The Effectiveness Of Using Corpora On Lexical Revision In L2 Writing

Elif Tokdemir Demirel
Karadeniz Technical University, Department of English Language and Literature
61080 Trabzon, Turkey
elif6171@gmail.com

Semin Kazazoğlu
Karadeniz Technical University Department of English Language Teaching
seminkazazoglu@gmail.com

ABSTRACT
This study reports on the results of classroom research investigating the effects of using data-driven learning methods by students in revising their writing errors. The main purpose of the study is to examine to what extent is consulting a corpus effective in correcting lexical errors in their writing. It has been found in previous research that data-driven learning can benefit students in the revision process and that it works better for certain error types (Tono et al, 2014). The targeted error category was lexical errors including formal errors or semantic errors. For the study, a small corpus of 44 student paragraphs written in a timed writing task was used. The corpus of student paragraphs was analyzed for the common lexical errors. The error classification used in the study was drawn from the lexical error taxonomy of James (1998). All lexical errors were hand tagged according to the taxonomy. From among the common errors, certain errors were chosen for revision activities. Students were given hand-on instruction on using an online corpus and its concordancing tools and were asked to revise the selected errors by referring to the corpus. The effectiveness of consulting a corpus while revising errors was compared for different lexical error types.

INTRODUCTION
A corpus is a systematize collection of language data. Generally corpora serve descriptive purposes that is to provide a picture of the subjected language in a selected time frame. Modern computerized corpora consist of large databases of language systematically divided into subgenres. Corpora such as the BNC (British National Corpus) and COCA (Corpus of Contemporary American English) have user friendly interfaces which can be used freely by both researchers and language learners. These corpora have their own built-in concordancing tools which make it easy to conduct searches on various language items. Originally developed for linguistics research purposes, corpora and their concordancing tools have started to attract the attention of language practitioners who have started to use them for teaching purposes. After a short period of training, language learners can become users of these tools and make their own discoveries about the language they are learning. It is believed that corpora provide valuable information about the appropriate and up-to-date use of vocabulary for language learners. Therefore students can benefit from consulting a corpus while revising their writing. Additionally this process could increase their self-confidence as learners and increase their autonomy in learning.

Studies on the effects of corpus use in error correction point to the fact that certain error types are more suitable against checking against a corpus. For example in a recent study with Japanese learners, Tono, Satake, Miura (2014) classified a total of 188 errors into three major categories: 'omission', 'addition' and 'misinformation'. Their study revealed significant differences in correction accuracy rates between these three error types. Whereas omission and addition errors were easily identified by learners, misinformation errors were low in correction accuracy.

There is a recent interest in the use of corpora tools, for example the use of learner corpora to facilitate L2 writing. For example Creswell (2007) has evaluated the effectiveness of Data-Driven Learning (DDL) (Johns 1994; Hadley, 2002) on writing achievement. Creswell's conclusion is that:
DDL…applied in the context of the communicative teaching of writing skills, is moderately effective, and that there is potential both for the further development of learner corpora in an evaluative role, and for use of a wider range of instrumentation. (p. 267)

Additionally, Lee, Shin and Chon (2009) have investigated the effect of corpus consultation on the writing performance of L2 writers. They utilized Concord Writer 2 to help for the lexical revision. Their results point to the positive impact of corpus consultation on L2 writing improvement as well as the ability to notice errors.

THE STUDY

In the light of previous research on the use of corpora as a tool for developing L2 writing, the present study investigated the following questions:

1. Which lexical error types are more frequent in L2 writing by Turkish non-native students?
2. How does the use of BNC as a reference tool affect students' lexical revision process in L2 writing compared to un-aided revision?
3. Is the use of BNC as a reference tool more effective on revision in certain lexical error types than others?

The participants of the study were 44 prep class students at KTU Department of English Language and Literature. The context was a preparatory class writing course where students are trained to write paragraphs and essays following a process approach. The students' English level ranges from intermediate to advanced. All participating students were native Turkish speakers.

A corpus based approach was followed in the study to determine the frequency of lexical errors to be targeted for revision activities. For this purpose, a small scale corpus of student paragraphs was compiled. This paragraph corpus consisted of opinion paragraphs written in a timed-writing task on the following topic: "Discuss the advantages and disadvantages of using a credit card." The resulting paragraph corpus consisted of 44 paragraphs which had 919 word types and 5655 word tokens. The paragraph corpus was hand tagged for lexical errors using an adapted version of James' (1998) error taxonomy. The frequency of errors in different categories were determined by using AntConc 3.2.4. Concordancing software. Figure 1 shows the concordance lines with error tagging displayed by AntConc.

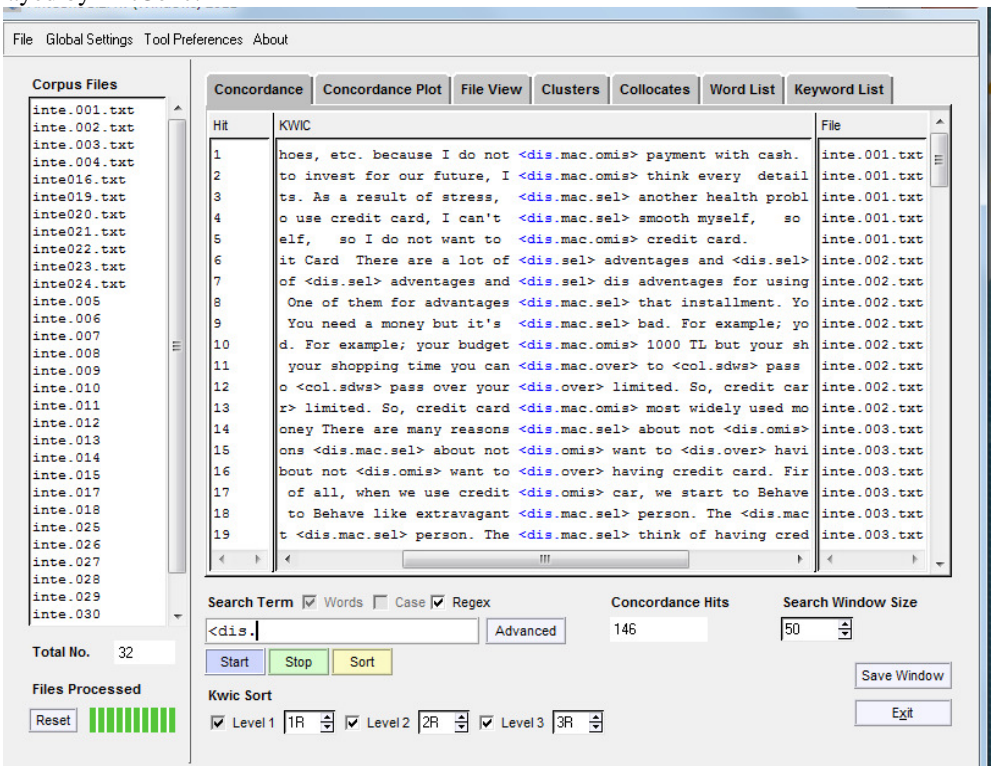

Figure 1. Hand-tagged concordance lines from the paragraph corpus

The BNC was used as a reference tool to aid students' revision process. The BNC website allows you to quickly and easily search the 100 million word British National Corpus (1970s-1993). The BNC was originally created by Oxford University Press in the 1980s - early 1990s, and now exists in various versions on the web. The BNC has its own built in tool which allows users to do searched and analyses. (see Figure 2)

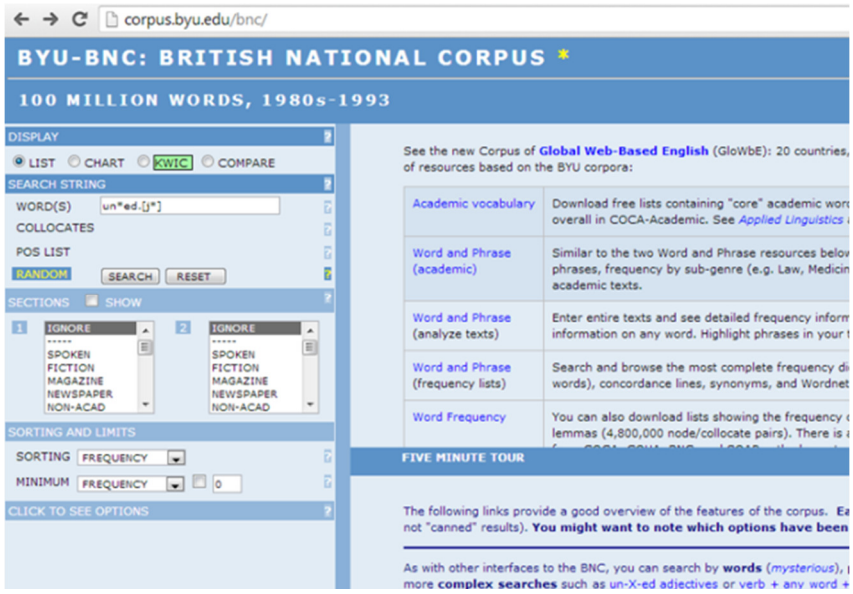

Figure 2. The BNC Corpus User Interface

The error taxonomy used in the study was developed by James (1998) and consists of two main lexical error categories of 'formal errors' and 'semantic errors' . (see Figure 3.) Formal errors category includes 'misselection', 'misformation' and 'distortion' errors. Semantic errors category includes 'confusion of sense relations', 'collocation', 'connotation' and 'stylistic' errors.

A Formal errors
- **1 Formal misselection**
  - 1.1 Suffix type
  - 1.2 Prefix type
  - 1.3 Vowel-based type
  - 1.4 Consonant-based type
  - 1.5 False friends
- **2 Misformations**
  - 2.1 Borrowing (L1 words)
  - 2.2 Coinage (inventing based on L1)
  - 2.3 Calque (translation from L1)
- **3 Distortions**
  - 3.1 Omission
  - 3.2 Overinclusion
  - 3.3 Misselection
  - 3.4 Misordering
  - 3.5 Blending

B Semantic errors
- **1 Confusion of sense relations**
  - 1.1 General term for specific one
  - 1.2 Overly specific term
  - 1.3 Inappropriate co-hyponyms
  - 1.4 Near synonyms
- **2 Collocation errors**
  - 2.1 Semantic word selection
  - 2.2 Statistically weighted preferences
  - 2.3 Arbitrary combinations
  - 2.4 Preposition partners
- **3 Connotation errors**
- **4 Stylistic errors**
  - 4.1 Verbosity
  - 4.2 Underspecification

Figure 3. Lexical Error Taxonomy (James 1998)

As a data collection procedure, a revision task was prepared based on the erroneous sentences chosen from the paragraph corpus. The participating students were randomly divided into an experimental group and control group. Each groups consisted of 10 students. The students in the control group were given a free revision task and were asked to correct the lexical errors depending on their intuitions. The students in the experimental group were given training on using the BNC online concordancing tool and were asked to make revisions after consulting the BNC corpus. In order to determine the correct revision of the incorrect student sentences an

answer key was prepared with the help of a native speaker university teacher with 10 years of teaching experience. The revision task was scored by using the answer key.

FINDINGS

Formal Errors

All lexical error types were hand tagged in the paragraph corpus. After the hand tagging, the frequency of lexical errors were determined by using AntConc. The frequency of lexical errors in different error categories are presented below.

In the formal error category there are three subdivisions: formal misselection, misformation and distortion. Sentence 1.(a) shows an example of a formal misselection mistake, specifically a suffix type error as the adverbial suffix '-ly' has been omitted.

1. (a) All in all, all these disadvantages are the most common examples and if you do not want to come across the bad result of credit cards, you should use it more **<for.suf> cautious**.

Table 1. Frequency of formal misselection errors.

FORMAL MISSELECTION	24
suffix.type	0
prefix type	0
vowel-based type	0
consonant-based type	24
total	

Table 1 shows the frequency of formal misselection errors in the paragraph corpus (n=24). As can be seen from the table, all errors in this category relate to the suffix; either omission of the required suffix or selection of wrong suffix.

Table 2. Frequency of misformation errors.

MISFORMATIONS	0
borrowing	0
coinage	13
calque	13
total	

As the second subdivision of formal errors misformations were determined in the paragraph corpus. The frequency of misformations is presented in Table 2. There are a total of 13 misformations which are categorized as calque (translation from L1). Sentence 1.(b) shows an example of calque error. Here the learner has translated from L1 since in Turkish a password can be 'solved', but in English instead of 'solve a password', 'break a password' is used.

1. (b) A computer hacker could easily **<mis.calq> solve its password** and they could use my credit card more than my limit.

The third subdivision of formal errors is distortion. At the distortion category, the James taxonomy was not found adequate as it only included letter level distortions but not word level distortions. Therefore, distortions were divided into two types: micro-level (those involving letter level distortions) and macro-level (those involving word level distortions). Table 3 shows the frequency of distortion errors both at the micro-level and macro-level. Most frequent type of distortion was found to be omission for both micro-level (n=18) and macro-level (n=38) distortion errors. The second most frequent error type is misselection and at both microlevel (n=12) and macro-level (n=37), however the frequency of macro level errors are higher for all error types.

Table 3. Frequency of distortions

DISTORTIONS	MICRO-LEVEL	MACRO-LEVEL
omission	18	38
overinclusion	10	16
misselection	12	37
misordering	1	14
blending	0	0
total	41	105

Sentence 1.(c) below shows an example of a distortion error at the micro-level, specifically an omission since a letter has been omitted when writing 'because' by the learner.

1. (c) To sum up, people should not use credit cards **<dis.omis> becuse** of these reasons.

Sentence 1. (d) below shows an example of a macro-level distortion, specifically a macro-level omission. Here the word 'become' has been omitted from the phrase 'become addicted to'.

1. (d) Moreover they **<dis.mac.omis> addict to** <for.suf> use credit cards.

Semantic Errors

In the semantic errors category there are only 8 errors in the confusion of sense relations error subdivision. 5 of these errors relate to using a general word for a restricted meaning. And 3 of the errors relate to using two near synonyms redundantly in the same sentence.

Table 4. Frequency of confusion of sense relations

CONFUSION OF SENSE RELATIONS	
superonym for hyponym	5
hyponym for superonym	0
inapproporiate co-hyponym	0
near synonym	3
total	8

Sentence 2. (a) below shows an example of near synonym error. I the sentence both unnecessary and extra have been used redundantly because both have very similar meanings.

2. (a) Secondly, when I use credit card, I have to pay its interest and what I say is that I pay **<sem.near> unnecessary extra** money.

When we consider collocation errors, we can see that the most frequent error type is semantically determined word selection. There are a total of 17 errors in this category. In terms of collocations learners also seem to have some difficulty in selecting the correct preposition partner for words, therefore fore there are 14 errors in the preposition partners category. In terms of arbitrary combinations there are only 3 errors detected.

Table 5. Frequency of collocation errors

COLLOCATION ERRORS	
semantically determined word selection	17
statistically weighed preferences	0
arbitrary combinations	3
preposition partners	14
total	34

Sentence 2. (b) shows an example of a semantically determined word selection error. Here the learner has used the word suicide as if it were a verb, however this word has a verb which closely collocates with it. This word is 'commit' but the learner has omitted the collocation.

2. (b) Meanwhile, there are a lot of <dis.mac.sel> person **who <col.sdws> suicide .**

Figure 3 shows the overall distribution of the error frequencies in the paragraph corpus. According to this distribution the most common error type is macro-level distortion errors, and the least frequent error type is confusion of sense relations. Overall formal errors are much higher in frequency compared to semantic errors.

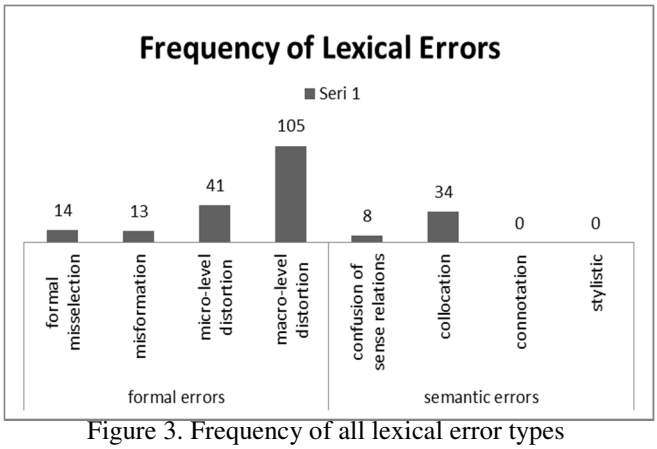

Figure 3. Frequency of all lexical error types

Comparison of Revision Accuracy Between Experimental and Control Group

At the last step of the study, the accuracy of revision were compared between the experimental and control groups through the completion of a revision task. Table 6 shows the accurate correction rates of distortion errors under the category of formal errors. According to the results, in this error category the corpus aided group (M: 80) performed better than the free correction group (M=52). The experimental group correctly revised 4 errors out of 5 errors; whereas the control group correctly revised 2,6 errors out of 5.

Table 6. Comparison of experimental and control group in terms of revision accuracy of formal misselection errors

Part A	**Corpus aided correction**	**%**	**Free correction**	**%**
	3	60	1	20
	3	60	3	60
	4	80	3	60
	3	60	2	40
	4	80	4	80
	5	100	3	60
	4	80	3	60
	5	100	2	40
	5	100	4	80
	4	80	1	20
average	**4**	**80**	**2,6**	**52**

Table 7 shows the revision accuracy rates of the experimental and control groups in different error categories. According to the overall results, in all lexical error categories, the learners scored higher in terms of revision accuracy. Among these, learners in the experimental groups were most successful in revising the formal misselection errors, followed by distortions and semantic errors.

Table 7. Overall comparison of revision accuracy between experimental and control group

Success in correction rates	**CAC***	**FC****
	%	**%**
Formal misselection	80	52
Misformations	46,67	22
Distortions	73,33	52
Semantic errors	46,67	26
Collocation errors	60	44
Total	61,33	39,2
Corpus aided correction* * Free correction*		

CONCLUSION

This study served to two main purposes: first determining the frequent lexical errors in student writing and second determining which error types are more suitable for revising with the help of a corpus tool. As a result of the study, it was found that L2 writers make most frequent lexical errors in the formal error category and most frequent of these errors are micro-level and macro-level distortions. In the semantic error category, the most frequent error type is collocation errors. These results show that Turkish L2 writers have most difficulty in selecting appropriate words contextually and also they have a lack of knowledge about collocation use.

The second research question investigated was the effect of BNC corpus as a reference tool in revising lexical errors in L2 writing. The results of the study shows that the BNC corpus serves as an effective tool which helps L2 writers greatly in revising their lexical errors compared to intuitive judgements. Although they can make accurate revisions to some extent depending on their intuitions, the level of accuracy is very low compared to corpus aided revision.

The third research question specifically enquired which error types are most suitable for revising with the use of a reference corpus. As an answer to this question, the revision accuracy rates show that most accurate revisions were done for formal misselection errors, distortions and collocations. On the other hand, the L2 writers have not benefited from reference corpus in revising misformation errors and semantic errors. Overall, these results point to the importance of corpus use and concordancing as an effective tool in helping L2 writers in revising their lexical errors, specifically related to contextual vocabulary selection and collocations. As an implication, the researchers greatly recommend the use of corpus tools and reference corpora as an aid in second language writing classes.

References

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.)Sage Publications, Thousand Oaks CA

Hadley, G. (2002). An introduction to data-driven learning. *RELC journal*, 33(2), (pp. 99-124).

James, C. (1998). *Errors in language learning and use*. White Plains, NY: Addison-Wesley Longman.

Johns, T. (1994). *From printout to handout: Grammar and vocabulary teaching in the context of Data-driven Learning*. Perspectives on pedagogical grammar, (pp. 293-309).

Lee, M., Shin, D., & Chon, Y. V. (2009). Online corpus consultation in L2 writing for in-service teachers of English. 영어교육, 64(2).

Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 26(02), (pp. 147-162).