NALANS
JOURNAL OF NARRATIVE AND LANGUAGE STUDIES | ISSN: 2148-4066

Journal of Narrative and Language Studies – December 2021, Volume 9 – Issue 18

A Learner Corpus Investigation of Formulaic Sequence Development in EFL Learners with a Focus on Native and Non-Native Corpora

Ali Şükrü Özbay
Karadeniz Technical University
alisukruozbay@gmail.com

Hanife Öztürk
Ağrı İbrahim Çeçen University
hanifeozturk00@hotmail.com

APA Citation:
Özbay, A.Ş. & Öztürk, H. (2021). A learner corpus investigation of formulaic sequence development in EFL learners with a focus on native and non-native corpora. *Journal of Narrative and Language Studies*, *9*(18), 412-437.

Abstract

In recent years, the advent of computer technology and software tools have made it available for more complicated and fully operational facilities for corpus linguistics. Thanks to these developments, the compilation of large collections of naturally occurring texts was made more accurately. In line with these developments, the current study aimed to investigate the usage patterns of three- to four-word sequences in a learner corpus composed of two semesters of written data from 85 English as a Foreign Language (EFL) learners. The data was analyzed by examining collective trends in terms of usage patterns of formulaic sequences across different time intervals. In the collection of data, the frequency approach was used. The most frequent three- and four-word recurrent formulas were extracted from each sub-corpus of the learner corpora in two groups. These sequences were classified structurally and functionally. Then, the use of these sequences was compared across native (LOCNESS) and non-native data by using the Sketch Engine corpus tool. The findings suggested that although formulaic sequences were used frequently in both learner groups, the frequency and type of these formulaic sequences were less diverse, and the number of formulaic sequences was limited when compared with the native data.

Keywords: Formulaic sequences (FSs); Learner corpus; EFL learners; multi-word units

Introduction

"The use of corpus for lexical investigation is not a recent phenomenon, but its full significance and value has, in the last decade, been realized especially after the introduction of computerized corpus tools by a much larger group of linguists all around the globe" (Özbay and Kayaoğlu, 2016, p.343). In foreign language learning and teaching, the use of multi-word combinations is crucial for language development. With computerized textual corpus tools and computer-assisted techniques, the focus of language studies has shifted to research on the recurrent multi-word sequences, and several large-scale studies were conducted to investigate native and non-native formulaic patterns. The significance of studying multi-word combinations, especially formulaic sequences, is that they provide a clearer understanding of how they are used in learners' texts. As noted by Chenu and Jisa, formulaic sequences “provide a stepping-stone into language development" (2009, p.27), and they are regarded as a cornerstone of L2 (Ellis, 1996; Oakey, 2002; Jones and Haywood, 2004; Schmitt and Underwood, 2004; Ellis et al., 2008; Wray and Fitzpatrick, 2008; Wray and Fitzpatrick, 2010). Thus, it is seen that tracking the usage patterns of FSs helps to notice the importance of these sequences in learner writing. Moreover, the importance of the current study as corpus-based research lies in its more profound exploration of collective trends. For instance, Simpson-Vlach and Ellis (2010) studied multi-word combinations and derived the list of FSs for academic writing and speech by conducting a corpus-based investigation, similar to that of Coxhead (2000), who generated an academic word list by compiling a corpus of written academic texts. Corpus-based research has provided evidence for revealing many different repeated patterns in language use and has ensured that native language is wealthy with respect to formulaic sequences (Schmitt and Carter, 2004). According to Schmitt and Carter (2004), corpus data has enlightened “the field by identifying formulaic language and describing how it is used in discourse” (2004, p. 11). Granger (2002) argued that "the area of linguistic enquiry known as learner corpus research has created an important link between the two previously disparate fields of corpus linguistics and foreign/second language research" (2002, p. 4). Learner corpus studies also assisted the studies investigating the acquisition, use and development of multi-word sequences and comparing the proper use of these sequences in different settings. The efficiency of the language formulas in both written and spoken language is perceived as a trace of competence in linguistic performance. As noted by Hyland (2008, p. 4), "multi-word expressions or formulaic sequences are important components of fluent linguistic production and a key factor in successful language learning”. Therefore, noticing the development of these sequences on non-native learners' texts provide an insight into how they progressed in writing. According to Ellis et al. (2008), language instructors should be aware of the extensive usage of sequences and their prominence in language, and they should inform learners about "which formulas should be prioritized for instruction with learners at different stages of development" (p. 379). Henceforth, it is essential to specify the development of language formulas in non-native students' texts. Although there have been ongoing issues related to usage patterns of formulaic sequences by L2 learners in their written outputs, the scope of these studies seems to be limited to specific formulaic patterns (Siyanova-Chanturia and Spina, 2020). In the context of these issues, what motivated this study is the need to understand the development of FSs in EFL learner corpora across different time intervals. We aimed to investigate the usage of 3- and 4-word sequences in 85 EFL learners' written outputs and focused on the examination of the structures and functions of these formulaic sequences in argumentative essays. These written productions were compiled during 2018-2019 (Fall and Spring), and two sub-groups of corpora were created. Obtained initial lists of these sequences were filtered, and the structural and functional characteristics of FSs have been listed by using the structural and functional framework of Biber et al. (1999, 2004).

Significance of Formulaic Sequences and Frequency of FSs

A thorough analysis of the existing literature has shown that the formulaic sequences were widespread in native speakers' language considering both frequency of occurrence and diversity (McCarthy, 1991; Schmitt and Carter, 2004) and they were the numerically predominant (roughly ten to one) in language use (Mel'cuk, 1998), while over half of the spoken and written discourse is comprised of these patterns (Erman and Warren, 2000). When compared to single morpheme lexical units, the dominance of FSs is noticed (Wong, 2012). It is important to note that these units are regarded as the primary carrier of meaning (Sinclair, 2008) and to become proficient writers in L2, it is required to have a high degree of proficiency in the use of FSs (Liou and Chen, 2018). It is also increasingly evident that FSs are valuable, and "their acquisition must become an essential feature of any model of language acquisition" (Schmitt and Carter, 2004, p.14). There were several reasons that indicated the significance of FS, such as its ubiquitous nature, its processing advantages and its role to realize the meanings and functions, its contribution to improving overall L2 production (Martinez and Schmitt, 2012).

The frequency of occurrence for both individual words (Nation, 2001; O'Keeffe et al., 2007) and formulaic sequences (Martinez, 2011) were used as one of the indicators of usefulness in language. Being a frequently-cited criterion, these sequences are conventionalized in language by native speakers (Schmitt and Carter, 2004), considered as "a salient, perhaps even a determining, factor in the identification of formulaic sequences" (Wray, 2002, p.25). According to Ellis (2013), frequency of usage is decisive for learning, memory and perception and the relationship between frequency and formulaicity is obvious when "that formulaic output is frequently called upon" (Wray and Perkins, 2000, p.7) as well as "retrieving and recognizing such multi-word units would facilitate the level of fluency" (Conrad and Biber, 2005, p.57).

Structural and Functional Characteristics

The sequences extracted from the corpus are categorized with regards to structures, functions and registers (Greaves and Warren, 2010). In this study, the taxonomy of Biber et al. (1999; 2004) was used for the classification of word sequences. The structural categories with three main types are verb phrase fragments, dependent clause fragments and noun phrase and prepositional phrase fragments, exemplified in Table 1 below.

Table 1

Structural Classification of Language Formulas

1. Verb phrase fragments	1a. (connector +) 1st/2nd person pronoun + VP fragment	*you don't have to, I'm not going to, well*
	1b. (connector +) 3rd person pronoun + VP fragment	*it's going to be, that's one of the, and this is a*
	1c. Discourse marker + VP fragment	*I mean you know, you know it was, I mean*
	1d. Verb phrase (with non-passive verb)	*is going to be, is one of the, have a lot of,*

	1e. Verb phrase (with passive verb)	*is based on the, can be used to, shown in fi*
	1f. Yes-no question fragments	*are you going to, do you want to,*
	1g. WH-question fragments	*what do you think, how many of you,*
2. Dependent clause fragment	2a. 1st/2nd person pronoun + dependent clause fragment	*I want you to, I don't know if, I don't know*
	2b. WH-clause fragments	*what I want to, what's going to happen,*
	2c. If-clause fragments	*if you want to, if you have a, if we look at*
	2d. (verb/adjective+) to-clause fragments	*to be able to, to come up with, want to do is*
	2e. That-clause fragments	*that there is a, that I want to, that this is a*
3.Noun phrase and prepositional phrase fragments	3a. (connector+) Noun phrase with of-phrase fragment	*one of the things, the end of the, a little bit of*
	3b. Noun phrase with other post-modifier fragment	*a little bit about, those of you who, the way*
	3c. Other noun phrase expressions	*a little bit more, or something like that*
	3d. Prepositional phrase expressions	*of the things that, at the end of,*
	3e. Comparative expressions	*as far as the, greater than or equal,*

Source: Biber et al., 2004, p.381

Functional categories are given with three main types, which are stance expressions, discourse organizers, and referential expressions. Stance expressions are identified as "epistemic evaluations or attitudinal/modality meanings" (Biber and Barbieri, 2007, p.270). Discourse organizers are used for indicating "the overall discourse structure and signal the informational status of statements" (Biber and Barbieri, 2007, p.271). Referential expressions can be described as "an entity or single out some particular attribute of an entity as essential" (Biber and Barbieri, 2007, p.271). The functional taxonomy created by Biber (1999, 2004) can be seen in Table 2.

Table 2

Functional Classification of Language Formulas

1. Stance expressions	A. Epistemic stance	*I don't know what/if/how/I, I think it was, the fact that*
	B. Attitudinal/modality stance	
	b1. Desire	*I don't want to, if you want to, do you want a,*
	b2. Obligation/ directive	*it is important/necessary to, you have to do,*

	b3. Intention/ Prediction	*I'm not going to, are we going to, going to be a,*
	b4. Ability	*to be able to, to come up with, can be used to, it is possible to*
2.Discourse organizers	A. Topic introduction/focus	*if you look at, want to talk about, let's have a look*
	B. Topic elaboration/clarification	*I mean you know, has to do with, on the other hand,*
3.Referential expressions	A. Identification/ focus	*of the things that, that's one of the, is one of the,*
	B. Imprecision	*or something like that, and stuff like that, and things like that*
	C. Specification of attributes	
	c1. Quantity specification	*have a lot of, how many of you, a little bit of, percent of the*
	c2. Tangible framing attributes	*the size of the, in the form of*
	c2. Intangible framing attributes	*in the case of, in terms of the, as a result of, on the basis of*
	D. Time/ Place/ Text reference	
	d1. Place reference	*in the United States, of the United States, the United States*
	d2. Time reference	*at the same time, at the time of*
	d3. Text-deixis	*shown in figure N, as shown in figure*
	d4. Multi-functional reference	*the end of the, the beginning of the, the top of the,*

Source: Biber et al., 2004: 384-388

Method

The current study addressed the following questions:

1.What are the most frequent 3- and 4-word sequences found in the learner corpus for two semesters of language development?

2. What are the structures and functions of the frequent formulaic sequences?

3. How similar or different are the frequent 3- to 4-word formulaic patterns produced by the learners from those found in native English corpora (LOCNESS)?

Participants

The participants of the study were freshmen studying in the English Department of a mid-size University in the northeast of Turkey. The students speak Turkish as their native language. The total number of participants consisted of 85 ELL students 59,99 % of them were females (n=51) and 39,99 % of them were males, (n=34) and whose ages range between 18-22. These participants studied one-year intensive English preparatory classes before their bachelor's degree, and the medium of instruction throughout their undergraduate education was English.

Instruments, Data Collection

The instruments employed in the study include one written native reference corpus; LOCNESS (The Louvain Corpus of Native English Essays), which involves 322 texts with a total of 360,685 tokens produced by native speakers of English who were between 17 and 23 years of age, and one written non-native corpus, the Learner Corpora. Native speaker corpus LOCNESS is made up of argumentative essays of British pupils' A level essays, British university students' essays and American university students' essays. It is important to note that LOCNESS as a control native speaker corpus is representative enough with regards to text types, sizes, participants ages and topic; therefore, it is considered to be comparable to the learner corpora used in this study. The data for learner corpora were gathered during two semesters of observation of academic writing courses for fall and spring terms 2018-2019 and were compiled every week and yielding 824 texts for two semesters of observation of the argumentative essays from 85 participants. Participants received in-class training 4 hours of classroom instruction per week, four weeks a month. After the training session, participants wrote an untimed essay. In the following week after each training, a course time was spent for teacher feedback sessions. At the beginning of the year, the participants were asked to write untimed essays on the assigned topic and their proficiency levels in English were rated by three instructors from two different universities with reference to rating criteria as an analytic scoring. In accordance with proficiency levels of learners, the participants were divided into two groups, and the essays were compiled into five sub-corpora whose profiles are given in Table 3.

Table 3

The Profiles of Learner Corpora in Two Groups

	Sub-corpus	***N* Participants**	***N* Texts**	***N* Tokens**	***N* Words**
	Sub-corpus 1	42	84	47,909	**42,582**
	Sub-corpus 2	42	81	51,573	**45,799**
Group 1	Sub-corpus 3	42	83	63,780	**56,930**
	Sub-corpus 4	42	80	72,065	**64,632**
	Sub-corpus 5	42	77	82,652	**73,884**
	Sub-corpus 1	43	85	48,800	**43,536**
	Sub-corpus 2	43	86	56,748	**50,448**
Group 2	Sub-corpus 3	43	84	65,920	**58,772**
	Sub-corpus 4	43	83	77,833	**69,610**
	Sub-corpus 5	43	81	79,612	**71,291**

Analysis

To find out the collective trends, frequency analyses were done with the compiled learner corpora, and Sketch Engine online corpus interface was used to analyze the data. To list the three- to four-word sequences in each sub-corpus of two groups, the researchers used n-gram function of Sketch Engine. Then, their structural and functional classifications following the taxonomy submitted by Biber et al. (1999) and Biber et al. (2004) were manually annotated. Lastly, Pearson correlation coefficient statistics were performed to reveal whether there is a relationship between the frequent common sequences among each group and native written corpora (LOCNESS).

Results and Discussion

The first research question asked the most frequent 3- and 4-word sequences. The lists of the most frequent shared 3- to 4-word FSs across both groups were extracted. Table 4 below illustrates the top 10 frequent formulaic sequences overtime in Group 1, listed depending on their normalized frequency.

Table 4

Top 10 Frequent Formulaic Sequences over Time in Group 1

Sub-corpus 1		Sub-corpus 2		Sub-corpus 3		Sub-corpus 4		Sub-corpus 5	
3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f
one of the	39,66	one of the	37,81	one of the	38,41	one of the	47,87	in the world	90,74
a lot of	29,22	a lot of	34,90	a lot of	38,41	the most imp.	36,08	one of the	42,35
there is a	26,09	in terms of	25,21	in order to	32,14	the end of	27,06	cannot be	25,41
day by day	21,92	it is a	22,30	in the world	24,30	acc. to the	27,06	acc. to the	25,41
should be given	17,74	the number of	22,30	because of the	24,30	it is not	26,37	it is not	24,20
to sum up	17,74	in order to	21,33	in terms of	22,73	in the future	23,59	I do not	22,99
there is no	17,74	acc. to the	20,36	the most imp.	22,73	a lot of	23,59	the fact that	21,17
it is not	16,70	first of all	19,39	it is a	18,03	there is a	22,20	of the world	21,17
they do not	15,65	want to have	18,42	first of all	16,46	in my opinion	20,12	as a result	19,96
have to do	14,61	is one of the	17,45	day by day	14,89	the quality of	20,12	there is no	19,36

When the list of three- to four-word FSs was retrieved, it was seen that the most frequent FS was *one of the* in each sub-corpus of Group 1. When looking at the length of these sequences, it is seen that nine out of 10 were 3-word FSs, with only one 4-word sequence in sub-corpus 2 of Group 1, whereas all of the sequences in sub-corpus 1, 3, 4 and 5 were 3-word FSs. This finding is supported by Conrad and Biber (2005), who found that 3-word FSs are more frequent in the corpora as evidenced in NES written corpora of academic prose.

Table 5

Top 10 Frequent Formulaic Sequences over Time in Group 2

Sub-corpus 1		Sub-corpus 2		Sub-corpus 3		Sub-corpus 4		Sub-corpus 5	
3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f	3-4 FSs	Norm f
one of the	35,86	a lot of	37,01	one of the	49,30	one of the	53,96	in the world	67,20
a lot of	29,71	one of the	35,24	a lot of	31,10	the most imp.	37,90	one of the	52,13
there is no	28,69	in terms of	35,24	in order to	30,34	in the future	34,69	there is no	31,40
they do not	23,57	the number of	32,60	in the world	22,75	it is not	31,48	in terms of	30,77
in order to	22,54	in order to	24,67	is one of the	21,24	the quality of	31,48	cannot be	30,77
to sum up	21,52	to sum up	21,15	to sum up	20,48	a lot of	23,77	there is a	30,77
we do not	21,52	in my opinion	20,27	one of the most	18,96	in my opinion	21,84	it is not	28,26
is one of the	19,47	to go to	20,27	in terms of	18,96	there is no	21,84	the fact that	25,75
day by day	18,44	they do not	18,50	of the most	18,96	in terms of	21,20	according to the	23,24
it is not	17,42	it is a	17,62	according to the	17,45	is the most	20,56	in order to	21,98

Table 5 displays the most common formulaic sequences in all sub-corpora of Group 2 over time. The most frequent three- to four-word FSs was *one of the.* This is also in line with the findings of Group 1. When looking at the length of FSs it was found that nine out of 10 were three-word sequences, with only 1 four-word sequence in sub-corpus 1 of Group 2. In sub-corpus 3, eight out of 10 were three-word sequences, with only 2 four-word sequences, whereas all of the sequences in sub-corpora 2, 4 and 5 were three-word FSs.

Table 6

Shared Frequent 3- and 4-Word Formulaic Sequences over Time in Group 1

3-4 Formulaic Sequences	Norm f Sub-corpus 1	Norm f Sub-corpus 2	Norm f Sub-corpus 3	Norm f Sub-corpus 4	Norm f Sub-corpus 5

one of the	39,66	37,81	38,41	47,87	42,35
a lot of	29,22	34,90	38,41	23,59	18,15
according to the	11,48	20,36	14,89	27,06	25,41
the most important	11,48	16,48	22,73	36,08	9,68
in order to	11,48	21,33	32,14	16,65	16,33
in terms of	14,61	25,21	22,73	11,79	16,33
it is a	14,61	22,30	18,03	17,35	18,75
there is a	26,09	15,51	12,54	22,20	14,52
as a result	14,61	17,45	14,89	18,73	19,96
it is not	16,70	9,69	7,84	26,37	24,20
in the world	7,31	11,63	24,30	7,63	90,74
there is no	17,74	11,63	10,19	15,26	19,36
day by day	21,92	15,51	14,89	10,41	8,47
I believe that	9,39	11,63	8,62	15,96	19,36
in my opinion	14,61	10,66	9,41	20,12	13,91
they do not	15,65	10,66	10,19	13,88	13,31
there are many	11,48	16,48	7,06	17,35	16,94
to sum up	17,74	16,48	13,33	9,71	7,86
the fact that	12,52	10,66	14,89	9,71	21,17
that it is	14,61	12,60	14,89	11,79	12,70
of the most	11,48	10,66	13,33	11,79	13,91
one of the most	10,44	9,69	11,76	11,10	13,31

Table 6 lists the most frequent 3- to 4-word sequences shared across five sub-corpora of Group 1. In the sub-corpora of Group 1, two of the (*to sum up* and *day by day*) shared formulaic sequences tend to be less frequent later over time. Some of the FSs less frequent in sub-corpus 1 tend to become more frequent in later sub-corpora such as *one of the*, *according to the*, *in order to*, *it is a* and *in the world*. In addition, the frequency analysis demonstrated a fluctuation in the normalized frequency scores of the majority of shared formulaic sequences. For example, *according to the* showed a fluctuating pattern in all sub-corpora.

Table 7

Shared Frequent 3- and 4-Word Formulaic Sequences over Time in Group 2

3-4 FSs	Norm f Sub-corpus 1	Norm f Sub-corpus 2	Norm f Sub-corpus 3	Norm f Sub-corpus 4	Norm f Sub-corpus 5
one of the	35,86	35,24	49,30	53,96	52,13
a lot of	29,71	37,01	31,10	23,77	19,47
there is no	28,69	7,93	12,89	21,84	31,40
they do not	23,57	18,50	15,93	14,13	12,56
in order to	22,54	24,67	30,34	11,56	21,98
to sum up	21,52	21,15	20,48	10,28	11,30
is one of the	19,47	13,22	21,24	7,71	13,19
it is not	17,42	9,69	16,69	31,48	28,26
there is a	16,39	15,86	15,93	14,78	30,77
as a result	16,39	13,22	14,41	7,71	16,96
in terms of	16,39	35,24	18,96	21,20	30,77
it is a	16,39	17,62	15,17	16,06	16,96
on the other hand	15,37	11,45	9,10	16,06	13,82
that it is	14,34	7,05	11,38	19,27	15,70
according to the	13,32	14,10	17,45	15,42	23,24
most of the	13,32	14,10	9,86	12,85	6,91
because of the	12,30	15,86	12,89	16,06	20,10
I strongly believe that	12,30	17,62	11,38	8,35	12,56
in my opinion	11,27	20,27	9,86	21,84	15,07
the fact that	9,22	7,05	11,38	16,06	25,75

Table 7 lists the most frequent 3- to 4-word sequences shared across five sub-corpora of Group 2. It is seen that two of the (*a lot of, they do not*) shared formulaic sequences tend to be less frequent later over time. Some of the formulaic sequences less frequent in sub-corpus 1 tend to become more frequent in later sub-corpora such as *one of the, in terms of, according to the* and *the fact that*. The frequency analyses illustrated a fluctuation in the normalized frequency scores of the majority of shared formulaic sequences. This is in line with the findings of Group 1. For instance, *a lot of* was less frequent in sub-corpus 1, increased in sub-corpus 2, and then dropped its frequency through sub-corpus 3, 4 and 5. While the sequence *that it is* was more frequent in sub-corpus 1, there was a steady decrease in sub-corpus 2, and then increased through sub-corpus 3 and 4.

In conclusion, when the top 10 shared three- to four-word formulaic sequences in Group 1 and Group 2 were retrieved and analyzed, it was found that the most frequent three- to four-word formulaic sequences were *one of the, a lot of, in terms of, it is not, in order to, the most important, there is no, there is a, according to the* and *it is a*. Among these, three sequences (*in terms of, in order to* and *there is a*) in the top 10 list were also common in the Academic Formulas List (AFL) developed by Simpson-Vlach and Ellis (2010).

Structural and Functional Analysis

The second research question asked the structures and functions of the frequent formulaic sequences that were evaluated through Biber's et al. (1999; 2004) taxonomy.

Structural Analysis

Table 8 below displays the structural classifications of the top 100 frequent formulaic sequences in sub-corpus 5 of Group 1 and Group 2.

Table 8

Structure of the Top 100 Frequent Formulaic Sequences in Sub-corpus 5 of Group 1 and Group 2

Structural Types	Group 1	Group 2
Personal pronoun+ VP (+complement-clause fragment)	*I do not agree, I believe that, I firmly believe that, he says that, and they can, I do not, they do not, they cannot, they are not, I think that, he claims that, I cannot, we look at, when we look,*	*I do not agree, I believe that, I firmly believe that, he says that, he fails to mention, I think that, I agree with, we can see, he thinks that, I strongly believe that, they do not, I do not, they cannot, they are not, I cannot, they want to, we cannot, we do not, and they are,*
VP with active verb	*do not agree with, should not be, do not have, will not be, cannot find, do not think, look at the, people do not, cannot be,*	*have the same, do not agree with, should not be, cannot be, does not mean, not want to, can be a, do not have, do not want to, will be a, will not be,*
With wh-clause fragments	*who is a, people who are,*	*who is a, who is the, when it comes to,*
Quantifier expressions	*a lot of, some of the,*	*a lot of,*
NP with of-phrase fragment	*one of the, the number of, one of the most, the problem of, because of the, the end of, the rate of, the use of,*	*one of the, the number of, of the world, because of their, because of the, of the most, one of the most, the problem of, the rate of,*
NP with other post-modifier fragment	*the fact that,*	*the fact that,*

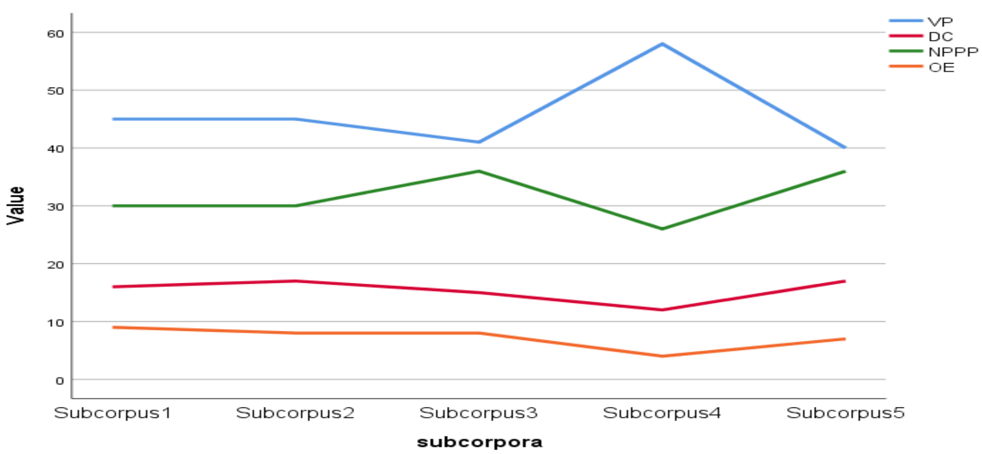
Other NP expressions	*the most important,*	*the most important,*
PP with embedded of-phrase fragment	*in terms of, as a result of,*	*in terms of, as a result of,*
Other PP fragment	*in the world, about this issue, on the other hand, in the same, around the world, about this topic, in this way, at the same time, of the most, in my opinion, in order to, for this reason, the reason for, in addition to, in his article, in the future, in the work, most of the, of the world, on the contrary, people in the,*	*in the world, about this issue, on the other hand, in the same, around the world, most of the, in my opinion, on the contrary, in order to, in other words, over the world,*
Anticipatory it+ VP/Adj.P.	*it is a, and it is, it is not, it can be, because it is, it does not,*	*it is a, it is not, but it is, and it is, it can be, it does not,*
Passive verb+ PP fragment	*based on my,*	*based on my, was published in,*
Copula be+ NP/ Adj.P.	*is one of, is not a, be the answer, is not the,*	*is one of the, is not a, is not the,*
(VP+) that-clause fragment	*that it is, that there are, that they are, that there is, can say that, is that the, claims is that, to say that,*	*that it is, that there is, fails to mention that, that they are, that there is a, that there is no, think that the,*
(verb/adjective+) to-clause fragment	*to be a, be able to, a solution to the, the answer to the, the only way to, to increase the,*	*need to be, to be a, a solution to the, be a solution to, be able to, to have a,*
Adverbial clause fragment	*day by day,*	*at the same time, day by day, in the future,*
Pronoun/noun phrase+ be (+ . . .)	*there is a, there is no, there are many, because they are, this is a,*	*there is a, there is no, there are many, this is a, because they are, there are some,*
Other expressions	*as well as, as a result, according to the, according to a, to sum up, due to the, thanks to the,*	*as well as, as a result, according to the, according to a, to sum up, according to my, according to his, as long as, in this way,*
Comparative expressions	*as much as,*	

There are 18 structural categories in sub-corpus 5 of Group 1 and 17 categories in Group 2. Whereas the most frequent structure was *other prepositional phrase fragment* with 21 types in Group 1, *personal pronoun + verb phrase (+ complement-clause fragment)* with 19 types was the most recurrent ones in group 2. For group 1, the followings were some of the examples of *other prepositional phrase fragment*: *in the world, on the other hand, in the same* and *around the world*. Examples of *personal pronoun + verb phrase (+ complement-clause fragment)* were: *I do not agree, I believe that, he says that* and *he fails to mention*. The second frequent sub-category was *personal pronoun + verb phrase (+ complement-clause fragment)* with 8 types in Group 1 while *verb phrase with active verb* and *other prepositional phrase fragment* was used in the same proportions with 11 types in Group 2.

The overall distribution of three- to four-word formulaic sequences in terms of structural categories in Group 1 is displayed in Figure 1 below.

Figure 1

The Distribution of the Type Frequency of the Structural Categories of the Top 100 Frequent Formulaic Sequences across Sub-corpora of Group 1

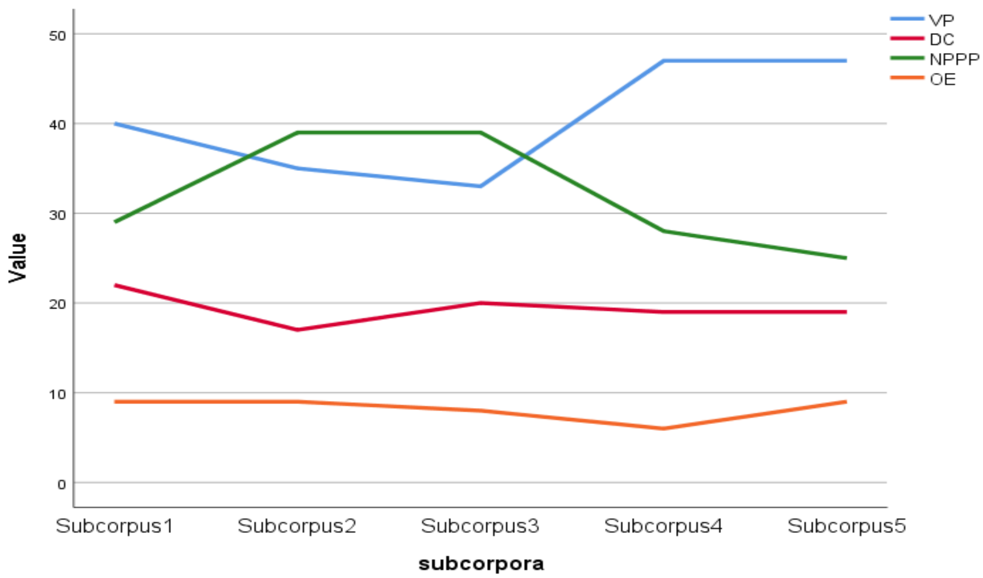


It is seen that the usage of three- and four-word sequences fluctuated over time. The majority of formulaic sequences found in each sub-corpus of group 1 took place in the category of *VP fragments* while *NPPP fragments* were the second common one, and it was followed by *DC fragments* and *OE* categories. In the category of *VP fragments*, while it was the same proportions in the beginning, they slightly decreased after sub-corpus 2. Then, it increased steadily in sub-corpus 4, but this steady increase was followed by a steady decrease in sub-corpus 5. It is significant to declare that there was no such steady increase or decrease in other categories, and though the types of *VP fragments* increased, the other three categories decreased in similar proportions. Similar to these findings, Biber et al. (2004) found that nearly 90 per cent of all formulaic sequences in conversation contained *VP fragments* in contrast and almost 70 per cent in academic prose comprised of NP fragments like *the nature of the*.

The overall distribution of three- to four-word formulaic sequences in terms of structural categories in Group 2 is illustrated in Figure 2 below.

Figure 2

The Distribution of the Type Frequency of the Structural Categories of the Top 100 Frequent Formulaic Sequences across Sub-corpora of Group 2



The use of three and four-word sequences in main categories from sub-corpus 1 to sub-corpus 5 in group 2 fluctuated. While the majority of formulaic sequences found in sub-corpus 1 were in the category of *VP fragments*, there was a change in sub-corpus 2 where *NPPP fragments* were used most frequently, and the use of *VP fragments* decreased. The category of *NPPP fragments* from sub-corpus 2 to sub-corpus 3 remained the same proportions. Then, it was seen a downward trend in this category, and *NPPP fragments* fell behind the *VP fragments*. There were some changes in the usage of formulaic sequences between these two categories across each sub-corpus. While *DC fragments* were the third most commonly used category, *OE* was used in the lowest proportions, and *DC fragments* and *OE categories* fluctuated over time. Overall, the common usage of the category of *VP fragments* and *NPPP fragments* were consistent with the findings of Fattani (2018), who found that "in terms of N-gram types, the VP- and PP-based forms are the most common structures (2018, p.113).

Functional Analysis

Table 9 displays the function of the top 100 three- and four-word frequent formulaic sequences in sub-corpus 5 of Group 1 and Group 2.

Table 9

Function of the Top 100 Frequent Formulaic Sequences in Sub-corpus 5 of Group 1 and Group 2

Functional Types	Group 1	Group 2

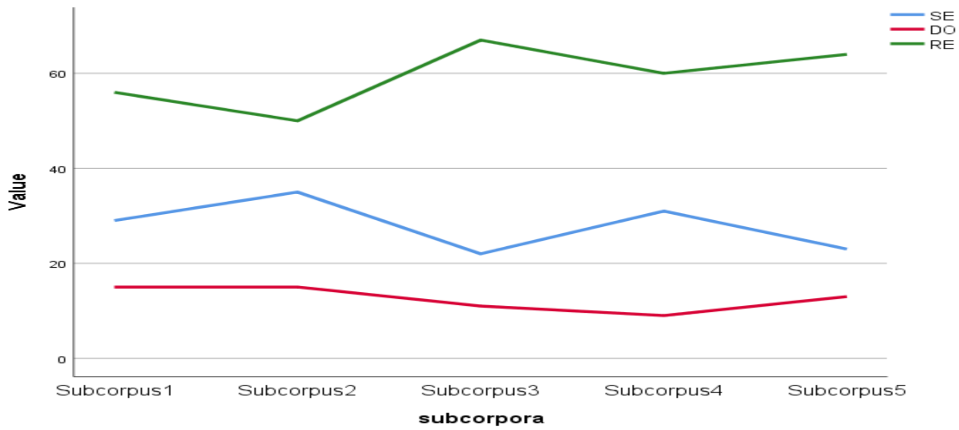
1. Stance expressions	*according to the, according to a, I believe that, I firmly believe that, I think that, the fact that, can say that, it can be, do not think, to say that, should not be, I do not agree, I do not, do not agree with, will not be, and they can, they cannot, cannot be, to be a, be able to, cannot find, I cannot, in my opinion,*	*according to the, according to a, according to my, according to his, I believe that, I firmly believe that, I think that, we can see, I strongly believe that, the fact that, think that the, do not have, not want to, do not want to, they want to, should not be, need to be, I do not agree, I agree with, I do not, do not agree with, we do not, will be a, will not be, they cannot, cannot be, to be a, be able to, can be a, I cannot, we cannot, he thinks that, in my opinion,*
2. Discourse organizers	*look at the, because they are, on the other hand, as well as, as a result, in order to, to sum up, reason for this, the reason for, as a result of, in addition to, on the contrary, thanks to the,*	*because they are, on the other hand, as well as, as a result, in order to, to sum up, in other words, on the contrary, when it comes to, as a result of, as long as, in this way,*
3. Referential expressions	*one of the, one of the most, it is a, and it is, it is not, is one of, that it is, that there are, that they are, that there is, is that the, there is a, there is no, the most important, because of the, they are not, do not have, they do not, who is a, in the same, is not a, of the most, a solution to the, be the answer, because it is, claims is that, due to the, he claims that, in the work, is not the, it does not, most of the, people do not, people in the, people who are, the answer to the, the only way to, this is a, to increase the, we look at, when we look, as much as, a lot of, there are many, the number of, some of the, based on my, in terms of, about this issue, the problem of, about this topic, in this way, the rate of, the use of, in the world, around the world, of the world, day by day, at the same time, in the future, in his article, the end of, he says that,*	*one of the, it is a, it is not, is not a, is one of the, that it is, that there is, that they are, that there is a, that there is no, there is a, there is no, because of their, they are not, they do not, this is a, have the same, who is a, who is the, in the same, is not the, most of the, a solution to the, and it is, was published in, be a solution to, to have a, it can be, it does not, of the most, one of the most, the most important, there are some, and they are, a lot of, there are many, the number of, the rate of, based on my, in terms of, about this issue, the problem of, in the world, of the world, over the world, at the same time, day by day, in the future, he says that, he fails to mention, fails to mention that, but it is, does not mean, around the world,*

In group 1, the most frequent function category of three- to four-word FSs was *referential expressions* with 64 types while *stance expressions* were the second most commonly used one with 23 types, followed by *discourse organizers* with 13 types. *One of the, one of the most, it is a, and it is* and *it is not* were some of the examples in the category of *referential expressions*. Examples of the category of *stance expressions* were: *cannot be, to be a, be able to, cannot find* and *in my opinion*. Lastly, the followings were among the examples of *discourse organizers*: *look at the, in order to* and *to sum up*. In the same vein with group 1, *referential expressions* were the most common function with 55 types while *stance expressions* were the second one with 33 types and followed by discourse organizers with 12 types. The examples of these categories in Group 2 were similar to Group 1, which means participants used similar three- to four-word formulaic sequences.

The overall distribution of three- to four-word formulaic sequences in terms of functional categories in Group 1 is illustrated in Figure 3 below.

Figure 3

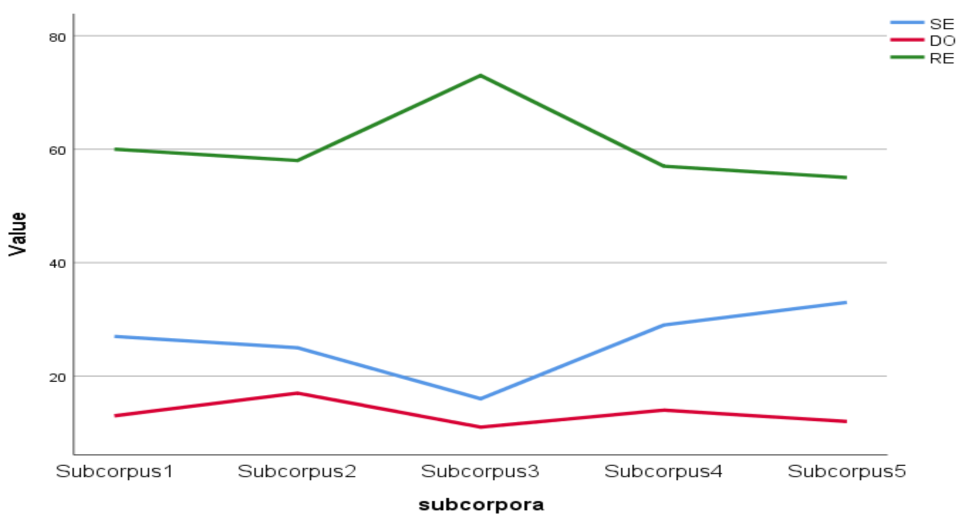
The Distribution of the Type Frequency of the Functional Categories of the Top 100 Frequent Formulaic Sequences across Sub-corpora of Group 1



The usage of three- and four-word formulaic sequences in the functional category fluctuated across each sub-corpus of Group 1. The majority of the three- to four-word formulaic sequences functioned as *referential expressions,* fluctuating over time. On the other hand, the use of this category increased steadily in sub-corpus 3 while it was a slight decrease in the other categories. In the last sub-corpus, the categories of *referential expressions* and *discourse organizers* increased but the *stance expressions* decreased. In the category of discourse organizers, there was no big fluctuation in contrast to two other categories and the fluctuations on the functional categories of three- to four-word formulaic sequences showed that the learners had various knowledge of *referential expressions* different from other categories.

Figure 4

The Distribution of the Type Frequency of the Functional Categories of the Top 100 Frequent Formulaic Sequences across Sub-corpora of Group 2



The use of three- to four-word formulaic sequences fluctuated from sub-corpus 1 to sub-corpus 5, and the results of Group 2 were in line with group 1. The majority of the formulaic sequences of Group 2 functioned as *referential expressions*, while the category of *stance expressions* was the second one and is followed by *discourse organizers*. The usage proportions of *stance expressions* and *discourse organizers* were similar over time, but *referential expressions* were the highest. In sub-corpus 3, there is a steady increase in the category of *referential expressions* while the other two categories slightly decreased. Bal-Gezegin (2019) indicated similar results in that the category of *referential expressions* are comprised the largest part, accounting for %75, followed by *discourse organizers*, accounting for %15 and *stance bundles* accounting for %8 in the academic writing of English L2 learners. In the same vein, Adel and Erman (2012) found that the majority of sequences functioned as *referential expressions* in advanced learner writing both by native speakers and non-native speakers.

Comparison of Native and Non-native Formulaic Sequences

The third research question explored the associations between learner corpora and LOCNESS in terms of the usage patterns of three- to four-word formulaic sequences. Based on the frequency analysis, the top 100 lists of both groups were extracted, their raw frequencies were normalized, and the Log10 was calculated to provide normal distribution. To measure the significance and strength level among shared formulaic sequences, a Pearson correlation analysis was conducted.

For Group 1, the findings of the Pearson correlation indicated that there was a significant and positive relationship, moderate in strength between Group 1 and LOCNESS ($r=.491$, $N=44$, $p=0.01$). In Figure 5, the scatterplot displays the results of correlation between each sub-corpus of Group 1 and LOCNESS.

Table 10

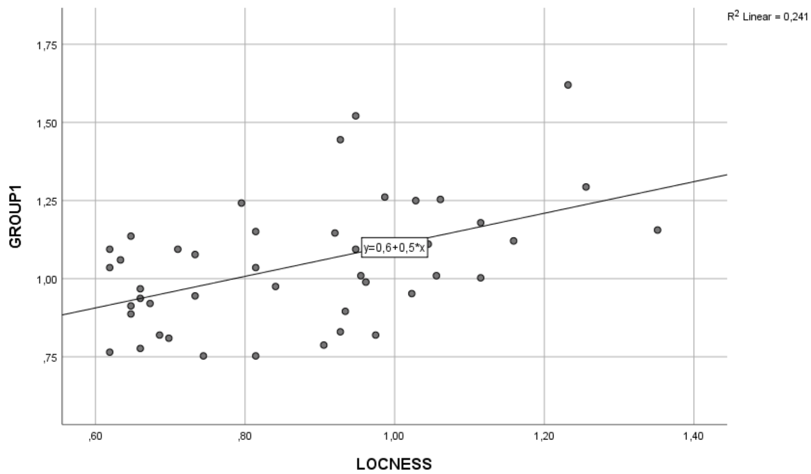
Correlations between Group 1 and LOCNESS

		LOCNESS	GROUP1
LOCNESS	Pearson Correlation	1	,491**
	Sig. (2-tailed)		,001
	N	44	44
GROUP1	Pearson Correlation	,491**	1
	Sig. (2-tailed)	,001	
	N	44	44

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5

Scatter Plot of the Relationship between Frequency Scores in LOCNESS and Group 1 in Learner Corpora



For Group 2, the findings of the Pearson correlation displayed that there was a significant and positive relationship, moderate in strength between Group 2 and LOCNESS (r=.449, N=39; r=0.01). In Figure 6, the scatterplot summarizes the results of the correlation between Group 2 and LOCNESS.

Table 11

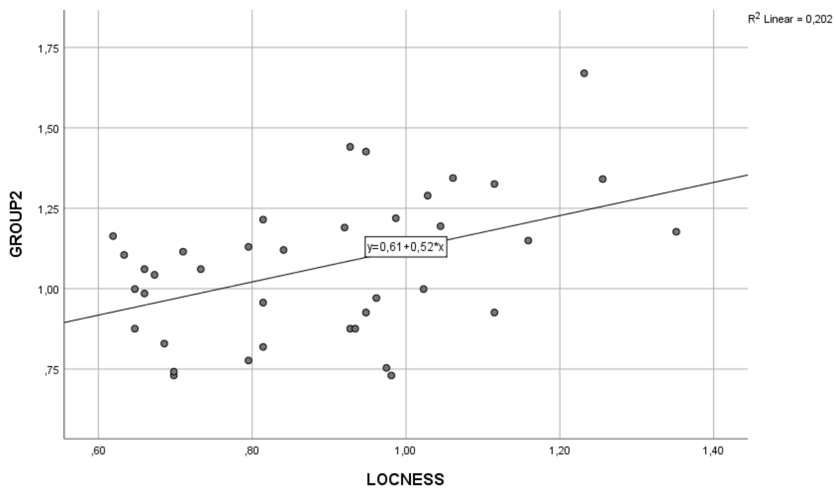
Correlations between Group 2 and LOCNESS

		LOCNESS	GROUP2
LOCNESS	Pearson Correlation	1	,449**
	Sig. (2-tailed)		,004
	N	39	39
GROUP2	Pearson Correlation	,449**	1
	Sig. (2-tailed)	,004	
	N	39	39

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6

Scatter Plot of the Relationship between Frequency Scores in LOCNESS and Group 2 in Learner Corpora



Discussion

We aimed to explore the use of three- and four-word formulaic sequences in learner corpora composed of EFL learners' essays across two semesters. Using a frequency-driven approach, the most frequent formulaic sequences were extracted in consecutive two semesters in 2018-2019. A total of ten argumentative essays for each participant were analyzed in terms of formulaic sequence content, frequency and type. The analysis included two groups of EFL learners with different levels of language proficiency. We used a frequency-based approach in the analysis of the data, and the frequencies obtained from the five sub-corpora from each group were normalized, and the resulting data were subjected to Pearson correlation analysis. The analysis also included structural and functional categorization of the FSs in the form of tables and graphics for each group. In doing so, we aimed to investigate collective trends in the use of FSs. All the analysis and the results yielded interesting results regarding the FSs developmental levels of EFL learners. The frequency analyses, correlation test, and the structural and functional analyses of the language formulas showed that the number and the range of FSs seemed to show an increasing pattern in number and type, as the learners were given more instruction and teacher feedback regarding their essays for each week during two semesters. As they get more teacher feedback, their general writing quality increases in line with the FSs development and this is also validated through correlation analysis.

The structural analysis showed how formulaic sequences were used by EFL learners. The structure of the majority of the frequent FSs in two semesters of observation contained the personal pronoun such *as we cannot, I cannot, we do not, they do not*, and *I believe that* (Biber et al., 1999). This finding is in agreement with Fattani's (2018) findings that VP-based structures are the most frequently occurring in the textbooks and the written AFL sub-list, and these structures account for the highest proportion of formulaic sequences. However, this finding is not consistent with the results of the study conducted by Cooper (2016), who investigated four-word multi-word units in the IELTS writing tests, student essays and published writing within the field of psychology. Cooper (2016) found that the dominant structural types in these corpora were prepositional phrase and noun phrase fragments, followed by verb phrase fragments. The Pearson correlation test revealed that the frequent FSs in the two learner corpora seemed to moderately correlate in Group 1 and 2 with the native learner corpora in the four sub-corpora.

The rationale behind the frequent usage of FSs such as *in order to, one of the, the use of, the fact that, there is a, there is no and on the other hand might be due to the fact that those units* is that they are highly frequent and, thus, salient in the written input of EFL learners. In this sense, the findings are consistent with the study of Biber et al. (1999), who stated that such formulaic sequences as *in order to, one of the, part of the, the number of, the presence of, the use of, the fact that, there is a, there is no* were the most common three-word sequences in academic prose while *in the case of* and *on the other hand* were the most common four-word lexical units in academic prose.

The frequency analysis carried out in the study is due to the fact that high-frequency FSs can be learned and processed more easily than less frequent FSs according to O'Donnell et al. (2013). Such features as frequency, familiarity, conventionality, prototypicality /stereotypicality (Giora, 2003) are the significant factors for EFL learners to learn them. According to Giora (2003), "the more frequent, familiar, conventional, or prototypical /stereotypical the information in the mind of the individual or in a certain linguistic community, the more salient it is in that mind or among the community members" (2003, pp.15-16). This is also supported by O'Donnell et al. (2013) who stated that "humans learn more easily and process more fluently high-frequency forms and "regular" which are exemplified by many types and which have few competitors" (2013, p.89). The findings clearly revealed that FSs were used saliently in the academic and expository essays in the two groups. Many FSs, especially in Group 1 sub-corpora were obtained with frequent FSs such as *one of the, a lot of, there is a* and *day by day*. Formulaic patterns which are similar to the academic ones such as *for this reason, as long as, it is undeniable* and *there will be* were not frequent in both groups since they demand extensive reinforcement not easily retrieved by the learners.

The examination of the most frequent FSs across two semesters gave the researcher various FSs differing in length and type. This finding is concurrent with the finding of Conrad and Biber (2005), who found that 3-word FSs are more frequent in the corpora as evidenced in NES written corpora of academic prose. The main 4-word formulaic sequences that were frequently used in learner corpora were *on the other hand*, *is one of the*, *I strongly believe that, I firmly believe that* and *one of the most*. This finding is also similar to the findings of Juknevičienė's (2009) and Pavesi's (2013). The frequent use of 3- and 4-word FSs may be given to the fact that they can be found in naturally occurring spoken and written language, and thus their exposure must have been easier for EFL learners. Regarding the types of FSs, our corpora gave frequent but limited FSs usage patterns. FSs were almost similar types of FSs or repeated across two semesters of both groups. This may be given to the L2 learners limited stock of FSs, and this conclusion is also supported by Ellis (2012), who stated that learners tend to use were common FSs and familiar constructions.

The results discussed so far in the study partly confirm our hypothesis in that EFL learners seem to show a reliance on frequent FSs. With this in mind, however, the researchers noticed that the type of FSs did not seem to increase to a great extent which suggests that with more instruction, the frequency of FSs increase far more than it does with types. From the functional perspective, the FSs indicated that writers used a high number of referential expressions followed by stance expressions and discourse organizers. The referential functions such as *one of the, it is not, it is a, is one of the, there is a* and *there is no* were among the most frequent ones across two semesters. One of the possible reasons for this may be that the learners were frequently exposed to these FSs functions in their previous instructions. FSs such as *one of the, a lot of* and *in terms of* are some of the most frequent examples from the referential category and can be easily remembered by learners while writing. Stance expressions, on the other hand, become slightly less frequent in all sub-corpora. Such stance expressions such as *the fact that, I do not want, do not agree with* and *should not be* were the most frequent. FSs of discourse organizers were the least employed ones by the EFL learners, and their frequency and types remain relatively lower during two semesters. From the early stages of L2 development, EFL learners rely on referential FSs. This finding is consistent with the results of Vidaković and Barker's (2009), who stated that "learning conventionalized word strings starts emerging after the lowest proficiency level" (2009, p.144). Still, the FSs employed so far were the most common and invariant ones. Our findings also showed that the *most frequent referential expressions* correspond to Vidaković and Barker's (2009) findings arguing that referential formulas were dominant in the written learner corpus at all levels. These results matched with the findings of Tomankova (2016), who revealed that referential expressions present the most frequently occurring functional type. A study conducted by Breeze (2013) investigating multi-word units employed in four legal corpora indicated the same results. Biber and Barbieri (2007) studied different registers and found strikingly different outcomes in terms of functional types in that institutional writing comprised approximately 70% of *referential expressions*, whereas written course management involved over 70% of stance expressions. They also claimed that referential expressions were dominant in academic writing (e.g., academic prose and textbooks). This finding is also in agreement with Kashiha and Heng (2014) findings which showed *referential expressions* were the most common functional type in the two disciplines, namely politics and chemistry lectures. Lastly, the study of Fattani (2018) is a good example of functional usage of formulaic sequences in different registers, and in the instructors' materials and the written AFL sub-list, the proportion of *referential expressions* was high compared to *stance expressions* and *discourse organizers*. In contrast to these two registers, the distribution of functional categories seemed to have different findings in that *stance expressions* were the most common in the textbooks. The stance expressions were the second, and the discourse organizers were the least employed across two semesters, corresponding with the findings of Biber and Barbieri (2007), indicating usage of over 10% of *stance expressions* and *discourse organizers*. The findings of the current study also were consistent with those of Kashiha and Heng (2014) that *stance expressions* were the second common functional type, including politics and chemistry of FSs.

Conclusion

The study aimed to investigate the usage of formulaic sequences in learner corpora composed of EFL learners' essays across two semesters of observation. To this aim, the analysis phase included group analysis. In the group analysis phase, the study examined collective usage of formulaic sequences over five sub-corpora of each group. The extracted formulaic sequences were classified into the Biber's et al. structural and functional taxonomy (1999; 2004). The usage of three- and four-word formulaic sequences were compared to frequent sequences found in the native corpus (LOCNESS). Group analysis yielded findings in terms of frequency, type, structure, and functions of frequent formulas used by the learners across two semesters. These findings can be attributed to the existence of FSs in the learner corpora in two groups. The learner corpora in the two groups contained many FSs, especially in the sub-corpus 3, 4 and 5, which were limited in type and length. It was seen in the learner corpora that three-word sequences were in majority.

On the other hand, the correlation analysis for Group 2 showed no relationship between sub-corpus 1, 2, 3 and 4 and native corpus, whereas there was a significant and positive relationship between sub-corpus five and native corpus, and the correlation was of moderate strength. Our study is limited to a corpus of EFL students registered at only one state university in Turkey, and so, another advanced level tertiary level EFL students were not included in this study. For this reason, the results of this study cannot be generalized to all advanced level university students in Turkey. Secondly, this study was strictly limited to investigating formulaic sequence aspects of the essays of tertiary level Turkish EFL students, which means that no other aspects (e.g., pragmatics, discourse markers, syntax) were targeted in this study. Thirdly, all the data collected with untimed essays were limited to a certain number and type, so they cannot be generalized to other essay design criteria such as timed essays. Another limitation was that operational restrictions of this study did not allow time for comparison between learner corpora and reference spoken corpora which would provide specific insight into spoken data with regards to how they are similar or different.

The results of this study have demonstrated that FSs usage patterns increase as the language proficiency of learners and teacher feedback increase in consecutive weeks. But it seems that there is a need for language teachers to create an air of repeated exposure in the classroom for the sequences. This must be done on a systematic basis by attaching importance to the FSs from different structural and functional categories. In other words, there is a need for immediate pedagogical focus on the use of FSs by the language teachers during their instruction.

References

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 84-91. DOI: 10.1016/J.ESP.2011.08.004

Bal-Gezegin, B. (2019). Lexical bundles in published research articles: A corpus-based study. *Journal of Language and Linguistic Studies, 15*(2), 520-534. https://doi.org/10.17263/jlls.586188

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286. DOI: 10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics, 25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English.* Longman.

Breeze, R. (2013). Lexical bundles across four legal genres. *International Journal of Corpus Linguistics, 18*(2), 229-253. https://doi.org/10.1075/ijcl.18.2.03bre

Chenu, F., & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition Et Interaction En Langue Étrangère, 1*, 17-38. https://doi.org/10.4000/aile.4506

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71. DOI: 10.1515/9783484604674.56

Cooper, P. A. (2016). *Academic vocabulary and lexical bundles in the writing of undergraduate psychology students* [Unpublished doctoral dissertation]. University of South Africa.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. https://doi.org/10.2307/3587951

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*(1), 91-126. https://www.jstor.org/stable/44487860?seq=1#metadata_info_tab_contents

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal Teddy Bear. *Annual Review of Applied Linguistics, 32*, 17-44. DOI: 10.1017/S0267190512000025

Ellis, N. C. (2013). *Construction grammar and second language acquisition*. Oxford University Press.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly, 42*(3), 375-396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk, 20*(1), 29-62. https://doi.org/10.1515/text.1.2000.20.1.29

Fattani, R. A. I. (2018). *Profiling lexical bundles in an EAP pre-sessional course: A corpus-based study on textbooks and instructors' materials* [Unpublished doctoral dissertation]. University of Sheffield.

Giora, R. (2003). *On our mind: Salience, context, and figurative language.* Oxford University Press.

Granger, S. (2002). A bird's-eye view of learner corpus research. In, S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 3-33). John Benjamins Publishing. https://doi.org/10.1075/lllt.6.04gra

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 212-226). Routledge. DOI: 10.4324/9780203856949.ch16

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21. DOI: 10.1016/j.esp.2007.06.001

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences acquisition, processing and use* (pp. 269-300). Benjamins.

Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra, 61*(3), 61-72. https://doi.org/10.15388/Klbt.2009.7638

Kashiha, H., & Heng, C. S. (2014). Discourse functions of formulaic sequences in academic speech across two disciplines. *GEMA Online® Journal of Language Studies, 14*(2), 15-27. http://dx.doi.org/10.17576/GEMA-2014-1402-02

Liou, H. C., & Chen, W. F. (2018). Effects of explicit instruction on learning academic formulaic sequences for EFL college learners' writing. *Taiwan Journal of TESOL, 15*(1), 61-100. DOI: 10.30397/TJTESOL.201804_15(1).0003

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy* [Unpublished doctoral dissertation]. University of Nottingham.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. DOI: 10.1093/applin/ams010

McCarthy, M. (1991). *Discourse analysis for language teachers Cambridge*. Cambridge University Press.

Mel'cuk, I. (1998). Collocations and lexical functions. In A.P. Cowie (Ed.), *Phraseology. theory, analysis, and applications* (pp. 23-53). Clarendon Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Oakey, D. (2002). Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation: Studies in corpus linguistics* (pp. 111-130). John Benjamins.

O'Donnell, B. M., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18*(1), 83-108. https://doi.org/10.1075/ijcl.18.1.07odo

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

Özbay, A. Ş., & Kayaoğlu, M. N. (2016). Computerized corpus based investigation of the use of multi word combinations and the developmental stages by tertiary level EFL learners. *EPiC Series in Language and Linguistics, 1*, 341-350. https://doi.org/10.29007/fs4s

Pavesi, C. (2013). *Recurrent sequences in a learner corpus of computer-mediated communication* [Unpublished doctoral dissertation]. Università Cattolica del Sacro Cuore.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1-22). John Benjamins.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 173-190). John Benjamins.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. https://doi.org/10.1093/applin/amp058

Sinclair, J. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407-410). John Benjamins Publishing Company. https://doi.org/10.1075/z.139.33sin

Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning, 70*(2), 420-463. https://doi.org/10.1111/lang.12383

Tomankova, V. (2016). Lexical bundles in legal texts corpora–selection, classification and pedagogical implications. *Discourse and Interaction, 9*(2), 75-94. https://doi.org/10.5817/DI2016-2-75

Vidaković, I., & Barker, F. (2009). Lexical development across second language proficiency levels: A corpus-informed study. In A. Harris & A. Brandt (Eds.), Language, learning & context: Proceedings of the 42nd annual meeting of the British association for applied linguistics (pp. 143-146). Scitsiugnil Press. https://www.baal.org.uk/wp-content/uploads/2017/12/proceedings_09_full.pdf

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier and S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 123-147). John Benjamins.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*(1), 1-28. DOI: 10.1016/S0271-5309(99)00015-4

Wong, H. P. (2012). *Use of formulaic sequences in task-based oral production of Chinese* [Unpublished doctoral dissertation]. Durham University.