larsen-Walker (2017) reports that L2 writing studies emphasize the importance of cohesiveness to fluent academic writing, but many writers are inclined to over-use linking adverbials (LAs), including both subordinating conjunctions (because) and transition words (however), which reduces the cohesion and readability of their texts. Her studies explore how Data Driven Learning (DDL) affects students' ability to use LAs correctly within their persuasive paragraphs. Posttest paragraph results indicate a change in correct use of LAs by the treatment group, from pretest (87.7% correct) to posttest (91.4%). Comparing the treatment and control groups, the mean percentage of correctly used LAs in the posttest paragraphs was 88% for the comparison group and 91% for the treatment group.

3. Methodology

3.1 Design of the Study

This research used a quasi-experimental design with one control and one experimental group. Both experimental and control groups received a pretest and a posttest, but the experimental group received the treatment while the control group did not.

3.2 Participants

In this research, the initial participants were selected based on convenience sampling from Rahamouz Shokouh Institute in Qa'emshahr, a city located in the north of Iran. They were female students at the intermediate level with the age range of 17 to 20 years. Out of seventy-three students who were given the language proficiency test, 52 were considered homogenized members based on the Nelson proficiency test. The homogenized members were divided into two groups: one as the control group and the other one as the experimental group (N= 26).

3.3 Instrument and Materials

The Nelson language proficiency test was administered to have homogenized participants. The Cronbach's Alpha reliability of the test was found to be .75. The test scores were statistically computed to select homogenized participants to be assigned into two groups of experimental

and control. As this study was quasi-experimental, the researcher was required to give a pretest at the beginning and a post-test at the end of the 10th session. Both the pre-test and post-test were sentences participants had to write based on the lists of collocations given to them by the researcher.

As for the materials, some texts on the internet were used by the participants in the experimental group as their sources of knowledge and concordance. The data source in this study were the Wiki-Corpus. This corpus encompasses around 1.9 billion words in 4.4 million web pages. The pages were downloaded from Wikipedia, and because of the open license, there are essentially no copyright restrictions for this site (or any other version of the corpus). This site had a wide variety of corpus of authentic language samples of any kind. The reason this corpus data based was used is that it searches for WICs, and all the corpuses are presented in a form of separate sentences in all language syntactic forms, which makes it easier for the participants to trace the directed sentences.

3.4 Procedure

Firstly, the researcher used the Nelson proficiency test in order to have homogenized participants. It should be noted that this test had been piloted at Shokoh language institute in Qa'emsher with 25 intermediate students. The obtained Cronbach's Alpha index showed that the test enjoyed a rather high reliability (r = .75). Next, the test was given to 73 English language learners in order to form two homogenous groups, one as an experimental and the other as a control group. The obtained scores of the participants were calculated using the SPSS (version 22) software and their means and standard deviations were reported. Out of 73 participants in the language proficiency test, 52 were selected based on their scores falling -1/+1 standard deviation around the mean. Finally, they were assigned to one experimental and one control groups. After that, a list of 30 collocations were given to both groups. The list was divided into 6 sets in which there were 5 collocations. The participants in both groups were asked to write 6 paragraphs using each set. The topics of the paragraphs were given to them by the researcher. Eventually, all the written paragraphs were corrected by the researcher to identify the errors made by the participants on the usage of given collocations in the sentences.

Totally, totally 10 sessions of treatment were administered to both groups. However, for the experimental group, DDL was implemented under the guidance of the researcher, and the participants were supposed to use the corpora in the Wikipedia corpus site to internalize the correct application of the given collocations. The control group did not have access to any corpora, neither online nor in print, and they only worked on what the teacher listed in their class as new collocations.

After the treatment sessions, the students were asked to write six passages again for a set of 5 collocations in the lists already given to them. The writings were then corrected and errors were taken into account and compared to the errors in the pre-test. The obtained data were statistically calculated using the SPSS software to find if there was a significant difference between the two test administrations.

4. Data Analysis and Results

4.1 Descriptive Statistics

As it was mentioned earlier, in order to have homogenous participants, the Nelson proficiency test was administered.

Table 1: *Descriptive Statistics of the NELSON Test*

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
NELSON	73	68.00	100.00	82.1644	7.58290	57.500
Valid N (listwise)	73					

As it can be seen in Table 1 above, the mean and standard deviation of the participants are 82.16 and 7.58 respectively. Therefore, the researcher chose those who scored -/+ SD below and above the mean as the homogenized members. The next table shows the result of the descriptive statistics of the homogenized members.

Table 2: Descriptive Statistics of the Homogenized Participants

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Homogenized	52	74.00	90.00	81.3462	4.27893	18.309
Valid N (listwise)	52					

Table 2 above shows that 52 participants who scored between 74 and 90 (rounded up) were chosen as the homogenized members. Next, there were assigned to a control and experimental group based on odd and even numbers (N= 26).

4.2 Answering the Research Question

The research question of the study investigated whether data driven learning (DDL) could have any significant role in reducing collocation errors in writing tasks.

In order to answer the research question, the researcher ran the ANCOVA test. The following table shows the descriptive statistics for the collocation scores of the two groups.

Table 3: The Descriptive Statistics for the Collocation Scores of the Two Groups after Adjustment

Group	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
control	7.214[a]	.347	6.519	7.908
experimental	4.486[a]	.347	3.792	5.181

As it can be seen in Table 3 above, the mean for the control and experimental groups related to their collocation scores are 7.21 and 4.48 respectively. Table 4 below shows the result of the ANCOVA test.

Table 4: *The Result of the ANCOVA for the Comparison of the Collocation Scores*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	431.120[a]	2	215.560	65.172	.000	.696
Intercept	5.740	1	5.740	1.735	.193	.030
Pre-scores	170.704	1	170.704	51.611	.000	.475
Group	**94.414**	**1**	**94.414**	**28.545**	**.000**	**.334**
Error	188.530	57	3.308			
Total	2673.000	60				
Corrected Total	619.650	59				

As Table 4 above shows, there was a statistically significant difference between the control and the experimental groups regarding their *collocation* scores, F *(1,57)* = 28.54, $p < .05$, partial $\eta^2 = .33$. As a result, the null hypothesis was rejected, confirming that the observed errors were significantly lower in the posttest of the experimental group compared to that of the control group.

5. Discussion and Conclusion

There have been some research studies similar to the present one. The findings of Braun (2007) are in line with the result of the present study showing that there is a significant relationship between DDL and English language learning in general. Boulton's (2010, 2012) findings also support the result of the present study where DDL proved to have significant effects, and was more favorable to teaching L2 activities. The corpus and concordance interface were originally perceived by and for linguists, so other users need to adopt the role of language researchers to make the most of them. Geluso (2014) tried to emphasize that in last two decades, there has been an increase in the integration of corpus-based language learning, or data-driven learning (DDL), as a supporting feature in teaching English as a foreign or second language. His research has mainly been on students' attitudes towards DDL as an instrument to facilitate writing, conforming that students believe DDL to be a useful and effective tool in the classroom. Smart's (2014) study, while supporting the application of DDL in language classroom, supports that there are measurable benefits to teaching ESL grammar inductively using paper-based DDL.

On the other hand, the present study is in some ways inconsistent with Yılmaz's (2015) research in which it was shown that DDL didn't have a significant impact on teaching or learning vocabulary in an EFL setting. Also, the present study is in some ways in contrast to the findings of Szymańska and Boulton (2015). Despite the benefits DDL may have for learners, as they mention, it has failed to fully catch on, both in the fields of corpus linguistics and language teaching in general.

This study has actually originated from the researchers' concern about finding a motivating technique to help language learners reduce their errors while using English collocations. The findings of the study revealed that the experimental groups, which used DDL method, had better performance as compared to the control group regarding the reduction of errors made while using the collocations. It was confirmed that data driven learning (DDL) is one of the best ways that can be implemented in language classes to boost the language learner's ability in different skills, particularly in writing. DDL helps the learners to delve into the authentic language body. By successively encountering the language forms, students can discover their language mnemonic and, consequently, come up with the correct usage of language. Having the blueprint in their minds, learners can discover the rules inductively and automatically eliminate their errors in any language areas and develop a well-formed body of language that is closer to the one existing in the mind of native speakers. Admittedly, working with the corpus should be under the guidance of a teacher who can direct learners towards what he or she wants them to go through.

These conclusions drawn above are of utmost significance for all who work as EFL materials developers, teachers, and learners in educational contexts. Materials developers are expected to insert various sections related to the use of collocations in their textbooks. It is hoped that the findings of this research will be beneficial for EFL teachers who try to familiarize students with new technologies and language corpora. Finally, students can save time and have extra practice at home by referring to the sites which contain a sort of language corpora.

References

Abdekhoda, M., Dehnad, A., Javad, S. J., Mirsaeed, G., & Gavgani, V. (2016). Factors influencing the adoption of E-learning in Tabriz University of Medical Sciences. *Medical Journal of the Islamic Republic of Iran*, *30*, 1-7.

Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *ReCALL, 21* (1), 37-54. doi: 10.1017/ S 0958344009000068.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60* (3), 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Boulton, A. (2012). What data for data-driven learning? *Proceedings of the EUROCALL 2011 Conference*, *20*, 23-27.

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, *19* (3), 307-328.

Celik, S. (2011). Developing collocational competence through web-based concordance activities. *Novitas Royal Research on Youth and Language*, *5*(2), 273-286.

Chang, J. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, *26* (2), 243-259. doi.org/10. 1017/S0958344014000056.

Chia, R., Wangb, D., Houc, Z, & Jin, S. (2012). *Data-driven optimal terminal iterative learning control*. Retrieved from https://doi.org/10.1016/j.jprocont.2012.08.001.

Ellis, N. (2006). *Cognitive perspectives on SLA: The associative-cognitive CREED*. *AILA Journal*, *19* (1), 100-121.

Flowerdew, L. (2008). *Corpus linguistics for academic literacies mediated through discussion activities.* Ann Arbor MI: University of Michigan Press.

Geluso, J. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, *26* (2), 225–242.

Hadley, G. (2002). *An introduction to data-driven learning*. Retrieved from https://journals.sagepub.com › doi › pdf.

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies*, *77* (4), 379–393.

Johns, T. (1988). *Whence and whither classroom concordance?* In T. Bongaerts (Eds.), *Computer applications in language learning* (pp. 9-27). Dordrecht, Holland: Foris.

Johns, T. (1991). *Should you be persuaded: Two samples of data-driven learning materials*. *ELR Journal*, *4*,1-16.

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

Larsen-Walker, M. (2017). *Can data driven learning address L2 writers' habitual errors with English linking adverbials?* Retrieved from www.elsevier.com/locate/system.

Lin, M., & Lee, J. (2015). Data-driven learning: changing the teaching of grammar in EFL classes. *ELT Journal.* doi:*10*.1093/ elt/ ccv010.

Maccio, D., & Cervellera, C. (2012). *Local models for data-driven learning of control policies for complex systems*. Retrieved from http://dx. doi.org /10.1016/ j.eswa. 2012.05.063.

Maftoon, P., Hamidi, H., & Sarem, S. N. (2012). The effects of CALL on vocabulary learning: A case of Iranian intermediate EFL learners. *Broad Research in Artificial Intelligence and Neuroscience, 3* (4), 19-30.

McCarthy, M. (1988). *Vocabulary*. Oxford: Oxford University Press.

Mishan, F. (2004). Authenticating corpora for language learning: A problem and its resolution. *ELT Journal*, *58*(3), 219-227.

Richard, J.C., & Rodgers, T. S. (2003). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Viola, S. R., Graf, S., & Leo, T. (2007). Investigating relationships within the Index of learning styles: A data driven approach. *Interactive Technology and Smart Education*, *4* (1), 7-18.

Schmitt, N. (2002). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *RECALL*, *26*, 184-201. http://dx.doi.org/10.1017/S0958344014000008.

Szymańska, L, A., & Boulton, A. (2015). Multiple affordances of language corpora for data-driven learning. *International*

Journal of Corpus Linguistics, 20, 560–569. https://doi.org/10.1075/ijcl.20.4.07gar.

Talai, T., & Fotovatnia, Z. (2012). *Data-driven learning: A Student-centered technique for language learning. Theory and Practice in Language Studies, 2* (7), 1526-1531. doi: 10.4304/tpls.2.7.1526-1531.

Yılmaz, E., & Soruç, A. (2015). The use of concordance for teaching vocabulary: A data-driven learning approach. *Procedia - Social and Behavioral Sciences, 191*, 2626-2630.

Appendix: *List of Collocations Used in the Pretest and Posttest*

Row	List of Collocations	Pre-test errors C-group	Post-test errors C-group	Pre-test errors E-group	Post-test errors E-group
1	**take a chance**	7.00	8.00	5.00	4.00
2	**take a look**	5.00	5.00	5.00	1.00
3	**take a seat**	12.00	13.00	10.00	5.00
4	**take a taxi**	9.00	7.00	8.00	3.00
5	**take notes**	4.00	8.00	4.00	6.00
6	**take an exam**	13.00	14.00	9.00	8.00
7	**take a picture**	3.00	5.00	1.00	1.00
8	**take care**	9.00	8.00	7.00	3.00
9	**take advantage of somebody**	7.00	8.00	5.00	4.00
10	**take a nap**	10.00	9.00	7.00	5.00
11	**take your time**	8.00	7.00	6.00	6.00
12	**take someone's temperature**	9.00	10.00	6.00	6.00
13	**take responsibility**	6.00	6.00	2.00	.00
14	**have a dream**	11.00	12.00	8.00	5.00
15	**go bald**	7.00	6.00	7.00	4.00
16	**go bad**	7.00	7.00	4.00	4.00
17	**go crazy**	6.00	8.00	5.00	3.00
18	**go abroad**	8.00	7.00	4.00	2.00
19	**go astray**	5.00	4.00	5.00	1.00
20	**go bankrupt**	13.00	11.00	9.00	6.00
21	**go out of business**	10.00	11.00	6.00	8.00
22	**go missing**	7.00	8.00	6.00	3.00
23	**go to the cinema**	9.00	8.00	7.00	4.00
24	**go port**	6.00	5.00	3.00	.00
25	**go on holiday**	8.00	3.00	8.00	.00
26	**go on a trip**	9.00	6.00	7.00	2.00
27	**go to college**	12.00	10.00	10.00	6.00
28	**go for a rest**	13.00	11.00	10.00	7.00
29	**go for a meal**	6.00	5.00	4.00	2.00
30	**go for a drink**	7.00	8.00	5.00	4.00