

A Corpus-based Study: EFL College Students Using COCA to Improve Lexical Usage in Written Production

Thao Thi Ngoc Nguyen

University of Science and Technology of Hanoi

doi: <https://doi.org/10.21467/proceedings.132.16>

ABSTRACT

Corpus linguistics has attracted growing interest from linguistics, researchers, and language educators. This area is potential for studies in linguistics as well as improvement in language learning. This study aims at investigating how a corpus-based website named <https://www.english-corpora.org/coca/> - or COCA - helps improve students' lexical usage in written production. Participants were EFL students from a university using COCA over two months for revising vocabulary in their writing assignments. Data included the measure of textual lexical diversity (MTLD) of the participants' revised assignments and their reflections/reviews on using COCA. Results reveal some positive changes in MTLD values, while participants' feedbacks suggest COCA is useful for gaining knowledge about lexis, and improving the use of collocations, synonyms as well as high-lexis despite some challenges of unfamiliar interface and overwhelming contents. The study implies the feasibility of applying corpora provided that there are sufficient facilities and careful training for learners.

Keywords: corpus linguistics, data-driven learning, COCA

1 Introduction

Corpus linguistics has become more influential in linguistics research and language education over the past decades. As the corpora utilize the natural texts and store them systematically, linguists can exploit the data or conduct corpus-based or corpus-driven studies related to language problems employing quantitative methods. The advantage of this storage, as claimed by Wallis (2021, p.3), can be seen in the case of research on language variation historically which "must rely on data". Wallis also asserts that corpora can also corroborate the evidence used for psycholinguistics studies. As for language education, the importance of corpus is reinforced under the communicative approach that has been adopted in recent years. Language classrooms focus more on instructing students in as authentic English as possible, which accordingly promotes data-driven learning, to increase learner exposure to a massive volume of language data taken from real-life contexts. The corpora, hence, is claimed to be a significant resource of data for foreign language learning and teaching (O'Keeffe et al., 2007, p.21).

In Vietnam, corpus linguistics and its applications are somehow modest in both research and classroom usage. Collocation dictionaries, writing samples, or traditional thematic vocabulary-teaching books are more dominant and regarded as more convenient for in-class users. While these materials are deemed useful and accessible to most language learners, it is not redundant to introduce alternatives that can support their learning more authentically, especially in the era where smart devices and high-speed Internet connection are becoming more approachable. Realizing that English language research and education in Vietnam is not the exception to this trend, this paper aims at contributing a critical account of one of the largest and most updated corpora currently - Corpus of Contemporary American English (COCA).



Particularly, the study aims at finding out the answers to the following questions:

1. How, if at all, does the measure of textual lexical diversity in participants' writing change after using COCA?
2. What are participants' perceptions of the use of COCA?

2 Theoretical Framework

2.1 Corpus Linguistics

A corpus, according to O'Keefe et al. (2007, p. 1), is "a collection of texts, written or spoken, which is stored on a computer [...] and can be stored and analysed using analytical software." These scholars also stress that a corpus has three features (pp.1-3):

- It is a principled collection of texts which must be representative of something;
- It is a collection of electronic texts stored on a computer which can be written texts or transcribed versions of spoken texts, or a mix of both.
- It is available for qualitative and quantitative analysis, allowing users to obtain results about a word's frequency in particular contexts or its concordance lines beyond the contexts.

Similarly, Wallis (2021, p.4) explains that corpora are "simply collections of language data processed to make them accessible for research purpose". The data are collected from uncontrolled and natural conditions, with texts not being separate or random but meaningful passages. Besides written corpus, spoken corpus containing several recordings can be built with orthographic transcription. This transcription provides richer data than written form, as other features of speech can also be annotated for further usage. In terms of linguistic research, Wallis believes corpora provide empirical evidence including factual, frequency, and interaction evidence (p.6), specifically the frequency evidence found when examining a language corpus can reveal language patterns or phenomena. O'Keefe and McCarthy (2010) claim that corpora can supplement many of the processes used in discourse analysis, while its comparison functions facilitate comparative studies in literature and translation studies. Besides, the two authors list other areas where corpora can be exploited. Corpus linguistics technique, for example, can allow filtering metadata of language users that are central in sociolinguistics or revealing noteworthy frequency in media texts which can evidence a hidden ideology in media discourse. Other disciplines such as forensic linguistics, pragmatics, or political discourse are also counted in this list.

In terms of pedagogical application, corpora can be used to create assignments or materials to provide learners with hands-on experience. There are two approaches to using corpora to language teaching and learning: direct and indirect (Römer, 2011, p. 207). Data-driven learning is the direct application of the tool in which teachers and learners engage with the resource to inform their language use, whereas indirect application involves works by syllabus developers, researchers, and material writers, as illustrated in Figure 1 below. Discussing the same topic, O'Keefe and McCarthy (2010) consider corpora can either be built from collected native texts or developed from texts produced by learners (learner corpora). Besides material development, corpora data have also been adopted in testing and teacher education.

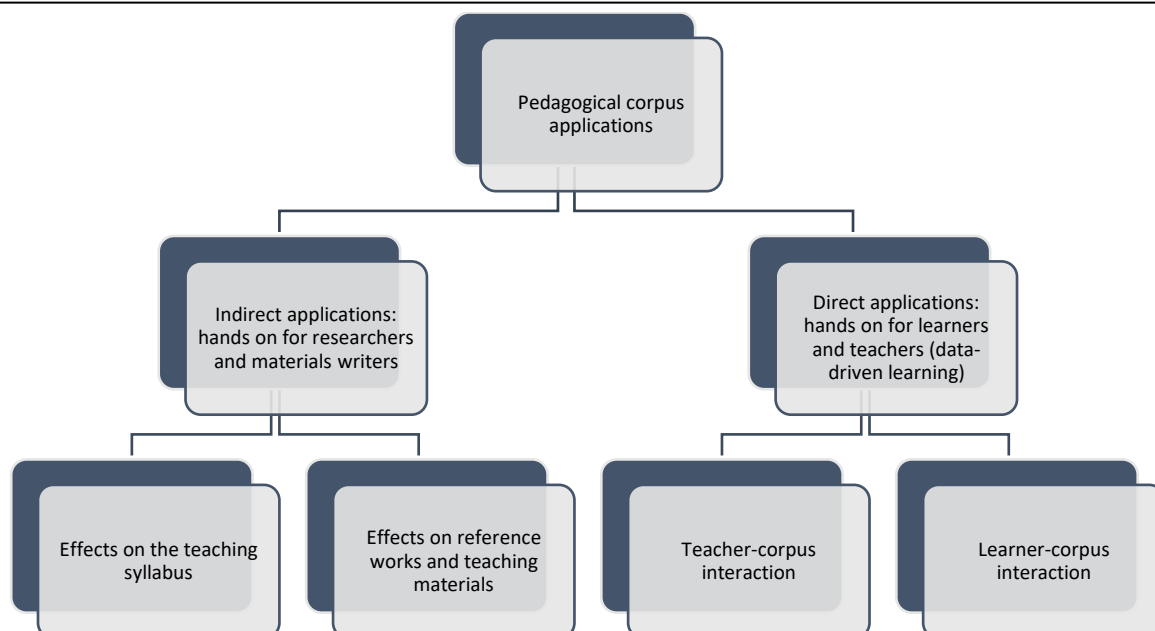


Figure 1. *Pedagogical corpus applications*

Several studies back up data-driven learning's benefits in promoting learner and teacher autonomy. The usage of learner corpora, according to Lewandowska (2014, p. 240), allows students to undertake interlanguage analysis among themselves, as well as reflect on course content and use that reflection to negotiate with their teacher.

Tono, Satake, and Miura (2014, p. 148) similarly assert that the introduction of data-driven learning has improved inductive teaching by allowing learners to learn from the data on their own. Data corpora, whether in printed or electronic forms, could be used in a variety of ways to provide a setting for teachers where they can organize interactive activities and for learners to have more control over their learning.

2.2 Corpus of Contemporary American English (COCA)

Corpus of Contemporary American English (COCA) is a corpus created in 2008 by Professor Mark Davies of Brigham Young University (Davies, 2008; Davies, 2010, p. 163). It is now "the first large-scale monitor corpus of any language, which is balanced between a number of different genres" (p.462). Its utility, according to Davies, is due to two features: continually updated data and balance in genres. The corpus architecture allows users to search "by substring, lemma, part of speech, collocates, synonyms" (ibid.), as well as see and compare frequencies.

According to the COCA website, COCA contains more than 1 billion words from 1990 to 2019 with nearly 500,000 texts updated periodically. Each genre contains approximately 120-130 words categorized into spoken, fiction, magazine, newspaper, academic, blogs, other webpages, and TV and movie subtitles. To access the data, users need to register with their email and can use the corpus' basic functions for free in online mode. To download or explore more from the resources, as well as to conduct a higher number of queries or retrieve more data, users need to upgrade to the paid premium account. In terms of the user interface (Figure 2), COCA's webpage has search boxes and query tools on the left side and tutorials on the right side. The dashboard can be navigated through the selection of one of the four tabs – Search, Frequency, Context, Overview. This can avoid the inconvenience caused by using the forward or backward buttons of the browser. Advanced queries which are Sections, Text/Virtual/, Sort/Limit, and Options are normally collapsed unless users click on them.

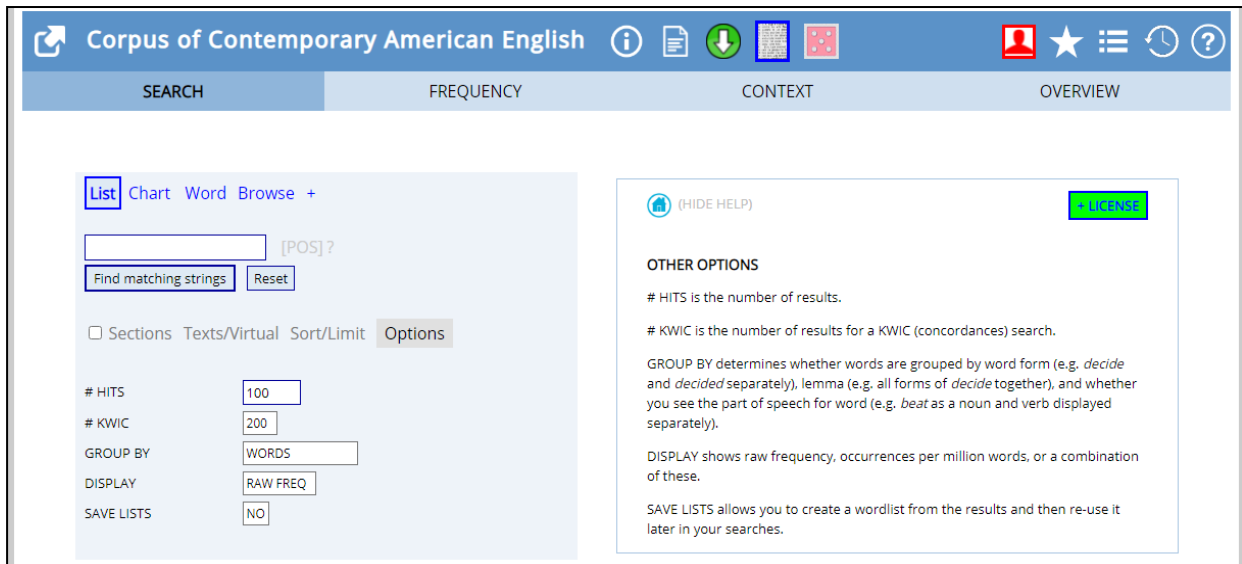


Figure 2. COCA's interface

The results of the searched keywords are displayed in the order of frequency and can be filtered by categories (in spoken language, in magazines, academic journals, et cetera) so users can narrow their search to a specific genre. COCA also includes plentiful query functions, specifically searching for collocation, comparing two confusing words (Figure 3), and showing keywords in context (KWIC). This KWIC function helps display articles, prepositions, conjunctions, and adverbs which are frequently used in conjunction with the searched keywords. Users can also see the original context of the keyword-containing sentence, as well as the preceding and subsequent sentences and information on its original publication. To illustrate, in Figure 3 below, the two words keyed in are money and currency. The corresponding result compares how frequently a word can go before “money” or “currency”, with statistics being drawn from the database of texts collected in the corpora. For example, before money, there are 4160 contexts using “save money” as a collocation, whereas there is no context using “save currency”. The majority of collocates preceding money are verbs such as “save”, “spend”, “raise”, and “borrow”, whereas those preceding currency are adjectives such as “single”, “convertible”, “weaker”, and “strongest”.

SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2)
 SORTED BY RATIO: CHANGE TO [FREQUENCY](#)

WORD 1: MONEY			WORD 2: CURRENCY		
WORD	W1	W2	WORD	W2	W1
SAVE	4160	0	SINGLE	304	3
SPEND	2596	0	CONVERTIBLE	73	1
SPENDING	2083	0	UNIFIED	24	0
RAISE	3747	1	WEAKER	23	0
YOU	1566	0	WIDE	22	0
TAXPAYER	1367	0	REFERENCE	19	0
BORROW	1091	0	CULTURAL	17	0
TAKE	1060	0	EURO	63	2
MAKE	8302	4	STRONGEST	15	0

Figure 3. COCA's word comparison

2.3 Using COCA in Language Education

As texts stored in COCA are in written or transcribed forms, it is understandable that a significant number of studies are about their benefits for teaching vocabulary, grammar, and writing skills.

In terms of indirect application, Yan (2012, p. 392) reveals that corpora could be a useful tool to expand one's vocabulary and understand how words are used in everyday situations. For material writers and developers, COCA is a great source providing a large volume of native-like data to include in the books, also saving time on materials compilation because all information about a word "including the meaning, collocation, colligation, and register" (p.393) is easily accessible after a click.

Regarding direct application, Yan (2012, p.392) and Yusu (2014, p.69) argue that COCA provides valuable assistance in acquiring understanding about morphological variation in words and their affixes or word roots usage. COCA can also assist students learn how words are collocated appropriately, and provide keyword-surrounding sentences illustrating how those words "in grammatically with other words" (Yusu, 2014, p.70). Students, according to Yan (2012, p.392), are better able to recognize how distinct meanings of a term are employed in different contexts. This idea is supported by Alshaar and AbuSeileek (2013, p. 73) because in their study the participants showed a preference for COCA and stated that the concordance help "identify useful phrases in context". Wu, Witten, and Franken (2010) also show that web-derived corpora can be helpful for students wanting to improve their collocation accuracy, in spite of some errors detected as some only chose more familiar words rather than those that match the context better (p.99). Similarly, COCA's word comparison function is mentioned in Yusu (2014, p.70) as an example of how effective the tool is in assisting the student in recognizing how two words and sentences differ in contexts. Furthermore, Alshaar and AbuSeileek (2013, p. 73) claim that their participants achieve more self-confidence "not only in writing, but at the same time reading, grammar, pronunciation, and vocabulary" and expect broader integration of the tool at tertiary level.

Despite the benefits, using COCA may represent several challenges. One challenge is the gigantic volume of data requiring a significant amount of time needed to process. Its large data volume requires teachers and learners to get familiar with online corpora and offline corpus tools (Römer, 2011, p. 214; Yusu, 2014, p.69). Processing data provided by the site, understanding vocabulary knowledge, and mastering the use of corpus technically really take time, according to Yusu (2014). Another problem concerns devices and facilities. As Yusu (2014, p. 69) points out, some places have inadequate amenities such as the computer lab and Internet access for corpora access. Römer (2011, p. 214) also adds there is a chance that learners will be uncomfortable using computers in their language. The third issue in using COCA is related to high-level words and complicated sentence structures, which challenges learners at the beginning level (Römer, 2011, p. 214). Yan (2012, p. 392) agrees that corpora's features may be effective mainly for those with advanced English proficiency to search and understand the results. To completely understand the meanings of words in diverse situations, low levels may need to consult their mother tongue's dictionary. Employing corpus therefore may increase burden while possibly lowering students' confidence and motivation to learn, as warned by Yusu (2014, p.69).

3 Methodology

3.1 Research Design

In this study, the qualitative approach and convenience sampling are employed to collected data from a small sample size over a lengthy period, such as a semester or several weeks. This is appropriate because qualitative research use small sample sizes and convenience sampling to get comprehensive and concentrated data (Friedman, 2011, pp. 185-186). Participants were 11 second-year undergraduates majoring in English for Business, or "the students". They were 18 to 20 years old and from the same class. Their English proficiency levels were B1 according to the Common European Framework of Reference -

CEFR. Their Writing lecturer was the researcher's former colleague and was willing to allow the study conducted.

The design involves students accessing COCA via <https://www.english-corpora.org/coca/> over two months to revise lexical usage in their four writing assignments before submitting them to their lecturer. Students used the tool outside classrooms due to limited class hours, and they all received training and support from the researcher during participation.

3.2 Data Collection and Analysis

3.2.1 Data Collection

Data were collected from two sources. First, participants' writing was used to calculate the measure of textual lexical diversity – MILD, and how this measure changed over several writing assignments, supposedly resulting from students' COCA consultation. This was to answer research question 1, and the values were obtained from those who edited at least 3 assignments with the help of COCA. Another source of data was reflections/reviews. After the students had finished correcting and submitting all writing assignments, they wrote the reflections including their experiences with COCA, an assessment of its impact on their lexis and other viewpoints. The students were provided with some guiding questions (Figure 4) for their reflections, but they were also encouraged to include more information outside the questions. To corroborate these reflections, the participants were advised to submit quick reviews after each time they used COCA because they might forget parts of their experiences especially if they did not use the tool regularly. It was suggested that the reviews should contain a list of words that participants altered after consulting with COCA, as well as what they liked or disliked about the tool. The style, formality, and frequency and length of the reviews were determined by the students' level of comfort and engagement. In total, there were 16 writing pieces by five participants, 18 quick reviews, and 10 formal reflections.

1. How frequently do you use the tool in revising your lexis in your writing assignments? Please indicate which assignment you used/did not use.
2. What do you use the tool for? Please list all purposes of your usage.
3. To what extent did the tool serve all those purposes? Please explain your evaluation in details with examples, if possible.
4. Do you see any changes in your range of vocabulary as a result of using the tool? Please specify.
5. Do you see any changes in your lexical diversity in your writing as a result of using the tool? Please specify.
6. Do you see any changes in your choice of word and collocations in written production as a result of using the tool? Please specify.
7. Do you see any other changes in your English learning as a result of using the tool? Please specify.
8. Do you choose to use the tool to support your English learning in the future? Please specify.

Figure 4 Guiding questions for reflection

3.2.2 Data analysis

To answer research question 1, the participants' lexical diversity (LD) measure was calculated using the web service <http://www.textinspector.com/>. Measuring lexical diversity provides information about how

learners produce language using vocabulary knowledge, according to Koizumi (2012, p. 60), which is best demonstrated by the measure of textual lexical diversity, or MTLT. This MTLT is developed by McCarthy and Jarvis (2010), which is defined as "the mean length of sequential word strings in a text that maintains a given TTR value" (p.384). The developers choose 0.72 as the default value. When compared to other measures at textual level, MTLT is said to produce accurate results for lexical diversity and is "least affected by text length" (Koizumi, 2012, p. 67).

As for research questions 2, the input from the participants' reflections and reviews was analysed. Content analysis namely the grounded theory method was employed, which was pioneered by Glaser and Strauss (1967) and frequently used in qualitative research (Lawrence & Tar, 2013, p. 30). The categories from the data was constructed through two major phases proposed by Charmaz (2006, p. 60). In initial coding, the researcher conducted word-by-word or line-by-line analysis by moving quickly through data and comparing data with data to identify as many ideas as possible. In focused coding, codes that are most prevalent to the study are selected and synthesized into larger conceptual categories. Table 1 shows an example of Charmaz's (2006) two-stage phase for analyzing participants' reflections.

Table 1 *Coding participants' reflections*

Focused coding	Initial coding		Examples of raw data (translated from Vietnamese)
	Categories	Sub-categories	
Evaluation of the effects of corpus consultation on productive lexis	Lexical diversity		My writing avoids word repetition.
	Word choice and collocation	Precision of word combination Changes in word level	I searched the word "prestigious" and it could collocate with the word "company". A lot of words at C1-C2 level
Evaluation of other aspects of the corpora	Contents	Amount of information	Results show many contexts of use but sometimes confuse me due to too many details.
		Types of information: collocation, synonyms, meanings, examples	I can check collocation and synonym at the same time.
	Interface	Text: colour, font size, highlight	Unattractive format due to small-size words, messy use of highlighting colours in examples.
		Layout	Corpus has many parts in its layout so I am confused about which part to look at first.
	Ease of use		The instruction is hard to understand.
	Needs satisfaction		Corpus did not serve any of my purposes
	Preference		If I can choose, I will not choose to use corpus.

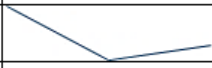
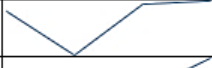
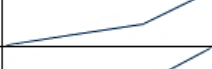
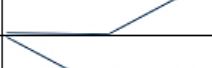
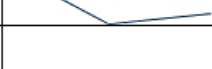
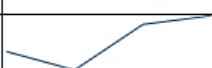

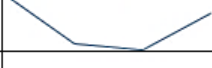
A system of identifying codes was employed for reporting the findings. The writing assignments were coded as W+a number ranging from 1 to 4, the participants' names were arranged in alphabetical order coded as S+a number ranging from 1 to 11, and the quick reviews/reflections were both coded as R+a number where the number represents the document's submission order. Students' ideas are also cited using this system, for instance, [S1.R1] stands for Student 1, Reflection/Review 1. Direct quotations are tagged with the word "translated" if they were originally in Vietnamese, for example, [S2.R4/translated] means Student 2, Reflection/Review 4 translated from Vietnamese into English.

4 Results

4.1 Research Question 1: How, if at all, does the measure of textual lexical diversity in participants' writing change after using COCA?

Changes in the measure of textual lexical diversity – MTLTLD – are detailed in Table 2 below. The majority of students used COCA for revising at least three assignments while three of them (S2, S7, S11) only used it one or two times. The changes were positive for half of the participants (S3, S4, S5, and S8) as the values increased from the first to the last assignments. Among these four people, the most remarkable improvement was in S4's MTLTLD which grew by 26.99 whereas the others had a more modest difference from 11.72 to approximately 17. The other four students (S1, S6, S9, and S10), however, saw a decline in their MTLTLD. S6's values showed the biggest discrepancy in the table as they dropped significantly by 39.91, whereas the smallest difference among all students belonged to S9 whose MTLTLD decreased by 10.64.

Table 2: MTLTLD value in participants' writing

Student	W1	W2	W3	W4	Difference	Trend
1	112.06	80.27	88.99		-23.07	
2	113.09					
3	137.61	87.17	145.56	149.33	11.72	
4	84.56	89.87	94.98	111.55	26.99	
5	89.22	88.82	105.07		15.85	
6	135.3	85.89	95.39		-39.91	
7	96.66			91.56		
8	96.31	87.47	109.17	113.34	17.03	
9	106.73	87.95	96.09		-10.64	
10	134.64	83.83	77.37	116.09	-18.55	
11	88.43					

4.2 Research Question 2: What are participants' perceptions of the use of COCA?

4.2.1 Lexical Aspect

Lexical diversity (LD)

There were three students (S7, S8, and S11) who gave no thoughts on LD change, whereas two others (S1 and S5) denied the impacts of COCA on the diversity of their productive vocabulary. Specifically, S5 said

"No change" [R2/translated] and S1 stated that she saw some changes in her lexical diversity but this was not due to the use of COCA.

Five students (S2, S3, S4, S6, and S9), however, reported the diversity level of their written lexicon had increased. S3 specifically stated: "I think my writing's vocabulary is more diverse. My writing avoids repetition, which I was unable to do before..." [R4/translated]. S4 similarly said: "I think there are some changes because Ms. X [the Writing teacher's name] said my vocabulary is much better than the previous semester" [R3/translated]. In another example, S9 and S6 also provided some synonyms they had learnt. S3 also confirmed the favorable effect despite some skepticism that the diversity degree was limited.

Collocation

COCA, according to four students, had little to no impact on their word choice and collocation use. S5 claimed that there was no difference whereas S3, S5, and S8 were doubtful and believed the corpora "made insignificant change" [S8.R3/translated]. S3 mentioned that the tool sometimes did not return any collocation results for the terms he keyed in, also examples in the concordance lines were so unfamiliar that he needed to use other collocation dictionaries.

On the other hand, five students (S2, S6, S9, S10, and S11) agreed that collocation usage in their compositions had improved. To explain how their collocation usage altered, four of these students used the phrase "more precise" [S2.R1, S9.R3, S10.R3, S11.R3/translated]. S11 stated that her long expressions could be condensed to be more concise and relevant to contexts. S6 also confirmed that she could avoid using spoken language, choose the right words for the text genre and use more high-frequency academic ones. S10 stated in her R1 that as she saw COCA showed the phrase "prestigious company", she gained more confidence in using this phrase as well as learned some other adjectives collocating with the word "company". Students also said that they were able to discover more advanced words to make some revision to their initial writing drafts. S9 also revealed that the level of COCA's collocations was near C1/C2 level (CEFR) but uncommon in a bilingual Vietnamese-English Dictionary, which was very useful to learn. This claim was supported by some examples of changes specified in students' quick reviews (Table 3), in accordance with Cambridge Online Dictionary (<http://dictionary.cambridge.org/dictionary/english/>):

Table 3 *Change of word level in participants' writing after corpus consultation*

	First Version	Revised Version	Change
S3.R2	Enough	Sufficient	A2→B2
S3.R2	Strengthen	Intensify	B2→C2
S3.R3	Strong similarities	Striking similarities	A2→ B2
S9.R1	Make a significant profit	Reap a significant profit	A2→C2
S11.R2	Nearby your house	In proximity to your accommodation	B1→C2
S7.R3	Has had many partnerships	Has established many partnerships	A2→C2

4.2.2 Other Aspects

The participants commented on a variety of other factor, which supports their perceptions of COCA’s effect on their productive lexis. Their opinions surround interface, contents, ease of use, demand satisfaction, and preference for using COCA.

Interface

Almost all of the comments on the interface were negative and detailed, which was an outstanding feature. For example, S4 expressed her dissatisfaction with COCA after using the tool for a while, whereas more specific criticisms were made by S5, S9, and S10. The usage of multi-colour was criticized first since it generated users’ confusion and made the contents more puzzled (Figure 5). S9 even remarked that she needed to take notes of each colour’s meaning whereas S5 believed that these colours dazzled her eyes. The second criticism concerns the improper layout of information. S10 felt perplexed since COCA exhibited multiple portions of the page on the screen at the same time, and she didn’t know which one to focus on first (Figure 6). Similarly, S3 suggested that information regarding Synonyms, Definitions, and Collocates be highlighted. The third problem is about the font size, which was considered to be too small to view and understand (S5, S10 and S3).

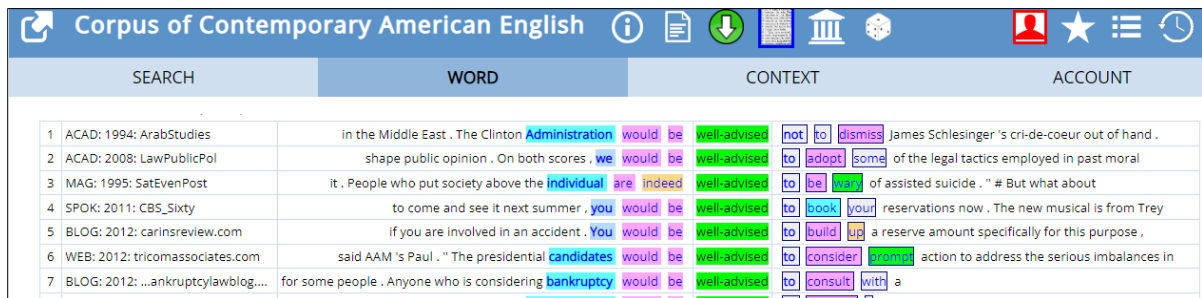


Figure 5 Dazzling colours



Figure 6 Overwhelming information

Contents

COCA received positive reviews for its contents. To illustrated, S10 said: "I found the corpus multifunctional. I could check collocation and synonyms at the same time, I do not need to spend more time checking different dictionary webpages" [R1/translated]. S10 and S6 agreed that that the superiority of COCA to other dictionaries they had used was the tool included word’s frequency of usage in different types of genres. Thanks to this feature, S6 could revise her writing more easily by choosing words more

frequently used in academic contexts. In her review, she continued: "Some other websites list out a word's synonyms but do not indicate which is equivalent to that word as a verb or a noun. But this corpus [COCA] could clearly show it." [R2/translated]. Besides, S3 recorded some improvements in his review as he learnt more meanings of the three words "indemnity, liability, administration" [R1/translated]. He also discovered expressions which sounded "more professional" [R2/translated]. The improvement is illustrated below:

Before consultation	After consultation
- The research suggests that + O + V (a long sentence)	- The research implies + Noun Phrase (a shorter one and sounds much more professional)

[extracted from S3.R2]

The contents, on the other hand, received some negative comments saying that it was overwhelming. According to S6, she found information related to concordance lines on the COCA's website more appropriate for those who conducted research in linguistics than those who only practiced writing skill like her, so she ignored the information. This inconvenience was confirmed by S9 who found the concordance line part inconvenient because of its separated sentences, and he labeled this part a "disorder" [R1/translated]. Another comment from S3 was that these concordance lines were "too unfamiliar" [R4/translated], so he checked the words by using another online dictionary.

Ease of Use

When it comes to the ease of use, participants' feedback was mostly negative. Some participants made general comments such as the corpora were "difficult to use" [S8.R3, S9.R3/translated] or other online dictionaries were "far more effective and easier to use" [S1.R3/translated]. Two participants (S3 and S6) said that COCA returned no results for several of the terms they typed, and they were unsure if this was due to a technical issue or the tool stored no data for the terms. S6 also confessed that she felt "upset" [R1] because she did not understand part of the information shown on the screen. The same confusion was confirmed by S1: "Sometimes the page has some errors, or maybe I do not understand how to use it because at first when I typed the words the page show errors, but later I checked it again and it went smoothly?" [R2/translated].

Users' Demand Satisfaction

COCA served several purposes mentioned in the students' reviews and reflections, although the ideas were not identical. According to S4, COCA assisted her in checking word meanings, also six students (S1, S4, S7, S9, S10, and S11) confirmed that COCA met their demand for collocation checking. Two students (S11 and S2) agreed that COCA could help them looking up synonyms while S11 added that she could identify "less common" [R3] synonyms. In contrast, two students (S1 and S8) said that the tool was ineffective in finding synonyms. Furthermore, two students (S3, and S6) used the corpora to verify the context of words, with S3 being the only to mention checking prepositions. Finally, while most of the participants felt that COCA can help them with at least one of their purposes, S5 claimed that the tool was useless. A summary of all these opinions is provided in the table 4.

Table 4 Participants' purposes of using the corpora

Student	Purposes				
	Find collocation	Find synonym	Check meaning	Check context	Others
S1	✓	✓			
S2		✓			
S3		✓	✓	✓	Check preposition
S4	✓	✓	✓		
S5					
S6		✓		✓	
S7	✓				
S8		✓			
S9	✓	✓			
S10	✓				
S11	✓	✓			

Preference for using COCA in the Future

The participant responded differently when asked if they want to use COCA for learning English in the future. Four students (S2, S4, S9, and S11) agreed they would continue to utilize the tool to check collocations and synonyms, three of them (S4, S9, and S11) even complimented that COCA's detailed and real-life contexts of the searched words were what they could learn and apply in their writing. The other seven students, however, preferred not to use the tool due to some reasons. Two students (S10 and S5) confirmed that they would not use COCA, while S1 considered using it only when being under no time pressure to revise an assignment. Others (S7 and S8) said they used the tool when the layout of information was improved, whereas S3 said he might employ COCA if he could not locate specific information about a word in other sources. Besides, S6 believed that only those who researched linguistics would find the tool useful, thus it was not for students learning English to prepare for exams. Finally, S7 pointed out that it was the lack of adequate facilities at university that hindered students from exploiting COCA, so it would be more preferable if the tool could be developed as an application to be downloaded to each student's personal device.

5 Discussion

5.1 Changes in MTLTD

There were only 8 out of 11 participants who maintained using COCA for proofreading their writing, therefore changes in MTLTD can only be obtained from these 8 participants. The limited volume of valid data appeared to reveal the time-consuming feature of COCA, which is confirmed in (Römer, 2011) and Yusu (2014). To explain this, time pressure might be the factor that discouraged COCA application. In fact, some of the participants admitted only focusing on the assignment deadline and forgetting to use the corpora [S2.R1], or procrastinating so close to the deadline that they prioritized completing the assignment [S11.R3] as COCA consultation would take more time.

Also, although previous studies claim that students reap some benefits from COCA with their vocabulary (Yan, 2012; Yusu, 2014), the changes in students' MTLTD in this study are not a uniform trend and may not

result from the use of COCA only. Some participants' results were on an upward trend while others experienced the opposite. These changes may also be impacted by the students' using other reference sources such as collocation, bilingual or synonym dictionary. Some students also revealed they combined COCA and other sources when revising their writing. As the writing assignments are a part of the participants' formative assessment, it is risky if they are asked to exclude other sources while the positive effect of COCA is yet to be confirmed in all cases. Therefore, the changes in MTLTD may be the outcome of students' combining both COCA and other resources, not solely from the use of COCA itself.

5.2 Participants' Perceptions of COCA

5.2.1 Lexical Aspect

One result comparable to the findings of Albushar & Albuseileek (2013) and Yan (2012) is the students' positive feedback for COCA's word meanings, collocations, and synonyms. This corroborates the comment that corpora assist students enhance their word knowledge, learn more meanings for a word, and understand how words collocate with each other or be replaced for greater lexical diversity in writing. The contribution of this study is discovering that thanks to COCA consultation, students can not only diversify their lexicon but also improve the difficulty level of the words by using less frequent words at higher levels C1-C2 (CEFR).

However, it is difficult to conclude if the students will be able to successfully use the words and phrases obtained from COCA into their written production because they may ignore the context. In this study most participants did not mention using COCA for checking word context, also an example in S4's reflection showed that although she had consulted COCA, the word she selected from the tool to replace the one used in her first version was not appropriate. Specifically, S4 replaced the word "refund" with "reimburse":

Table 5 *Difference in meanings of "refund" and "reimburse"*

First version	Revised version
Hurrah will make up for Mr Danson and Mr MacKenzie by refunding unexpected expenditure.	Hurrah will make up for Mr Danson and Mr MacKenzie by reimburse the unexpected expenditure.

[extracted from S4.R1]

refund	to give somebody an amount of money especially because that person is not happy with a product or service they have bought
reimburse	to pay back money to someone who has spent it for you or lost it because of you

The two words "refund" and "reimburse" have different connotative meanings, according to the Online Cambridge Dictionary (Table 5). S4's writing was originally a report to a customer service department detailing a problem and proposing a solution to satisfy the customer. The word "reimburse" does not match the context of S4's writing, hence it should not be used to replace "refund". This example verifies the study by Wu et al. (2010). According to Wu et al. (2010, p.99), one reason for this inappropriateness is students change words before or after the keyword but forget to check the whole context when consulting the corpora.

Also, the participants' high appreciation for the C1-C2 level words is partly because these words may help them achieve more fruitful grades for their writing assignments or tests. These high-level words, however, are not complimented by Yusu (2014) as they pose a challenge for low-level students or even lead to possible demotivation. However, this scenario is not present in this study, as none of the participants mentioned having difficulty understanding the results on COCA. Given these participants' proficiencies are at the B1 level and no reports on difficult language levels are recorded, the study suggests that the minimum level of users to process and benefit from COCA's lexis is B1.

5.2.2 Other Aspects

As can be seen from the findings, the participants' feedbacks on COCA's interface and contents reinforce and contribute further details to the literature. The overwhelming amount of information has been mentioned in Römer (2011, p. 214) and Yusu (2014, p.69). Under the effect of the layout of information, font size, text colours, and highlighting effects, this amount appeared to be more difficult to proceed. One possible explanation for the students' critical accounts is that they are more accustomed with traditional dictionaries whose interfaces are simpler, information are filtered and examples have been adjusted. Therefore, it is understandable that they felt overwhelmed and even irritated when confronted with a large amount of information shown after a single click. This feeling also obstructed them from reading the lines with adequate meticulousness.

Also, participants' reflections echo an idea related to ease of use mentioned in the literature. According to Römer (2011) and Yusu (2014), students may find it inconvenient to use a new tool or work with computers, because this requires time to get familiar with. Similarly, some participants in this study simply stated that COCA was "difficult to use" without further explanation. One possible reason for this comment is that the participants generally are not comfortable using computer or any new tool, plus their reflections were written several days/weeks after they had stopped using COCA. Therefore, they forget some details of the difficulty they experienced, leaving them with a general impression that COCA was "difficult to use". This suggests that participants need more careful training not only before but also during their usage, especially on how to examine the contexts and focus on important information. The training can be best maintained by letting students use the tool in class and the lecturer can provide instant help for any problems encountered, but this can only be possible if classroom facilities are adequate with computers and on-site stable Internet connection. These conditions were not accessible to the participants of this study, which ratifies the viewpoints of Yusu (2014, p.69). The suggestion is the university offers suitable infrastructure so that the integration of COCA and other corpora will be more successful, and hopefully avoid issues that hinder students' utilization of the tool to revise their lexicon.

6 Conclusion

This study investigates how COCA assists learners with productive vocabulary in Writing. The findings suggest that for EFL learners whose English proficiency is of B1 level or above, COCA can be useful to help them develop their knowledge about words, as well as usage of collocations, synonyms, and high-lexis despite some inappropriateness with regards to context. However, an unfamiliar interface and overwhelming contents may hinder ease of use. This implies more proper training and upgraded facilities needed in their learning environment.

For future research, a higher number of participants and research contexts which enable participants to use the tool in class under stricter supervision of the researchers/lecturers is needed to yield a result with higher reliability and generalizability.

References

- Alshaar, A. A., & AbuSeileek, A. F. (2013). Using Concordancing and Word Processing to Improve EFL Graduate Students' Written English. *JALT CALL Journal*, 9(1), 59-77.
- Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. 2006 London. UK SAGE.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.
- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4), 447-464.
- Davies, Mark. (2008). The Corpus of Contemporary American English (COCA). Available online at <https://www.english-corpora.org/coca/>.
- Friedman, D. A. (2011). How to Collect and Analyze Qualitative Data Research Methods in Second Language Acquisition (pp. 180-200): John Wiley & Sons, Ltd.
- Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching. *The Routledge handbook of corpus linguistics*, 359-370.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine Pub. Co.
- Koizumi, R. (2012). Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens? *Vocabulary Learning and Instruction*, 1(1), 60-69.
- Lawrence, J., & Tar, U. (2013). The use of grounded theory technique as a practical tool for qualitative data collection and analysis. *The Electronic Journal of Business Research Methods*, 11(1), 29-40.
- Lewandowska, A. (2014). Using corpus-based classroom activities to enhance learner autonomy. *Konińskie Studia Językowe*, 2(3), 237-255.
- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*: Cambridge University Press.
- O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge handbook of corpus linguistics*. Routledge.
- Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 31, 205.
- 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), 147-162.
- Wallis, Sean (2021). *Statistics in Corpus Linguistics Research – A New Approach*, New York, London: Routledge.
- Wu, S., Witten, I. H., & Franken, M. (2010). Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge. *ReCALL*, 22(01), 83-102.
- Yan, Z. (2012). Using Corpus Technology to Assist in Vocabulary Acquisition of ESL Learners. Retrieved on September 4th, 2021 from http://cblle.tufs.ac.jp/assets/files/publications/working_papers_08/section/389-396.pdf
- Yusu, X. (2014). On the Application of Corpus of Contemporary American English in Vocabulary Instruction. *International Education Studies*, 7(8), 68.