TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
TẠP CHÍ KHOA HỌC
KHOA HỌC GIÁO DỤC
ISSN: 1859-3100 Tập 16, Số 8 (2019): 275-300
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
EDUCATION SCIENCE
Vol. 16, No. 8 (2019): 275-300
Email: tapchikhoahoc@hcmue.edu.vn; Website: http://tckh.hcmue.edu.vn

**Research Article**

# USING CORPORA TO TEACH COLLOCATIONS IN A UNIVERSITY CONTEXT

***Nguyen Thi Thanh Huyen***
*Faculty of English – Hanoi National University of Education*
*Corresponding author: Nguyen Thi Thanh Huyen – Email: nguyenthanhhuyen2111@gmail.com*
*Received: April 18, 2019; Revised: July 13, 2019; Accepted: July 18, 2019*

**ABSTRACT**

*English as Foreign Language students tend to learn vocabulary in word isolation, not in chunks or collocations which produces meager results in students' collocational competence and lexical resources. In addition, a corpus-assisted method is used in this project because of its significant effectiveness in bringing real-world language use or authentic materials in teaching and learning collocations. Therefore, this article investigates the potential role of using corpora and concordances in teaching and learning collocations with a view to improving university students' collocational competence in academic writing. To do this, an experiment was conducted among 30 third-year students in the English Faculty of Hanoi National University (pseudonym) who had little or no previous knowledge of collocations as well as corpora. Students were in both the experimental group in a six-week English unit which a corpus-assisted method was applied for the experimental group and a traditional (or rule-based) method was used for the control group to find out the differences and improvement among groups of students. They were required to take part in different tests in different time periods including before, immediately after and two weeks after the course. The results of these tests were analyzed carefully in terms of learners' collocational use in academic writing, specifically premodifier-noun collocations. Results indicated that while both groups experienced improvements in their academic writing skill, the students of the experimental group displayed a holistic improvement regarding the use of collocations with fewer collocational errors and more academic collocation patterns. It is, hence, concluded that the application of corpora exerts a tremendous influence on developing learners' collocational competence as well as language proficiency.*

***Keywords:*** Corpus (Corpora), corpus-based (corpus-assisted), collocations, lexical approach.

## 1. Collocation

### *1.1. Definition of a collocation*

***Cite this article as:*** Nguyen Thi Thanh Huyen (2019). Using corpora to teach collocations in a university context. *Ho Chi Minh City University of Education Journal of Science, 16*(8), 275-300.

It is commonly believed that collocation has been of paramount importance in the field of language in recent years and exerts a tremendous influence of learners' collocational competence. According to Lewis et al (1997), "collocation forms a central feature of a lexical view of language and noticing collocation is a central pedagogical activity". It is worth being paid more attention to as "language knowledge is collocational knowledge" and "all frequent and appropriate language use requires collocational knowledge" (Nation, 2000b). So, what is collocation? Among linguists and educators, what is called "collocation" still remains controversies and it sparks two opposite views. On the one hand, collocation is considered "the co-occurrence of words at a certain distance, and the distinction is usually made between co-occurrences that are frequent and those that are not" (Nesselhauf, 2004). This view, as a result, has been called "frequency-based approach" or "statistically oriented approach". Firth - a father of chunks and collocations – shared the same opinion with this latter view and defined collocation as "the company words keep their relationships with other words and it is the way words combine in predictable way" (as cited in Lewis & Conzett, 2002). He argued that "the meaning of a word is as much a matter of how it combines with other words in actual use as it is of the meaning it possesses in itself" (O'Keeffe, McCarthy, & Carter, 2007a). To put it simply, when it comes to collocation, it can be understood as "two or more words that tend to occur together (collocate)" (Lewis & Conzett, 2002) which means the way one word frequently comes together with other words for no specific reasons.

In terms of classification of collocations, there still remains quite a few different ways to divide collocations. To Lewis's way of thinking (2000), he classified collocations into four main groups: *unique collocations, strong collocations, weak collocations* and *medium-strength collocations*. According to Hill, he believed that:

> …the main learning load for all language users is not at the strong or weak ends of the collocational spectrum, but in the middle -those many thousands of collocations which make up a large part of what we say and write.
>
> (2000, as cited in Michael Lewis & Conzett, 2002, p. 64)

Medium-strength collocations are one that each individual word may be known to language learners, but they probably may not acknowledge the whole collocation and are likely to express their thoughts word by word or phrase by phrase. For example, most learners can know the meaning of each single word "hold" and "conversation", however, they may not know that they can express a collocation as "hold a conversation" due to their lack of collocational competence. They may express their sayings in an unprecise way like "keep a conversation" or "maintain a conversation" or so on, which means that the collocation "hold a conversation" is not stored as a single item in their mental lexicons and they may make some mistakes related to collocations. Thus, it is understandable why

medium-strength collocations are of prime significance in expanding learner's mental lexicons as well as collocational competence. And one question about the reason why educators or teachers need to know about the classification of collocations, especially collocational strength, is put forward.

In terms of **premodifier-noun collocations**, based on the definition in the Oxford dictionary, they are defined as a combination of a premodifier and a noun to form a collocation. Premodifier is a word, especially an adjective or noun, which is placed before a noun and describes or restricts the meaning of that noun in some way. Thus, the premodifier-noun collocations can be easily understood as a "noun phrase" which combines a noun or an adjective and a noun and they express a complete meaning. There are two main types of premodifier-noun collocations and they are classified based on whether the premodifier is an adjective or a noun. For instance, "reasonable price" is considered as a premodifier-noun collocation as it is formed by a combination of an adjective "reasonable" and a noun "price" to express a fixed meaning in terms of price. Or "job orientation" is also considered as a premodifier- noun collocation because this collocation consists of two main parts, namely a noun "job" and another noun "orientation". The reason why the researcher decided to choose premodifier-noun collocations for deeper research is that they are commonly used in many authentic texts.

### *1.2. The importance of collocations*

As in aforementioned parts, it is obviously undeniable that collocations play a pivotal role in the pedagogical field and there are a host of reasons below which answer the question why I decided to opt collocations as a core for my research thesis.

The first and foremost obvious reason is that the lexicon is not arbitrary and "the way words combine in collocations is fundamental to all language use" (Lewis & Conzett, 2002). Firth (1951) emphasized that collocations are not arbitrary word combinations which are frequently uttered by native speakers whereas other combinations which share the same expression, meaning and equal grammatical point are not accepted (Nation, 2000b). The second reason worth mentioning in terms of the importance of collocations is fluency. It is clearly evident that collocations have a considerable bearing on the fluency of learners in all four skills including Speaking, Writing, Listening and Reading as they help learners constantly recognize multi-word units rather than process every speech or text word by word, which is time consuming and has an adverse effect on learners' time processing. This merit of collocations is also advocated by Nation who shared the same opinions about time processing related to learning collations. He stressed that:

> The main advantage of collocations is reduced processing time. That is, speed. Instead of having to give a close attention to each part, collocation is seen as a unit which represents a saving in time needed to recognize or produce the item… it is treated as a basic existing unit.
>
> (Nation, 2007, p. 520)

So, it is easily seen that collocations treated as a unit can support learners considerably in reducing the processing time and learners tend to be able to think faster and more accurately. It is proved that "collocation allows us to name complex ideas quickly so that we can continue to manipulate the ideas without using all our brain space to focus on the form of words" (Lewis & Conzett, 2002, p. 55). Even advanced students are not likely to become more fluent by giving more chances to be fluent. As a result, it is undeniable that "collocation is an important key to fluency" (Nation, 2000b) and collocation has a tremendous influence on learners' language proficiency. Another reason supporting for the importance of collocations is that complex ideas are often expressed lexically. Thus, collocation should be treated equally as an important factor contributing considerably and majorly to language learners' collocational competence as well as language proficiency.

### *1.3. Collocations and teaching*

There is no doubt that collocations are important building blocks and have an inextricable relationship with language teaching. To illustrate obviously this point, a criterion called "learning burden" is given for deeper understanding. "Learning burden" is learner's effort to learn vocabulary; thus, in order to reduce learning burden for language learners, teachers had better "pay attention to the systematic patterns and analogies within the second language and point out the connection between the second and first language" (Nation, 2000a). The principle of learning burden applies just as much to collocation as it does to individual words. It is widely acknowledged that "its learning burden is light if it follows regular predictable pattern" (Nation, 2000a).

In terms of pedagogic value of collocation, from the observations of noticing, recording and learning, there are two main crucial values for teaching language. On top of that is that words "are not normally used alone and it makes sense to learn them in a strong, frequent, or otherwise typical pattern of actual use" (Lewis & Gough, 1997). Additionally, collocation is more efficient to learn the whole and break it into parts, than to learn the parts and have to learn the whole as extra arbitrary item. Thus, it can be easily seen that collocations have a considerable bearing on teaching and learning language.

## 2. Literature review

### *2.1. The lexical approach*

The Lexical approach is discussed in this section since it is considered as a theoretical framework for teaching vocabulary in general and collocations in particular. It has emerged and officially introduced since 1993 by Lewis, which stimulated wide and lively debates among linguistics and educators all over the world. An enormous number of colleagues have written with queries, disagreements, support and practical suggestions for taking this approach in the classroom. The standard norm dictates that language is divided into "grammar" (structure) and "vocabulary" (words), which are separately taught and transcended to the language learners. As can be easily seen, most of the teachers, at that time, advocated for the former and laid strong emphasis on teaching grammar only. Vietnam is a case in point. It is undeniable that a host of Vietnamese teachers paid too much attention to teaching grammar and ones who were good at grammar were considered as talent students in learning English. That was a preconceived notion that needed bettering radically and positively. With that being said, the Lexical approach challenges this fundamental view of language and argues that "language consists of chunks which produces continuous coherent text when combined" (Lewis & Gough, 1997).

### *2.2. The relationship between corpus linguistics and language teaching*

An indeed important feature that needs taking into consideration in this field is the correlation between corpus linguistics (CL) and language teaching (LT). Over the past two decades, "the contribution of corpus linguistics to the description of the language we teach is difficult to dispute" (O'Keeffe, McCarthy, & Carter, 2007b). Corpora, definitely, have brought to light features about language which had eluded our intuition. So, the significant use of corpus has recommended a host of pedagogical corpus applications.

Looking at the Figure 1 "The relationship between corpus linguistics and language teaching", it is obviously seen that there is an indispensable relationship between corpus linguistics and language teaching. On the one hand, the CL provides many resources, methods and insights for the LT which are very useful in the context of language pedagogy. On the other hand, the LT gives needs-driven impulses to CL, which is of great significance. Moreover, when it comes to types of pedagogical corpus applications, a useful distinction can be made between "direct" and "indirect" applications depending on who and what is affected by the use of corpus methods and tools.

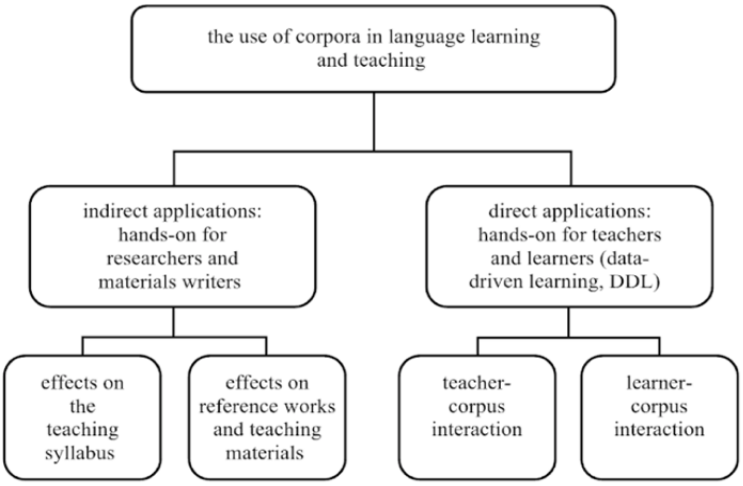


***Figure 1.*** *The relationship between corpus linguistics (CL) and language teaching (LT)*

It is evident that two types of corpus applications are absolutely different to each other and each type includes their own features as illustrated in Figure 2 "Applications of corpora in language teaching". As compared to indirect ones which lay an emphasis on the impact of corpus evidence on syllabus design or teaching materials and is concerned with corpus access by researchers and material designers, the direct ones focus more on the teacher-corpus and learner-corpus interactions so they are more suitable to teachers and learners in the language classroom. This tends to facilitate learners opportunities of being "linguistic researchers" (Gavioli, 2006, as cited in Lüdeling & Kytö, 2009).

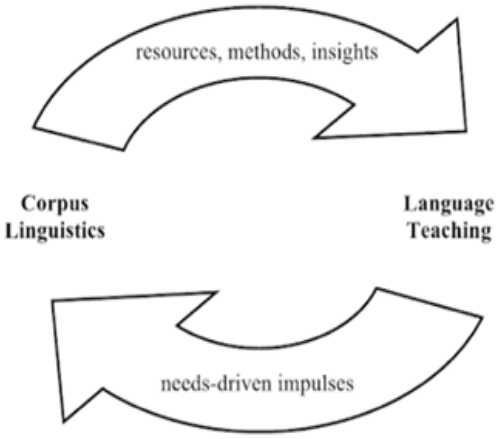


***Figure 2.*** *Applications of corpora in language teaching*

In terms of advantages of corpora related to pedagogic view, it is obviously evident that corpora "have changed the way we look at language and, for teachers at least, the way we see our own role" (Hunston, 2010). As new concepts such as the "unit of meaning" are dependent on the availability of large quantities of language which can be manipulated electronically. And a corpus gives learners not only definitions and a few examples like ordinary dictionaries, but samples of concordance lines which facilitate learners deeper understanding of lexis. So, the relationship between corpus linguistics and language teaching, as of late, has been inextricable and needs more attentions from language teachers as well as researchers. Language teachers should pay attention to the application of corpus linguistics in language teaching as it "supports the use of examples of real language in classroom" and "corpus data can provide language teachers and learners with illuminating guidance as to frequent collocations" (Reppen, 2010). Regarding the historical background of the application of corpus linguistics in the pedagogic field, there is no lack of corpus-assisted research informing the teaching of collocations, but many of them focus on an indirect application of corpora in classroom settings. As mentioned above, the indirect application means that it is handed on for material writers or researchers for syllabus design or collocation dictionaries, not for teachers and learners in the classroom. A host of research and materials associated to indirect application has been carried out by Chen (2013) or McGee (2012) who paid a lot of attention to develop materials of collocations and chunks. So, the result of implementing corpus-assisted method for the effect seems to be less positive, for it cannot reach the "deeper layers" or, in fact, "teachers" and "learners". On the flip side, it is quite rare to observe the direct application of corpora in the language classroom to develop learners' collocational competence because it tends to challenge both learners and teachers with some possible hinders. Although corpora are universally acknowledged to be a valuable resource in describing language, "there is less consensus on the value of corpus findings in the description of language for learners or on the use of corpus-based material in language classrooms" (Hunston, 2010). As cited in "Corpora in Applied Linguistics", Hunston (2010), Widdowson (2000) and Cook (1998) posed several challenges when an direct application of corpora use in language classrooms.

Despite those aforementioned obstacles, the direct application of corpus on language classrooms facilitates a wide range of merits to both language teachers and learners. There exists a variety of research and studies that have experimented the direct use of corpus associated to teaching collocations in language classrooms. Ly (2017) has demonstrated the effectiveness of corpus application in teaching verb-preposition collocations among Chinese postgraduates and the findings revealed that one group of learners who had intense exposure to corpus application showed better in writing essays with perfect related

collocations and they could even remember these collocations for a longer time than the other group who learned collocations in a traditional method. Rafael (2009) shared the same idea with Ly when he implemented a research to test the effectiveness of corpus-assisted method in teaching collocations among EFL students. He realized that using corpora helped students get better awareness of collocations and they could hold their memory about collocations for a finite period. Moreover, the result of his research also reported that learning collocations through corpora facilitated his students' potential to communicate better in daily conversations. With the principles of data-driven learning (DDL), McEnery & Wilson (2011) argued that the lexical approach with a data-driven corpus-based methodology in language teaching "can enrich the learners' language experience and raise their language awareness while bringing out the researcher in them". Or in another study, Varley (2009) indicated that his students had a positive response to corpus consultation in teaching collocations and syntactic patterns, which contributed to the significant role of corpus-based method on teaching and learning vocabulary. Faghih and Sharafi (2006) shared the same opinion in his research on the role of collocations on Iranian language EFL leaners' interlanguage. They strongly pointed out that most of errors that learners made in their tests were rooted in their deficiency of collocational knowledge and this raised an alarming bell for learning collocations to improve their mind. Similarly, Lüdeling & Kytö (2009) demonstrated that the adoption of a web-based collocational concordance promoted the learners' ability of using collocations correctly in a writing course. Thus, it can be easily seen that there is a flaw in those aforementioned researches, which means that the real effectiveness of using corpora in language classrooms is not definitely embedded for a long-term period. And my thesis, to some extent, will fill this gap to explore the feasibility of incorporating direct application of corpora into a curriculum to teach collocations, especially on a long-term process.

## 3. Research Methodology

The main purpose of this article is to investigate the positive role of corpus application in EFL learners' collocational competence in academic writing. There are two main primary research questions proposed to serve this purpose:

- *How does the corpus-assisted method used in teaching and learning premodifier-noun collocations?*
- *How does the corpus-assisted method promote learner's development of premodifier-noun collocational competence in academic writing?*

With a view to answering these two questions, an experimental design – a traditional approach to conducting quantitative research – is implemented. Regarding definition of this case study, an "experimental design" can be easily acknowledged as an idea (or practice or procedure) which is tested to determine whether it influences an outcome or

dependent variable. The researcher has to decide on the first idea which to "experiment", assign individuals to experience it, and then determine whether those who experienced the idea (or practice or procedure) performed better on some outcome than those who did not experience it (Creswell, 2012).

The underlying reason why the author decided to opt an experimental design for this research is justifiable. In this experiment, main methods of teaching vocabulary for university students are desired to be differentiated, namely the traditional and the corpus-based one; and then are compared in terms of teaching effectiveness and students' collocational competence. This means that the author attempted to control all the variables that influence the outcome except for the independent variables. Moreover, experiments are highly controlled, so they are the best of quantitative designs to use to establish probable cause and effect. Experimental design creates a favorable condition for the researcher to control all the variables that might influence the outcome except for the difference in types of teaching (traditional or corpus-based method). By comparing and contrasting two groups (experimental and control group) with the same condition and same time period, the author found it convenient to find out the result and draw a conclusion about students' collocational competence in academic writing.

One more thing should be laid emphasis on is that there are two different types in the experimental design, including "true experiment" and "quasi experiment". In my thesis, "*quasi experiment*" was chosen as it includes assignments, but not random assignments of participants to groups. Before considering how to conduct an experiment, it is of paramount importance to understand in more depth several key ideas central to experimental research. These key characteristics exert a tremendous influence on the author's decision of choosing experimental design as a method for this article. Not only do they contribute to the author's way of thinking about different steps but also play a crucial role as a "frame" for accessing criteria in this thesis, including random assignment, pretests and posttests, group comparisons and threats to validity.

### *3.1. Overview of research procedure*

In this thesis, an experimental research was conducted to investigate the effectiveness of teaching collocations based on corpus with a view to developing the EFL learners' collocational competence. This research was carried out between two groups of third-year students at English Faculty of Hanoi National University (pseudonym) who had no or little previous knowledge of corpora and collocations; and they are called "the experimental group" and "the control group". Both groups were required to complete a course in linguistics lasting for six continuous weeks, with the former using the corpus-based method and the latter using the traditional (or rule-based) one. The skill tested was writing and the chosen topic for this study was "Health". The English essays written by

both groups from different time periods (before, immediately after and two weeks after the course) were collected and analyzed in terms of the use of premodifier-noun collocations. In the following parts, the detail information about the participants, the different phases they took part in, the data used for analysis and the procedure for carrying out the research is mentioned and discussed.

### *3.2. Participants and different phases of the research procedure*

#### *3.2.1. Participants*

In this experimental study, the participants are 30 Vietnamese sophomores in English Faculty of Hanoi National University who have no or little previous knowledge of corpus. They are all majoring in English linguistics and their main subjects at university are Reading, Listening, Writing and Speaking. According to the language frame of CEFR, their current level of language ability is at around B1 level and their target in this semester is B2. It seems evident that all selected participants possess a basic knowledge background in terms of grammar and practice skills (as they could pass the university entrance exam of Ministry of Education and Training last year); however, what renders them from achieving higher level (B2 level) is that they cannot upgrade their use of lexical resources, especially collocations or chunks.

#### *3.2.2. Different phases of the research procedure*

In order to carry out more effectively, the researcher divided this research procedure into three main phases.

- ***Phases 1:***

The first phase (Phase 1) was the pre-test for all students for group classification. They were required to take part in a writing mini-test (an around 200-word essay on the given topic) to assess their entrance level. This test was compulsory for all the participants as it was the best way to evaluate the initial level of each participants and the writing test marking was based on the assessment criteria (see Appendix A). Finally, based on their writing performance, 15 students were assigned to the experimental group and the other 15 to the control group. This initial assessment helped to make sure that the average level of participants in each group were quite similar and balanced.

- ***Phase 2:***

In the second phase (Phase 2), after classifying all the participants, two different six-week courses were applied into two groups. The former was introduced and taught about corpus and the corpus-assisted method, while the latter learnt the traditional method without an introduction of corpus with a rule-based style. The main purpose of this course is to develop students' ability in language analysis and their English language proficiency. At the same time, 10 articles and texts about the topic "Health" were collected and given to

students for analysis during this course. Most of the articles are academic ones and were collected from several reliable websites such as the Guardian, the Medium or BBC News. They were all converted to plain texts and put on a corpus named "***Health Articles***". However, one problem arises was how the researcher acknowledges of whether one collocation selected from the corpus is the strongest and the most certain one or not before introducing them to the whole participants. To answer this question, the ***Mutual Information Score*** (***MI-score***) and ***t-score*** were calculated carefully with detail formulas in order to give the precise strength and certainty of each collocation in ten selected articles.

**MI score**: An MI-score measures the amount of non-randomness present when two words occur. It is a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of the two words. An MI-score of 3 or higher can be taken to be significant.

The MI-score is the Observed divided by the Expected, converted to a base-2 logarithm:

$$\log_2 \frac{f_{AB}N}{f_A f_B}$$

**t-score**: t-score reveals the certainty of a collocation which is calculated by subtracting the Expected from the Observed and dividing the result by the standard deviation. A t-score of 2 or higher is normally taken to be significant.

$$\frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

In which:

*N = Corpus size.*

*fA = Number of occurrences of the keyword in the whole corpus (the size of concordance)*

*fB = Number of occurrences of the collocate in the whole corpus*

*fAB = Number of occurrences of the collocate in the concordance (number of co-occurrences)*

The important differences between *MI-score* and *t-score* is that while the former is a measure of strength of collocation, the latter is a measure of certainty of collocation. It is obvious that the value of an MI-score is not particularly dependent on the size of the corpus. However, for the t-score, corpus size is important as the amount of evidence is being taken into account. Thus, MI-scores can be compared across corpora, even if the

corpora are of different sizes, but t-scores cannot be compared across corpora because the size of the corpus will have effect on t-score (Hunston, 2010).

All steps from how to calculate the MI-score and t-score, and how to see all of the most frequent adjective collocations in the corpus "Health articles" were implemented thanks to the application named ***Sketch Engine*** (sketchengine.eu). Sketch Engine is a tool for discovering how language works which helps the learners or researchers easily discover what is typical or frequent in the language. It has many tools to identify and analyze collocations, especially frequency word lists of English single words or multi-word expressions of various types can be generated, which is of great significance in this thesis. So, it is justifiable that the researcher could generate a list of the most frequent words (including "*premodifier* + *noun*" as this thesis aspired to adjective-noun collocations only); and then calculated the MI-score and t-score to make a decision of which collocations should be selected from the given list. Figure 3 is a list of top twenty frequent multi-words generated from Sketch Engine. The reason why the researcher chose multi-words instead of single ones as it created more opportunity to identify collocations in the whole ten selected articles.

MULTI-WORDS

| | Word | Focus corpus | Reference corpus |
|---|---|---|---|
| 1 | diet culture | 5 | 0 |
| 2 | body image dissatisfaction | 4 | 1 |
| 3 | image dissatisfaction | 4 | 1 |
| 4 | counterfeit food | 4 | 2 |
| 5 | s sleep | 4 | 2 |
| 6 | clean eating | 4 | 53 |
| 7 | diet soda | 4 | 60 |
| 8 | non-dieting eating | 3 | 0 |
| 9 | body-mass index | 3 | 6 |
| 10 | melatonin production | 3 | 6 |
| 11 | psychological well-being | 3 | 32 |
| 12 | cheat day | 3 | 39 |
| 13 | sleep quality | 3 | 44 |
| 14 | body image | 6 | 373 |
| 15 | xenophobic violence | 2 | 0 |
| 16 | pre-pregnancy size | 2 | 4 |
| 17 | soda tax | 2 | 4 |
| 18 | diet industry | 2 | 5 |
| 19 | psychological wellbeing | 2 | 5 |
| 20 | expired food | 2 | 7 |

***Figure 3.*** *Top 20 frequent multi-words generated from Sketch Engine*

After creating a list of top frequent multi-words, some collocations from the above list were eliminated as they are either meaningless (such as number 5) or too terminological (such as number 10 and 15). At the same time, some were added for score calculation as they are quite ubiquitous and easy to apply in academic writing. Then, MI-score and t-score for each collocation from the above list were calculated carefully for more detail selection. All the indexes are illustrated in the Table 1.

***Table 1.*** *Statistics (MI-score and t-score) for each collocation*

| *Number* | *Collocation* | *MI-score* | *t-score* |
|---|---|---|---|
| 1 | Diet culture | 2.18 | 3.02 |
| 2 | Image dissatisfaction | 1.76 | 3.89 |
| 3 | Counterfeit food | 4.98 | 5.73 |
| 4 | Clean eating | 1.2 | 3.67 |
| 5 | Diet soda | 1.14 | 2.34 |
| 6 | Non-dieting eating | 2.85 | 4.67 |
| 7 | Body-mass index | 4.53 | 6.52 |
| 8 | Healthy diet | 4.82 | 2.13 |
| 9 | Ultra-processed food | 5.75 | 7.45 |
| 10 | Psychological well-being | 3.11 | 3.42 |
| 11 | Sleep quality | 1.17 | 2.12 |
| 12 | Soda tax | 1.01 | 1.9 |
| 13 | Diet industry | 2.19 | 1.65 |
| 14 | Mentally taxing | 4.21 | 6.25 |
| 15 | Expired food | 3.89 | 5.12 |

The next step after calculating the MI-score and t-score for each selected collocation (as can be seen in table 1) was choosing which collocations worth introducing for participants in the whole corpus. According to the aforementioned part, a collocation which has the MI-score of 3 or higher means strong one. Similarity, a collocation which has the t-score of 2 or higher means certain one. Based on the calculated statistics, there were some chosen collocations, namely "*diet culture, image dissatisfaction, counterfeit food, body-mass index, healthy diet, ultra-processed food, psychological well-being, mentally taxing and expired food*". Having said that, this list was used as reference, and if there is any collocation arising during the process of running the corpus, the researcher will note down and calculate these two mentioned scores like this.

In terms of *the experimental group*, a short explanation about what corpus linguistics is was introduced before they jumped into the main part of the course: using corpora to discover and learn collocations. For this group, the researcher decided to use ***LANCSBOX 4.0*** application which is one of the latest one in corpus linguistics recommended by a host

of educators in one of the most reliable learning websites named *Futurelearn.com*. This application contains many useful and convenient tools for both teachers and students to learn collocations such as ***KWIC*** (***Key word in context***) and ***Graph Coll***. The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance and includes many concordances which are "relatively simple piece of computer software which allows a constructive search of large amounts of text of a particular words or phrases" (Micheal Lewis, 2006). For example, with the topic "Health", all the selected articles were put in a corpus named "Health" and then let it run for a few minutes. After that, by using KWIC tool, the key word "food" (for example) was searched and all the concordances with the word "food" appeared after a mouse click.


As can be seen from the picture, the search word "foods" is placed in the middle of the page where it is easily recognized. Moreover, there is only a single line of text is listed for each example and these are usually not complete sentences. Students can easily acknowledge many adjective-noun collocations with the word "foods" from 10 selected articles such as "counterfeit foods", "junk foods", "whole foods" or "highly/ultra-processed foods" accompanied with detailed contexts. Thus, it can be concluded that a corpus with concordances provide much help richer sources of co-textual information than

dictionaries and "they can lead to a more efficient exploration of the collocates of a word" (Lewis, 2006).

On the flip side, regarding *the conventional group*, a traditional method (or a rule-based method) was applied to teach vocabulary items, particularly collocations. First, ten articles were handed on for all 15 students in this group and then what they had to do was skimming and scanning all the texts to find new vocabulary items associated to the topic "Health". They noted down and accessed to dictionaries with a more-centered approach and no corpus use.

One more thing should be paid attention is that written informed consent was obtained from all the participants who kindly allowed their essays to be used for research purposes. Moreover, they were informed that these essays would be a means to monitor their progress in English academic writing for the mid-term test.

- ***Phase 3:***

The last but not least phase is the third one (Phase 3), which was carried out 2 weeks after finishing the course. In this phase, all the students from two groups were required to write an approximate 300-word essay about a chosen topic to evaluate again their collocational competence in writing skill and how many percent they could remember all the collocations they had learned a few weeks ago. All the essays were collected and put into Lancsbox application for corpus analysis.

### *3.3. Data analysis:*

In the aforementioned part, each participant is requested to write three essays in different time periods (before, immediately after and two weeks after the course). For each essay, the students were instructed to write on a specific topic for around 200-300 words. In addition, they are allowed to access any tools or materials they use in the course, namely corpora for the experimental group and dictionaries or other learning materials for the control one.

Next, 90 essays were processed to anonymize participants' personal information and then tagged in terms of part of speech (POS). The corpus is referred to as "the Corpus of Students' Essays" (COSE), consisting of 12780 tokens. Furthermore, this COSE was divided into six sub-corpora to distinguish texts from different groups and different time periods (Figure 4). All the mentioned sub-corpora were analyzed in details in LancsBox 4.0 application. The abbreviations in the figure 4 "The construction of the corpus for analysis", "EG" and "CG" stand for "Experimental group" and "Control group" respectively.

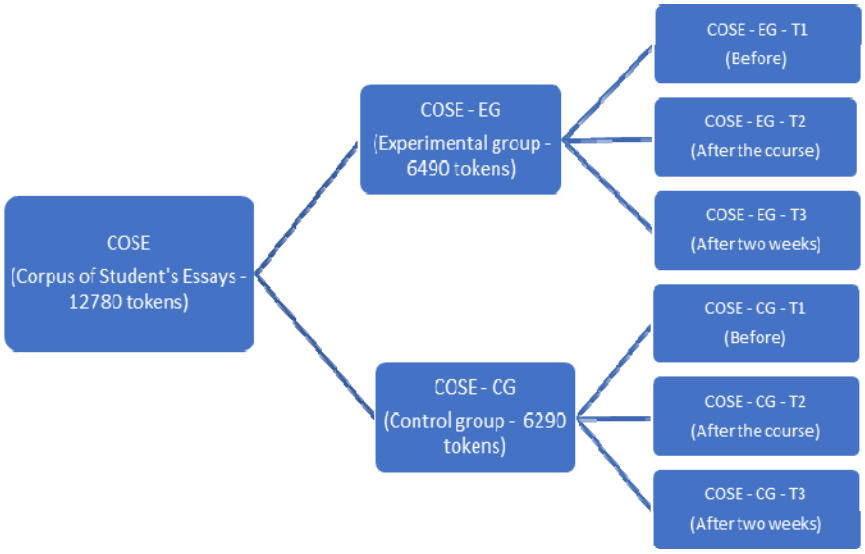


***Figure 4.*** *The construction of the corpus for analysis*

After building the corpus construction for analysis, four main indexes were chosen and put emphasis on for the key criteria with a view to comparing the differences between two groups at different time periods.

- ***Type-token ratio (TTR)***

"Token" is an academic term for any single, particular instance of an individual word in a text or corpus; as compared to "type" which is a single particular wordform. Any difference of form makes a word a different type. All tokens comprising the same characters are considered to be examples of the same type. The type-token ratio (TTR) is a measure of vocabulary diversity in a corpus, equal to the total number of types divided by the total number of tokens. The closer the ratio is to 1 (or 100%), the more varied the vocabulary is.

- ***Academic Word List index (AWL index)***

AWL is carefully and rapidly calculated on website *Lextutor.ca*, and the researcher can recognize how many percent of academic words ranged from above B1 level students can embrace in their writings in different periods. Additionally, the researcher is able to compare the ratio of academic words used in student's writings in different groups and phases; and then draw a conclusion whether students make improvement on collocational competence or not. The higher AWL index, it is clearly evident that the more improved students achieved.

- **Lexical Density index (LD index)**

LD is a useful measure of the difference between texts or corpora. In order to calculate the LD index, a distinction between lexical words (or information-carrying words) and function words (words that bind together a text) within the word classes of English must be obvious. The LD is calculated based on this formula:

***LD = (Number of lexical words/ total number of words) x 10***

In terms of findings, the analysis process was divided into three main phases as mentioned below with specific statistics for each one.

- ***Phase 1: Before the course***

The main purpose of the first phase is for group classification and facilitates the researcher to get a general overview of students' level. Students were required to take part in a writing mini-test, specifically writing a 250-word essay about a topic with a view to accessing their entrance level. The type of essay is "Problem-Solution Essay" (as it is included in students' curriculum) and the chosen topic is "*These days, there is a decrease in the number of people choosing teaching as their profession. What are the problems and what are the possible solutions?*". However, for some personal reasons, 2 students could not submit the given task before deadline, so there were only 28 students participating in this research. After collecting 28 essays, these indexes were concluded based on these indexes mentioned in Methodology part. 28 collected essays are put into LanscBox 4.0 and Lextutor.ca application to run for the index of TTR, LD and AWL which are illustrated in Table 2.

***Table 2.*** *Statistics for TTR, LD and AWL indexes in phase 1*

| INDEX | STUDENT'S WRITING | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** | **S9** |
| TTR | 12.1% | 15.6% | 24.4% | 35.1% | 16.1% | 34.4% | 25.1% | 29.8% | 13.4% |
| LD | 0.53 | 0.42 | 0.4 | 0.67 | 0.21 | 0.53 | 0.64 | 0.21 | 0.43 |
| AWL | 4.1% | 5.03% | 6.14% | 5.34% | 2.46% | 5.34% | 4.2% | 5.1% | 6.04% |

| INDEX | STUDENT'S WRITING | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **S10** | **S11** | **S12** | **S13** | **S14** | **S15** | **S16** | **S17** | **S18** |
| TTR | 15.1% | 38.7% | 41.9% | 32.2% | 23.3% | 18.4% | 27.1% | 31.3% | 29% |
| LD | 0.32 | 0.58 | 0.64 | 0.31 | 0.38 | 0.29 | 0.3 | 0.56 | 0.45 |
| AWL | 5.43% | 4.1% | 2.54% | 5.86% | 5.74% | 5.5% | 6.73% | 6.82% | 5.52% |

| Index | STUDENT'S WRITING | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **S19** | **S20** | **S21** | **S22** | **S23** | **S24** | **S25** | **S26** | **S27** | **S28** |
| TTR | 19.2% | 23.1% | 22.4% | 23.2% | 34.5% | 46.1% | 23.4% | 31.6 % | 25.4% | 21% |
| LD | 0.46 | 0.53 | 0.67 | 0.19 | 0.56 | 0.54 | 0.64 | 0.51 | 0.31 | 0.4 |
| AWL | 2.01% | 7.54% | 5.52% | 5.81% | 7.63% | 6.82% | 5.46% | 4.3% | 3.33% | 4.1% |

According to these statistics, those participants with higher TTR, LD and AWL index are listed in the Experimental group (EG) and they will learn collocations through the corpus-based method during the six-week course. The reason why the researcher decided to choose those who had higher statistics in their writings for the EG was justifiable. These indexes TTR, LD and AWL reveal students' diverse vocabulary items, the total percentage of academic words ranged from above B1 level, and the distinction between the lexical and function words respectively. Therefore, the higher these indexes are, the more proficient students are in terms of using vocabulary. That can be considered as one of the initial evidences for group classification in this research. As a result, two groups are classified with the equal number of participants (14 for each), namely the Experimental group and the Control group.

The main function of this first phase is for group classification which creates a favorable condition for the researcher to carry out different teaching methods for each group in a six-month course for a deeper analysis of the next phases. By analyzing carefully, the very first major finding the researcher can recognize is that students have a tendency of ***misusing collocations***, which means they cannot recognize collocations or in other words, they cannot determine which words should go together to form a chunk. Adapted from students' essays, some collocations were highlighted such as "*hard schedule, busy schedule, outside activities, male career* and *significant workloads*". Based on the MI-score which measures the amount of non-randomness present when two words occur, each collocation was calculated to find out the MI-score so that the researcher could decide whether each one is a significant collocation or not. After calculating, the MI-score of all mentioned collocations is all under 3, which means they cannot be considered as significant collocations. As a result, it can be evidently inferred that the collocations used in students' writing are misused or in other words, they are used in an inappropriate way. This finding plays a pivotal role in orientating the researcher to propose some suitable teaching methods to tackle the problem of misusing collocations among participants. Back to those aforementioned collocations, based on the COCA (Corpus of Contemporary America) corpus, they can be corrected or replaced by more significant collocations such as "*hectic schedule, extracurricular activities, male profession* and *heavy workloads*"; and the MI-score for those are above 3, which means they are truly significant collocations and best used in academic texts.

- ***Phase 2: After the course***

After classifying groups, two different teaching approaches were applied for each group in 6 weeks. The corpus-based method was for the EG and the traditional one for the CG. After six-week course, an essay was given for both groups to check for the efficiency of each

method. The type of the essay was still "Problem-Solution essay", but the topic for both groups in this phase was "*Obesity is becoming more and more alarming, particularly among the youth. What are the main causes? What are solutions to address this issue?*". In this phase, the researcher analyzed statistics of each group separately, and then compare the two results to access the efficiency between two groups. Additionally, in each group, each student's essay was applied to run for separate statistics and then the average statistic for each group was calculated in the total of essays in each group. Figure 5 below is an illustration of running a student's essay for detail statistics, namely TTR, AWL and LD thanks to the website Lextutor.ca. Besides, all the analysis figures of Control group, Experimental group and both goups are detailed in table 3, 4 and 5 respectively.

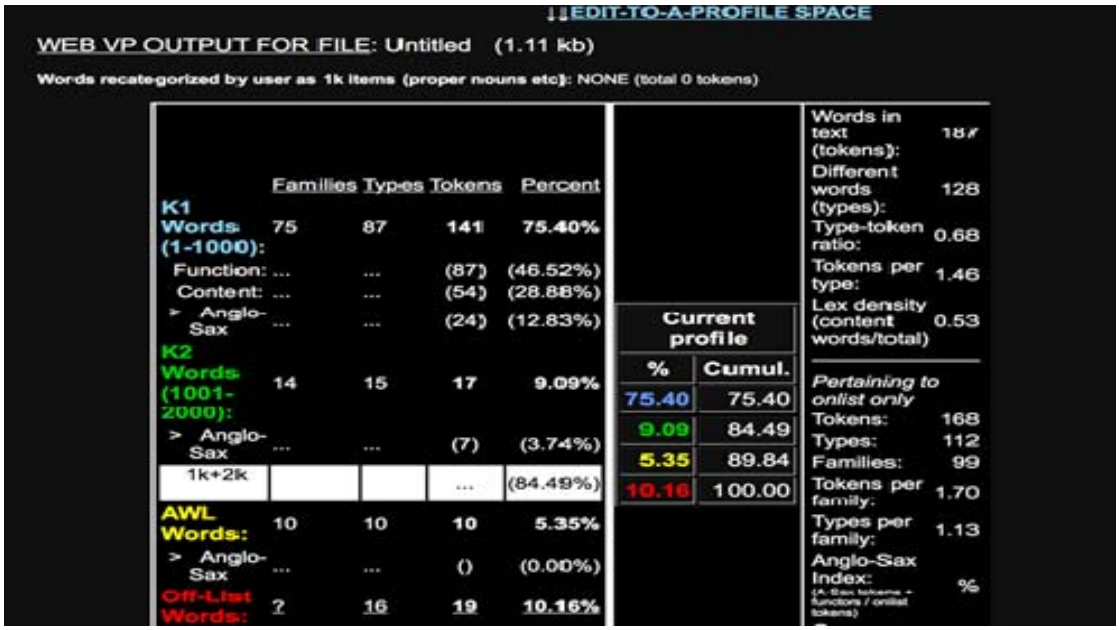


***Figure 5.*** *An illustration of running a student's essay for detail statistics in the website Lextutor.ca*

***Table 3.*** *Analysis figures of CG in Phase 2*

| Control Group – Phase 2 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Index** | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** | ***S7*** | ***S8*** | ***S9*** | ***S10*** | ***S11*** | ***S12*** | ***S13*** | ***S14*** |
| **TTR** | 0.76 | 0.81 | 0.7 | 0.72 | 0.67 | 0.73 | 0.59 | 0.65 | 0.86 | 0.73 | 0.75 | 0.72 | 0.74 | 0.66 |
| **AWL** | 4.58 % | 3.17 % | 6.02 % | 2.54 % | 4.9% | 1.8% | 1.13 % | 4.7% | 3.9% | 4.21 % | 3.39% | 2.65 % | 4.04 % | 3.89 % |
| **LD** | 0.56 | 0.56 | 0.56 | 0.53 | 0.6 | 0.57 | 0.51 | 0.6 | 0.6 | 0.49 | 0.57 | 0.52 | 0.64 | 0.52 |

***Table 4.*** *Analysis figures of EG in Phase 2*

| Experimental Group – Phase 2 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Index** | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** | ***S7*** | ***S8*** | ***S9*** | ***S10*** | ***S11*** | ***S12*** | ***S13*** | ***S14*** |
| **TTR** | 0.73 | 0.87 | 0.75 | 0.81 | 0.86 | 0.9 | 0.71 | 0.65 | 0.73 | 0.82 | 0.72 | 0.86 | 0.8 | 0.71 |
| **AWL** | 5.35 % | 3.98 % | 4.93 % | 6.1% | 5.93 % | 7.2% | 6.83 % | 5.7% | 5.84 % | 8.2% | 8.01 % | 4.9% | 5.03% | 6.1% |
| **LD** | 0.82 | 0.73 | 0.84 | 0.76 | 0.9 | 0.83 | 0.71 | 0.7 | 0.8 | 0.69 | 0.92 | 0.73 | 0.74 | 0.69 |

***Table 5.*** *Average statistics of both groups in Phase 2*

| **Index** | **Control Group** *(including 14 files, 1685 tokens, 533 types)* | **Experimental Group** *(including 14 files, 1883 tokens, 622 types)* |
|---|---|---|
| **TTR** | 31.6% | 33.03% |
| **AWL** | 3.63% | 5.98% |
| **LD** | 0.56 | 0.81 |

As can be obviously seen in table 3, 4 and 5 above, the indexes for different group are calculated and summarized. The deviation between two groups in terms of TTR, AWL and LD index creates a condition for the researcher to analyze and draw a general comparison about the effectiveness of each method on students.

- **TTR**: TTR index of CG and EG are 31.6% and 33.03% respectively. The figures are nearly the same so it can be easily illustrated that students in both groups are aware of using a variety of words and avoid using repetitive ones to make sure the Lexical Resources criterion (LR) in Writing assessment list. However, the figures are quite low (compared to 100%) so it can be acknowledged that students' ability to use a wide range of academic vocabulary items is still limited.
- **AWL**: The AWL index of CG and EG are 3.63% and 5.98% in turn, which means the index of the latter is nearly as twice as the former's. Hence, it is feasible to infer that the total number of academic vocabulary items above B1 level of EG is much higher than that of CG group.
- **LD**: The LD indexes for CG and EG are 0.56 and 0.81 respectively. The deviation is not such a huge gap, however, these figures can still reveal the fact that the quality of participants' essays in EG is better than that of CG in terms of academic words and phrases.

Apart from these initial conclusions based on the disparity between two groups in terms of TTR, LD and AWL scores, the second major finding in this thesis was revealed. By looking closer at student's writing in each group and comparing them to the first phase, it is evidently obvious that there is ***a dramatically decrease in the ratio of misusing collocation*** among participants in both groups. This finding can be considered as a tremendous evidence for great improvement in students' collocational competence as well

as for the effectiveness of both teaching methods applied for each group. Take some pieces of students' writings as examples to illustrate this point.

After a six-week course, students in both groups have raised their awareness of using correct collocations. They also paid more attention to applying collocations in their essays in order to boost their lexical resource score. There are a host of significant collocations related to the topic "Obesity" that can be easily found in those above examples, namely "*sedentary lifestyle, hectic schedule, high-processed food, ultra-processed food, fatal diseases, epidemic sicknesses, work-life balance, food safety*" and so on. They are those collocations taught by the researcher during the six-week course and they were applied flexibly and correctly in students' essays, which illustrates the improvement in student's collocational competence in writing skill. However, when it comes to comparison between two groups in this phase with a view to comparing the effectiveness of these two teaching methods (corpus-assisted method and traditional one), there is no doubt that the collocations used by students in EG ***are more significant and at higher level according to the CEFR*** than those used by students in CG. This finding is of fundamental importance, and it is concluded based on the t-score of each collocation (t-score measures the certainty of a collocation). The reason why MI-score is effective in this comparison is that Mi-score can be measured across different corpora. So, it is justifiable to compare across two corpora in terms of the strength of collocations.

The comparison between each pair of collocations (with the same meaning or illustration) based on MI-score is described in the table 6 below.

***Table 6.*** *MI score of each pair of collocation in students' writing*

| EG student's writing | CG student's writing |
|---|---|
| Sedentary lifestyle (MI = 4.85) | Unhealthy lifestyle (MI = 2,78) |
| Ultra-processed food (MI = 5.6) | Fast food (MI = 3.43) |
| Whole food (MI = 4.21) | Unhealthy food (MI = 2.64) |
| Adverse impact (MI = 4.56) | Negative effect (MI = 3.02) |

It can be inferred that the collocations used by students in EG are stronger and have a closer relation than those in CG thanks to the calculation of MI-score.

On the flip side, besides the above finding, by analyzing carefully the students' essays in both groups, the researcher found out that the students in EG have a tendency of ***using more compound premodifiers as adjective*** than those in CG. Compound premodifiers are words that are connected together with a hyphen and illustrate a general meaning, namely "*fast-paced, budget-friendly, health-conscious, work-life, far-reaching, ultra-processed*" and so on. These compound premodifiers tend to act as adjectives supporting the nouns in terms of meaning.

As can be seen clearly from essays of students in EG group, with the assistance of corpus-based method, they have applied flexibly and appropriately the use of compound

premodifiers as adjectives in their writing. Some typical collocations in the examples are "*health-conscious person, fast-paced life, far-reaching repercussion, work-life balance, budget-friendly fast food joints, on-a-daily-basis meals*" and so on. The reason why the researcher paid attention to compound premodifiers is that they are such high-level vocabulary items (above B1 level) and they are occasionally used in academic articles, which are of great importance in the lexical resource marking criteria. Compared to CG, by careful observation, students still cannot notice the use of compound premodifiers. It is quite easy to understand as the corpus facilitates the users the concordance lines which show clearly all the compound words, while the traditional method does not.

- ***Phase 3: Two weeks after the course***

The process in this phase was carried out the same as phase 2, even the topic for writing. However, in this phase, participants were required to be at the same place and at the same time, and what they had to do was to write an essay (at least 250 words) in a limited time (40 minutes). Similarly, like phase 2, they were not let to know the main purpose of evaluating their essay (collocational competence), and they just knew this writing as their mid-term test. After 40 minutes, all writings of the two groups were collected separately and used for statistics analysis. The evaluation of collocational competence between EG and CG was still based on three main indexes: TTR, AWL and LD. As the same as phase 2, the analysis figures of Control group, Experimental group and both groups are illustrated in table 7, 8 and 9 in turn below.

***Table 7.*** *Analysis figures of CG in Phase 3*

| **Control Group – Phase 3** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Index** | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** | ***S7*** | ***S8*** | ***S9*** | ***S10*** | ***S11*** | ***S12*** | ***S13*** | ***S14*** |
| **TTR** | 0.68 | 0.53 | 0.64 | 0.81 | 0.61 | 0.5 | 0.72 | 0.85 | 0.68 | 0.71 | 0.73 | 0.67 | 0.71 | 0.64 |
| **AWL** | 1.86 % | 3.14 % | 5.03 % | 2.45 % | 4.98 % | 2.31 % | 1.89 % | 4.2 % | 4.5 % | 3.46 % | 3.24 % | 2.76 % | 4.35 % | 2.01 % |
| **LD** | 0.53 | 0.51 | 0.56 | 0.65 | 0.52 | 0.57 | 0.68 | 0.7 | 0.6 | 0.42 | 0.48 | 0.51 | 0.63 | 0.49 |

***Table 8.*** *Analysis figures of EG in Phase 3*

| **Experimental Group – Phase 3** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Index** | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** | ***S7*** | ***S8*** | ***S9*** | ***S10*** | ***S11*** | ***S12*** | ***S13*** | ***S14*** |
| **TTR** | 0.64 | 0.82 | 0.63 | 0.71 | 0.82 | 0.88 | 0.64 | 0.7 | 0.69 | 0.81 | 0.67 | 0.91 | 0.85 | 0.74 |
| **AWL** | 5.93 % | 4.05 % | 5.81 % | 6.1% | 6.12% | 7.8% | 6.81 % | 7.5% | 4.4 % | 8.25 % | 7.01 % | 4.94 % | 5.33 % | 7.05 % |
| **LD** | 0.86 | 0.82 | 0.73 | 0.76 | 0.85 | 0.79 | 0.73 | 0.8 | 0.7 | 0.69 | 0.98 | 0.83 | 0.91 | 0.65 |

***Table 9.*** *Average statistics of both groups in Phase 3*

| **Index** | **Control Group** *(including 14 files, 1842 tokens, 612 types)* | **Experimental Group** *(including 14 files, 1792 tokens, 634 types)* |
|---|---|---|
| **TTR** | 33.2 % | 36.8% |
| **AWL** | 3.62% | 6.87% |
| **LD** | 0.63 | 0.88 |

According to the statistics collected in phase 3, the researcher continued to take a closer look at the deviation between the indexes of two groups (EG and CG) in terms of different criteria, namely TTR, AWL and LD.

- **TTR**: The TTR index of CG and EG are 33.2% and 36.8%, so it can be seen that they make no difference with these of phase 2 and they are nearly the same to each other. Compared to TTR index of both groups in the second phase, which are 31.6% and 33.03% in turn, this can be inferred that the the ratio of vocabulary diversity in each corpus remains unchanged.
- **AWL**: The AWL index of EG is still as nearly twice as that of CG like phase 2, which means the total number of academic vocabulary items above B1 level of EG is much higher than that of CG group; and the participant's competence of using collocations in their writings as well as the more effectiveness of the corpus-assisted method compared to the traditional one after two weeks of the courses.
- **LD**: Like phase 2, the deviation of LD index for both CG and EG is not much (0.63 and 0.88 respectively for each group). These figures can still reveal the fact that the quality of participants' essays in EG is better than that of CG in terms of academic words and phrases.

## 4. Conclusion

As mentioned in collocation part, *collocation* is a language phenomenon that arouses many insurmountable obstacles for students in the process of learning language. According to Lewis & Gough (1997), the lexical approach always aspires to learning and teaching vocabulary in chunks, especially collocations, so it is sensible to conclude that collocation is of significant importance and it exerts a tremendous influence on learner's language competence, especially in terms of both receptive and productive use of the language to L2 learners.

In this thesis, based on the statistical analysis, the researcher discovers that the participants hold positive attitudes toward the teaching and learning collocations in the classroom, even if the teaching method is traditional or corpus-based. All participants are gradually becoming more confident with their lexical resources thanks to acknowledgement of collocations and chunks, however, the EG report that they are absolutely impressed by the use of concordance lines in corpus-based method. They describe a corpus as "a living dictionary" which is very useful for them because a corpus gives them a host of authentic examples with real information, so that they can understand deeper about the collocations as well as become more well-informed for their essays. It is such good news as the development of technology has made it possible for students to explore corpora of authentic language and obtain samples of texts from these corpora with a concordance.

In addition, the original purpose of this thesis is to discover the effectiveness of corpus-based method in the development of students' collocational competence, or in other words, is how corpus has assisted learners in learning collocations to improve their lexical resources. By comparing and contrasting two groups called CG and EG, based on the statistical analysis, it can obviously be inferred that the EG is able to learn and remember collocations better than the CG through the authentic examples (or authentic concordance lines). The collocations that were used in EG participants' essays are much more academic and nativelike, compared to the CG ones. Especially, after two weeks finishing the courses, the EG can still remain the better result in terms of collocation uses which are stated clearly in the statistics. Hence, the researcher can draw a conclusion that despite positive effect of teaching collocations in both groups, it is irrefutable that the corpus-based method is more effective for students in the process of developing collocational competence than the traditional one. Additionally, the corpus-based method also creates more favorable conditions for learners to broaden their horizons and memories.

## 5. Limitation

This research attempts to understand the effectiveness of the corpus-based method on teaching and learning collocations in a university context. Despite the research findings above, the quasi-experimental design is not without limitation. In this study, the researcher uses a relatively small sample (10 articles and 28 selected participants), so the generalizability of these participants' perceptions to other populations with different educational backgrounds or teaching methods may be limited.

## 6. Pedagogical suggestion

It is undeniable that all the participants express their positive attitudes toward learning collocations and they confirm that collocations are of fundamental importance to their writing skill. However, some confess that it is, sometimes, difficult for them to identify the right key words. Moreover, when they find the collocations in some general corpus which contains millions of types and tokens such as BNC (British National Corpus) or COCA (Corpus of Contemporary American), they feel quite confused as they are huge amount of related information in concordance lines in that corpus. This is attributed to the lack of students' collocation concept as well as their low level of language proficiency. That is the reason why I highly recommend the use of Sketch Engine application, which helps users create a list of multi-words and then the learners will use Corpus application to search for key words and observe them carefully in concordance lines. One thing can be suggested is output task. Teachers should provide learners a corpus containing of many articles/ texts related to the same topic, allow them to use corpus application to identify collocations/ chunks and then require students to apply those collocations/ chunks into their writing or speaking skills (productive skills). By doing this, students are enabled to remember collocations well and use them flexibly, even in daily communication.

❖ ***Conflict of Interest:*** *Author have no conflict of interest to declare.*

**REFERENCES**

Creswell, J. W. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (4th ed). Boston: Pearson.

Hunston, S. (2010). *Corpora in applied linguistics* (7. print). In *The Cambridge Applied Linguistics Series* (7. print). Cambridge: Cambridge University Press.

Lewis, Michael, & Conzett, J. (Eds.). (2002). *Teaching collocation: further developments in the lexical approach*. Boston: Thomson, Heinle.

Lewis, Michael, & Gough, C. (1997). *Implementing the lexical approach: putting theory into practice* (Nachdr.). Andover: Heinle Cengage Learning.

Lewis, Micheal. (2006). *Teaching collocation: Further developments in the lexical approach*. Cambridge University Press.

Lüdeling, A., & Kytö, M. (Eds.). (2009). *Corpus linguistics: an international handbook*. Berlin; New York: Walter de Gruyter.

Li, S. (2017). Using corpora to develop learners' collocational competence. *Language Learning and Technology*, *21*(3).

McEnery, T., & Wilson, A. (2011). *Corpus linguistics: an introduction* (2. ed., repr). In *Edinburgh Textbooks in Empirical Linguistics* (2nd ed., repr). Edinburgh: Edinburgh University Press.

Nation, P. (2000a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2000b). *Teaching vocabulary*. Oxford: Oxford University Press.

Nation, P. (2007). The Four Strands. *Innovation in Language Learning and Teaching*, *1*(1), 2-13. https://doi.org/10.2167/illt039.0

Nesselhauf, N. (2004). *Collocations in a learner corpus*. In *Studies in Corpus Linguistics*, *14*. Amsterdam; Philadelphia: J. Benjamins Pub. Co.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge; New York: Cambridge University Press.

Reppen, R. (2010). *Using corpora in the language classroom*. In *Cambridge Language Education*. New York: Cambridge University Press.

# ỨNG DỤNG NGÔN NGỮ HỌC KHỐI LIỆU TRONG VIỆC GIẢNG DẠY CÁC CỤM TỪ CÓ TÍNH CHẤT BỀN VỮNG TRONG TIẾNG ANH Ở MÔI TRƯỜNG ĐẠI HỌC

***Nguyễn Thị Thanh Huyền***

*Trường Đại học Sư phạm Hà Nội*

*Tác giả liên hệ: Nguyễn Thị Thanh Huyền – Email*: *nguyenthanhhuyen2111@gmail.com*

*Ngày nhận bài: 18-4-2019; ngày nhận bài sửa: 13-7-2019; ngày duyệt đăng: 18-7-2019*

## TÓM TẮT

*Sinh viên học ngôn ngữ tiếng Anh thường có xu hướng học từ vựng đơn lẻ, không theo cụm từ. Chính điều này đã ảnh hưởng không nhỏ đến khả năng tư duy ngôn ngữ theo cụm từ nói riêng và vốn từ vựng của sinh viên nói chung. Trong nghiên cứu này, phương pháp dạy từ vựng dựa vào ngôn ngữ học khối liệu đã được tiến hành áp dụng bởi phương pháp này có ý nghĩa rất lớn khi đem lại cho người học cơ hội tiếp xúc với ngôn ngữ đời thực hoặc những tài liệu dạy và học từ vựng có tính chất thiết thực. Chính vì thế, bài báo này tập trung nghiên cứu tiềm năng của việc sử dụng khối liệu và các dòng ngữ liệu trong việc dạy và học các cụm từ có tính chất bền vững, nhằm mục đích cải thiện vốn từ vựng của sinh viên áp dụng trong kĩ năng viết học thuật. Để đạt được mục đích nghiên cứu, một thử nghiệm đã được tiến hành với 30 sinh viên năm 3 thuộc Khoa Tiếng Anh – Trường Đại học Quốc gia Hà Nội (tên giả), và hầu hết những sinh viên được chọn đều không biết hoặc biết rất ít về ngôn ngữ học khối liệu. 30 sinh viên được lựa chọn và được chia thành hai nhóm bằng nhau: nhóm thử nghiệm và nhóm kiểm soát. Trong một khóa học sáu tuần liên tiếp, trong khi nhóm thử nghiệm được áp dụng phương pháp dạy từ vựng dựa vào khối liệu, thì nhóm kiểm soát lại được áp dụng phương pháp dạy truyền thống dựa vào các quy tắc. Mục đích của việc chia nhóm là tìm ra điểm khác biệt và so sánh xem liệu rằng phương pháp dạy dựa vào khối liệu có thực sự đem tới hiệu quả cho người học hay không. Tất cả sinh viên đều bắt buộc phải tham gia những bài kiểm tra vào các thời điểm khác nhau: trước khóa học, ngay sau khi kết thúc khóa học và 2 tuần sau khi khóa học đã kết thúc. Những bài kiểm tra được được phân tích cụ thể, kĩ càng dựa vào các tiêu chí đánh giá nhằm mục đích kiểm tra khả năng dùng từ vựng theo cụm của sinh viên. Kết quả cho thấy cả 2 nhóm đều đạt được sự tiến bộ trong việc sử dụng các cụm từ có tính chất bền vững trong văn viết học thuật; tuy nhiên, nhóm thử nghiệm với phương pháp giảng dạy dựa vào khối liệu cho thấy hiệu quả vượt trội so với việc giảng dạy theo phương pháp truyền thống. Dựa trên so sánh đối chiếu kết quả của việc áp dụng 2 phương pháp khác nhau trong việc dạy các cụm từ có tính chất bền vững, tác giả đưa ra kết luận việc ứng dụng ngôn ngữ học khối liệu trong thiết kế các dạng bài tập khác nhau (từ cơ bản đến nâng cao) trong quá trình tăng tính hiệu quả và sáng tạo đối với việc dạy các cụm từ có tính chất bền vững.*

***Từ khóa:*** Ngôn ngữ học khối liệu, phương pháp giảng dạy dựa vào khối liệu, dòng khối liệu, các cụm từ có tính chất bền vững, năng lực sử dụng từ vựng theo cụm.