

*The development of productive bigram knowledge in  
EFL: Exploring the bigram continuum*

**Brahim Ait Hammou**

School of Arts and Humanities, Moulay Ismail University, Meknes, Morocco

<https://orcid.org/0000-0001-9977-5209>

[hammou76@gmail.com](mailto:hammou76@gmail.com)

**Mohammed Larouz**

School of Arts and Humanities, Moulay Ismail University, Meknes, Morocco

<https://orcid.org/0000-0003-2017-8976>

[m.larouz@flsh.umi.ac.ma](mailto:m.larouz@flsh.umi.ac.ma)

**Mustapha Fagroud**

National School of Agriculture, Moulay Ismail University, Meknes, Morocco

<https://orcid.org/0000-0002-8084-8532>

[mfagroud@hotmail.com](mailto:mfagroud@hotmail.com)

**Abstract**

Using a written corpus of 529 texts, this study is the first to examine undergraduate university learners' productive bigram knowledge in the Moroccan EFL context. Internationally, while previous research focused on strongly associated bigrams, the current study adopts a continuum perspective and examines strongly and moderately associated bigrams as well as wrong word combinations, using a cross-sectional approach with three adjacent groups. The analysis of strongly and moderately associated rare bigrams shows that learners tend to increase their MI mean score towards the production of more strongly associated bigrams, as we compare lower to upper proficiency groups. For frequent bigrams, the results show that there is a *reversed effect* of proficiency on learners' performance as learners have a tendency to produce less fixed frequent combinations as they gain better proficiency, which suggests that they are using frequent word combinations in novel, more flexible ways. For wrong combinations, the three groups manage to form more

accurate combinations as they move towards higher proficiency levels both when rare and frequent bigrams are considered. The results indicate that the development of phraseological knowledge follows a continuum trajectory. The study has implications for learning and teaching EFL.

*Keywords:* bigrams; collocations; EFL; writing; phraseological knowledge

## 1. Introduction

Investigating the role of word co-occurrences has been introduced into linguistics with the advent of phraseology, mostly during the second half of the 20th century. Sinclair (e.g., 1991, 1996, 1998) was among the first to highlight the idiomatic nature of language and to note that language is made up of patterns. Similarly, Wray (2000) maintains that the notion of a native-like proficiency in a foreign (or second) language is no longer based on the speaker's knowledge of a set of linguistic rules, but it is rather based on their knowledge of a set of pre-fabricated expressions and formulas. A wide variety of terms are used to refer to the phenomenon of word combinations in language among which are *holo-phrases*, *pre-fabricated routines*, *routine formulaes*, *lexical chunks*, and *lexical phrases* (Schmitt, 2000). However, as Weinert (1995, p. 182) notes, "while labels vary, it seems that researchers have very much the same phenomenon in mind."

Another concept that has gained wide recognition in phraseology is the concept of *n-grams*. Paquot and Granger (2012, p. 138) used *recurrence* to refer to n-grams, which they define as "the repetition of contiguous strings of words of a given length." The term n-gram is reserved only for contiguous word combinations. Another characteristic of n-grams is related to their syntactic form. Biber and Conrad (1999, p. 183) highlight that n-grams may not be complete structural units. The researchers view lexical bundles as extended collocations which occur in continuous sequences. Extending the use of lexical bundles, Hyland (2008) maintains that *lexical bundles*, *clusters*, or *chunks* are extended collocations which appear in language more frequently than expected by chance. He states that lexical bundles "appear to represent a psychological association between words and reflect a very real part of users' communicative competences" (2008, p. 5). In the present study, we use bigrams to refer to any two contiguous words which re-occur more frequently than could be predicted by chance. While the term 'wrong bigram' is used for erroneous word combinations. We may also use the term *collocation* interchangeably with the term *bigram*, while keeping in mind the contiguous criteria as a defining aspect for this type of collocation.

While previous studies which examined the development of bigram or collocational knowledge focused mainly on strongly-associated combinations, the present study looks at the development of bigram knowledge as a continuum. It, thus, highlights the importance of strongly associated and moderately associated bigrams as well as erroneous word combinations in the development of FL learners' phraseological competence. To our knowledge, the category of moderately associated bigrams has long been ignored in previous research. Hence, the present study introduces the importance of this category to research on phraseology. Similarly, while previous research which examined learners' bigram knowledge, as reviewed in this study, was conducted in contexts where English is either a 'real' second language (abundantly available in the context outside home) or a foreign language which is introduced to learners earlier in their schooling years, the present study is conducted in a context where English as a foreign language is taught only late at the last four years of the secondary school. This shall provide a different view of how EFL learners develop their phraseological knowledge in contexts where English is not available to them in the outside context.

## 2. Literature review

### 2.1. Previous studies on bigram knowledge

In this section, the studies which mainly focused on the production of n-grams in learners' corpus are reviewed. Durrant and Schmitt (2009) investigated the use of two-word collocations in essays written by post-graduate students on pre-sessional EAP courses at a British university and first year undergraduate students on in-sessional EAP courses at an English-medium university in Turkey. The study concluded that non-native writers developed better knowledge of frequent collocations (as attested by t-scores) earlier while they showed less use of strongly associated rare items (as attested by MI scores). A similar finding was reported in Durrant (2014), who concluded that native speakers show some ability to acquire infrequent but strongly associated collocations earlier compared to non-native speakers who do not show sensitivity to low-frequency but strongly associated collocations. Similarly, Granger (2018) noted that intermediate learners tend to overuse high frequency collocations such as *hard work* and to underuse lower-frequency, but strongly associated, collocations such as *immortal souls*.

O'Donnell et al. (2013) compared the use of collocations by L2 learners to that of expert writers. The study concluded that L2 learners produced more frequency-defined formulas than their native speaker peers. Surprisingly, the study showed that all non-native groups outperformed native speakers in their use of MI defined formulas. A similar finding was outlined by Kim et al. (2018), who

investigated the use of a variety of lexical features in L2 learners' writing. The results indicated that L2 learners began to produce a greater proportion of particularly frequently used n-grams overtime.

Other studies reported that the production of collocations is related to the amount of input which students receive. Li and Schmitt (2010) investigated the production of adjective + noun co-occurrences by Chinese learners of English. The study showed that the learners as a group did not change much over the academic year. Similarly, Laufer and Waldman (2011) concluded that there was no relationship between the proficiency level of learners and the number of erroneous collocations they produced. However, Groom (2009) found a high correlation between the number of statistical co-occurrences and the amount of time spent in an English-speaking country. Also, Paquot (2017) examined adjectival collocations produced in the writings of intermediate and advanced French EFL students of linguistics. The results indicated that more advanced writers produce collocations with significantly higher MI mean scores for all three types of adjectival combinations.

The results of the reported studies point to one main direction in the use of collocations. Learners are reported to mainly produce high-frequency collocations as highlighted by the t-score earlier compared to their production of less frequent but strongly associated collocations, which are highlighted by the MI score. In terms of quantity, the studies which examined the proportion of bigrams in learners' texts indicated that learners make more use of high-frequency bigrams compared to lower-frequency but strongly associated ones. Similarly, the results of these studies show that learners' proficiency level may not affect much the acquisition of rare collocations. Most studies indicated that learners tend to demonstrate better knowledge of high-frequency collocations earlier compared to lower-frequency collocations.

## 2.2. Measures of bigram strength of association

To measure the strength of attraction between to linguistic items, including bigrams and other types of collocations, researchers rely on measures of strength of association which derived from contingency tables (e.g., Evert, 2009; Gries et al., 2005; Gries & Stefanowitsch, 2003, 2004; O'Donnell et al., 2013). The use of strength of association measures aims at rejecting the null-hypothesis of independence or complete absence of association between the words (Evert, 2009). Two main measures of strength of association are widely used: MI statistic and the t-score statistic. Durrant (2014, p. 456) maintains that "MI is a measure of the extent to which the probability of meeting one word increases once we encounter the other. T-score, on the other hand, is a hypothesis testing technique,

which evaluates how much evidence there is that a particular combination occurs more frequently than we would expect by chance alone, given the frequencies of its component parts."

The MI score, which is interpreted as the number of bits of "shared information" between two words (Evert, 2009) is calculated as  $MI = \log_2 \frac{O}{E}$ . MI highlights collocations which are composed of rare words such as *exultant triumph*. Chen and Baker (2014) maintain that the MI score highlights n-grams made up of words that are rarely found independently of each other. Evert (2009) reports that an MI value of 0 bits corresponds to a word pair that co-occurs just as often as expected by chance; 1 bit means twice as often as expected by chance, 2 bits means 4 times as often, 10 bits about 1000 times as often, etc. while a negative MI value shows that a word pair co-occurs less often than expected by chance. Higher MI scores indicate that there is stronger attraction between the words forming the collocation. Qian (2019, p.3) reports that "the MI-value measures how strongly words are attracted to each other. A high MI is indicative of a collocation that is idiomatic or of high quality." A negative MI score shows that the two words are not attracted to each other frequently enough in the corpus. In their study, Bestgen and Granger (2014) concluded that negative MI values, assigned to bigrams such as *everything are*, correspond to bigrams that co-occur in the reference corpus less frequently than chance would predict. The top-scoring bigrams identified on the basis of MI scores, however, all contain much less frequent words.

A second measure of strength of association is the t-score. Bestgen and Granger (2014) reported that the top scoring bigrams identified on the basis of the t-score are composed of frequent grammatical words (pronouns, prepositions, auxiliaries, determiners) and high frequency lexical verbs such as *have, think, want, get, and say*. The t-score highlights collocations which are relatively frequent in the corpus and it is calculated as:  $t\text{-score} = \frac{O-E}{\sqrt{O}}$ . The t-score measures the degree of confidence about the existence of an association. Compared to MI statistic, the t-score highlights the number of times a bigram has been observed. Hence, the statistic underscores frequently occurring bigrams, mostly made up of frequent words.

### 3. Objective of the study and research questions

#### 3.1. Objectives of the study

The current study has two main objectives. First, it aims at highlighting how EFL learners develop their knowledge of bigrams in a context where English is introduced to learners only at the end of the secondary school. While addressing this objective, the present study adopts a continuum perspective through which it

examines how learners of different proficiency levels enhance their knowledge both of strongly associated and the long-ignored moderately associated bigrams. In the meantime, by examining the produced wrong word combinations, the current study aims at examining if there are any trends in how learners improve their knowledge of how to form accurate word combinations. As a second objective, the current study highlights the pedagogical implications which can be concluded from how learners develop their phraseological implications in EFL contexts.

### 3.2. Research questions

To address the aforementioned objectives, the present study is guided by the following three questions:

1. Do EFL university learners of adjacent proficiency groups differ in their productive knowledge of strongly associated bigrams?
2. Do EFL university learners of adjacent proficiency groups differ in their productive knowledge of moderately associated bigrams?
3. Does the use of inappropriate word combinations reflect any developmental tendency in learners' knowledge of bigrams?

## 4. Method

### 4.1. Participants

This study adopted a cross-sectional design. The design is, however, quasi-longitudinal in the sense that the three groups are adjacent, which shall reflect a perspective which is similar to how longitudinal research is conducted. A corpus of 529 written essays was collected from three main groups of undergraduate EFL learners. The participants of the three groups belong to different classes and they were taught by different instructors. The participants are from semester 1 (S1), semester 3 (S3) and semester 5 (S5) students who are studying English as a Foreign Language (EFL) in a Moroccan university. At the time of administering the writing test, semester 1 students just started their 1st year at university, S3 their 2nd year, and S5 their 3rd year. The phraseological knowledge of semester 1 learners, therefore, reflects the knowledge they had built over the last four years of being exposed to English in the secondary school. Their inclusion in the study aims at monitoring the development of the phraseological knowledge for the other two levels (S3 and S5). This study included 167 (31.65%) participants from semester 1, 188 (35.53%) from semester 3 and 174 (32.89%) from semester 5. Similarly, as Table 1 shows, there is a balanced representation both of females and males within each proficiency level.

Table 1 The numbers and percentages of participants in the study

Proficiency Level		Frequency	Percent	Valid percent
S1	Male	79	47.3	47.3
	Female	88	52.7	52.7
	Total	167	100.0	100.0
S3	Male	76	40.4	40.4
	Female	96	51.1	51.1
	NI (not indicated)	16	8.5	8.5
	Total	188	100.0	100.0
S5	Male	79	45.4	45.4
	Female	92	52.9	52.9
	NI (not indicated)	3	1.7	1.7
	Total	174	100.0	100.0

## 4.2. Data collection

The present study used as its data a corpus of 529 essays which were written by undergraduate EFL students. Collecting the data was done at the beginning of the school year. The participants were asked to write an expository essay of around 300 words about a unified topic. Writing the essays was done in students' officially scheduled classes to make sure that the students could attend the class and no re-scheduling was required. Also, the students wrote their essays in regular classes which required some kind of dissertation writing such as their composition or applied linguistics classes. Before taking the test, the participants were given instructions about the objective of the test and the task was clearly explained. We also made sure that the students were invigilated by their assigned teachers. This helped a lot in controlling for any issues which might have affected the quality of the test such as the use of mobile phones or copying from peers. All the groups were given one hour to complete their writing which was a response to the following prompt:

*Write an essay about the reasons which push young people to pursue their university studies besides the aim of getting a job after their graduation.*

## 4.3. Data analysis

To prepare the texts for the bigram analysis, a number of procedures were carried out. First, we corrected all the spelling and punctuation mistakes. Moreover, we deleted proper nouns from the texts, mainly the repetition of the words *Morocco* and *Moroccan*. Also, contracted forms were changed into their full forms. Because we aimed at examining learners' errors in forming bigrams, no attempt was made to correct the words which are wrongly used or even the grammar

mistakes. Finally, since this study relies on the Corpus of Contemporary American English (COCA) as its reference corpus, we normalized all the words to the American English spelling. The students' texts were then saved as a .txt format in order to allow for their analysis by the COCA Parser tool (Wolk et al., 2017). All the analyses which were used in the present study relied on computing MI and t-score statistics for the bigrams in each individual text.

In this study three categories of bigrams were analyzed. First, we analyzed the category of strongly associated bigrams. Following previous studies (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009), we opted for an MI score which equals or is more than 3 and a t-score which equals or is more than 2 for our selection of strongly associated bigrams. Second, the category of moderately associated bigrams was also examined. This category contains bigrams with an MI score which ranges from 0 to 2.99 and bigrams with a t-score which ranges from 0 to 1.99. Finally, to have a complete picture of the development of learners' bigram knowledge, the category of inappropriate or wrong word combinations, those which are associated with negative MI or t-scores, was also examined.

To obtain the lists of bigrams of each category, we relied on the automatically generated lists and their strength of association scores from each individual learner's text. The lists were sorted out depending on their strength of association scores which were used to compute the means and standard deviations. In the current study, the bigrams were examined as *types* rather than *tokens*. That is to say, only one single appearance of each bigram was used in the analyses. For the third research question, the erroneous word combinations were studied and classified based on the possible source of the error in the combination.

## 5. Results

### 5.1. Analyzing strongly associated bigrams

This section analyzes the category of strongly associated bigrams and provides the results which are related to the first research question. Table 2 provides the descriptive statistics for the performance of the three groups in their production of strongly associated rare and frequent bigrams.

For the category of rare bigrams (the MI statistic), we observed that the mean MI score increases considerably from lower to higher proficiency levels. S1 learners had a mean of 4.37 and S3 learners had a mean of 4.41. For S5, the mean is considerably higher (mean = 4.47). The mean difference was statistically significant ( $F(2, 526) = 6.30, p = .002$ ). Using Tukey HSD multiple comparisons test, the results showed that only the difference in means between S1 and S5 is statistically significant ( $p = .01$ ) while the mean difference between S1 and S3 ( $p = .36$ ) and



also between S3 and S5 ( $p = 0.65$ ) was not statistically significant. This suggests that lower proficiency learners relied more on the production of less strongly associated bigrams. The results also suggest that higher proficiency learners tend to produce more strongly associated combinations. These combinations mostly involve content words such as *minimal wage*, *intrinsically motivated*, *grammatically correct*, *motivating factor*, *gravitate towards*, *dominant role*, *daily life*, etc.

Table 2 Learners' performance in strongly associated rare (MI) and frequent (t-score) bigrams (in types)

Strongly associated bigrams: Descriptive statistics							
Proficiency level		N	Minimum	Maximum	Sum	<i>M</i>	<i>SD</i>
S1	Strongly associated bigrams (MI $\geq 3$ )	167	3.19	4.88	730.87	4.37	0.30
	Strongly associated bigrams (T-score $\geq 2$ )	167	0,000	180.51	20724.16	124.09	21.18
S3	Strongly associated bigrams (MI $\geq 3$ )	188	3.87	5.05	830.03	4.41	0.19
	Strongly associated bigrams (T-score $\geq 2$ )	188	77.57	171.08	23043.14	122.56	18.85
S5	Strongly associated bigrams (MI $\geq 3$ )	174	3.93	7.47	779.23	4.47	0.29
	Strongly associated bigrams (T-score $\geq 2$ )	174	74.72	177.07	20900.71	120.11	16.19

For frequent bigrams as highlighted by the t-score, we observed a completely different trend. First, the means were higher compared to those of the MI statistic and second a reversed effect of proficiency was observed. As for the t-score means, S1 had a mean ( $M = 124.09$ ) which is higher compared to S3 ( $M = 122.56$ ) and also S5 ( $M = 120.11$ ). Similarly, S3 had a mean which is higher than that of S5. These mean differences were not statistically significant ( $F(2, 526) = 1.94, p = .14$ ). Strongly associated bigrams which are highlighted by the t-score usually involve function words as in the examples: *it is*, *to be*, *they are*, *they have*, *want to*, *to have*, etc.

## 5.2. Analyzing moderately associated bigrams

The second research question examined the differences between the three groups in the category of moderately associated bigrams (those with MI = [0-2.99] and t-score = [0-1.99]). Examining bigrams with moderate strength of association might give us an indication of learners' attempt to expand their collocational knowledge and to move away from the production of only salient, strongly associated collocations to the production of more flexible combinations.

Table 3 Learners' performance in moderately associated rare (MI) and frequent (t-score) bigrams (in types)

Proficiency level		N	Minimum	Maximum	Sum	M	SD
S1	moderately associated bigrams (MI = [0-2.99])	167	0.72	1.86	251.23	1.50	0.14
	moderately associated bigrams (t-score = [0-1.99])	167	.000	1.98	154.54	0.93	0.55
S3	moderately associated bigrams (MI = [0-2.99])	188	1.03	1.92	291.55	1.55	0.12
	moderately associated bigrams (t-score = [0-1.99])	188	.000	1.99	169.37	0.90	0.51
S5	moderately associated bigrams (MI = [0-2.99])	174	1.14	1.84	271.67	1.56	0.12
	moderately associated bigrams (t-score = [0-1.99])	174	.000	1.85	148.95	0.86	0.45

The descriptive statistics indicate that there were differences between the three groups in moderately associated bigrams. For the bigrams which are highlighted by the MI statistic, S1 had a mean score of 1.50, S3 had a slightly higher mean of 1.55. The mean for S5 ( $M = 1.56$ ) is not very different from that of S3. These MI means suggest that learners' productive knowledge of moderately associated bigrams tends to move *increasingly* towards the production of more strongly associated rare bigrams. The increasing MI scores from S1 to S5 indicates that similar to strongly associated bigrams, learners tend to increase their overall MI mean scores in the category of moderately associated bigrams. Using an analysis of variance test, we checked the differences in group means, we concluded that the mean differences between the three groups were statistically significant for the MI ( $F(2, 526) = 9.82, p < .05$ ). To check where the differences lie between the three groups in moderately associated rare bigrams, we used Tukey HSD post-hoc test. The results show that the difference was statistically significant between S1 and S3 ( $p = .002$ ) and also between S1 and S5 ( $p < .05$ ). The observed slight difference in the mean MI statistic between S3 and S5 learners was not statistically significant ( $p = .708$ ). This suggests that within the category of moderately associated bigrams, learners of different proficiency groups show some improvement towards using more strongly associated bigrams by increasing their mean MI scores. However, the difference is significant only when we compare learners whose difference in instruction range is two years (S1 to S5). Examples of moderately associated bigrams which are highlighted by the MI score include: *something that, educated enough, their language, people choose, the way, etc.*

For moderately associated bigrams which are highlighted by the t-score, we observed again that there was some reversed effect of proficiency. S5 had the lowest t-score mean ( $M = 0.85$ ), followed by S3 with a slightly higher mean

( $M = 0.90$ ) and then S1 learners with the highest mean ( $M = 0.92$ ). The trend is very similar to that which was observed for the three groups in the category of strongly associated bigrams with the t-score. Similar to what was observed for strongly associated bigrams with a t-score, this difference in means between the three groups was not statistically significant ( $F(2,526) = 0.828, p > .05$ ). These results suggest that the three groups were significantly different in their productive knowledge of bigrams which are composed of rare, mostly content, words while their knowledge of bigrams which are composed of words which are of high-frequency was very similar. Instances of these bigrams are: *cannot be, parents asked, fight illiteracy, their exam, etc.*

### 5.3. Analyzing negatively scored combinations

This section deals with the third research question. It examined the list of bigrams which appeared in learner data with negative strength of association scores and we checked if there was an effect of grade level on the performance of learners on their knowledge of how to form more accurate bigrams.

Table 4 Learners' performance in negatively scored bigrams (in types)

Erroneous word combinations: Descriptive statistics							
Proficiency level		N	Minimum	Maximum	Sum	M	SD
S1	word combinations (negative MI score)	163	-3.23	1.96	-284.48	-1.74	0.52
	word combinations (negative t-score)	167	-497.96	-21.12	-19692.66	-117.92	76.22
S3	word combinations (negative MI score)	188	-2.62	1.51	-292.37	-1.55	0.38
	word combinations (negative t-score)	188	-2269.85	-12.49	-19816.42	-105.40	171.08
S5	word combinations (negative MI score)	174	-2.1	0.61	-245.11	-1.40	0.30
	word combinations (negative t-score)	174	-206.09	138.07	-14020.46	-80.57	42.79

Table 4 shows that there were some differences in the means between the three proficiency groups. For S1, the negative MI mean score was -1.74. S3 students had a lower negative mean score of -1.55 and S5 students had the lowest score of the negative combinations compared to the two groups with a negative MI mean score of -1.40. An analysis of variance showed that this difference was statistically significant ( $F(2, 522) = 28.57, p < .05$ ).

A similar trend was observed for bigrams which are associated with negative t-scores. Comparing the performance of the three groups showed that the

negative t-score means consistently increased towards the positive side of the scoring scale. For S1, the mean t-score was -117.92, followed by S3 with a lower t-score mean of -105.40 and then S5 with a mean t-score of -80.57. The difference in means was also statistically significant ( $F(2, 526) = 4.82, p < .05$ ). Using Tukey post hoc test, the results showed that for negatively scored combinations with an MI score, the differences between the three groups were statistically significant ( $p < .05$ ) while for negatively scored combinations with a t-score, only S1 and S5 differed significantly ( $p < .05$ ).

The results of the analysis of negatively scored bigrams both for the MI and t-score statistic indicate that each higher proficiency group showed some improvement in their knowledge of bigrams as they increased their mean scores towards the positive MI and t-score statistic compared to their immediate lower group. Examples of negatively scored bigrams are: *reason according, to went, is have, be they, of it, special that, it we, gifts they, etc.*

## 6. Discussion

In order to examine the development of EFL learners' phraseological knowledge, this study looked at how learners develop their productive bigram knowledge as a continuum wherein learners learn how to better avoid erroneous word combinations while enhancing their repertoire towards the production of more accurate and native-like combinations. The first research question examined the differences between the three proficiency groups in the category of strongly associated bigrams. For the MI score which highlights rare bigrams, the results show that the three groups differ significantly and that lower proficiency groups have lower MI means compared to the adjacent higher proficiency group. Because our results and those of previous research (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Qian, 2019) showed that bigrams with higher MI scores are mainly composed of relatively rare content words, we may conclude that higher proficiency learners (S5) have better knowledge of rare word collocations, as their significantly higher mean MI score showed. Similarly, the mid-proficiency group (S3) scored better than S1 and lower than S5 in their mean MI score, which also indicates that better language proficiency is associated with better knowledge of strongly associated rare word collocations.

It seems from our results that as learners gain better knowledge of language, they tend to approach native-like competence in the use of rare bigrams. In this respect, Durrant and Schmitt (2009) investigated the use of premodifier-noun combinations and found that native speakers demonstrated a significant use of low-frequency, strongly associated collocations. Also, Granger and Bestgen (2014) investigated non-native speakers' use of three different word pairs.

Their study concluded that compared to advanced learners, intermediate ones make use of less strongly associated combinations which are highlighted by the MI score. Similarly, our results are similar to those of Paquot (2017), who concluded that more advanced writers produce collocations with significantly higher mean MI scores. The results of these studies indicate that more advanced learners (both native and advanced non-native) make much more use of lower-frequency but strongly associated collocations. Because our advanced learners showed better use of the bigrams which are highlighted by MI scores, we may conclude that their knowledge of bigrams which are composed of less frequent but strongly associated word combinations is approaching native-like use of low-frequency combinations.

This knowledge which is gained as learners move towards the end of their undergraduate studies (S5) is not paralleled by a significant difference in the use of frequent bigrams, as highlighted by the t-score. It was observed that the three proficiency groups demonstrate very little difference in their performance in the mean t-score statistic. This result probably indicates that over their first two years at university, learners of different proficiency groups exhibit very similar knowledge in the category of frequent bigrams, most of which are function words. This similar performance between lower (S1) and higher proficiency learners (S3 and S5) might suggest that using high-frequency bigrams develops at a much higher rate compared to rare bigrams. Because even semester 1 learners have the highest t-score mean, we may tend to say that learners have already attained an advanced level knowledge of strongly associated frequent bigrams, and this association tends to move towards more flexible combinations by the end of the second year at university (semester 5), which explains the lower (though not significant) t-score mean for S5 students. Similar to our result, previous studies (e.g., Bestgen, 2016a, 2016b; Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014) showed that compared to the mean MI score, the average t-score is less successful in distinguishing between advanced and beginner learners in their bigram knowledge. A similar conclusion is highlighted by Qian (2019) who concluded that the average t-score for the five learners who were investigated longitudinally did not change much, compared to the MI values which underwent noticeable changes.

The decreasing t-score mean as learners move from lower (S1) to higher proficiency levels (S3 then S5) indicates that when they first join university from secondary school, learners use more strongly associated bigrams which are composed of frequent words. This is so, probably because they mainly rely on the use of fixed combinations (e.g., *have to*, *should have*, *will do*, *have been*, *going to*, etc.) at the beginning of their EFL instruction, and they gradually start improving their collocational knowledge by experimenting with a wider range of less fixed combinations by the end of semester 5. Compared to learners' knowledge

of strongly associated rare word bigrams, knowledge of bigrams which are composed of frequent (mostly function) words seems to develop faster. This may not be the result of only the abundant form-focused instruction which learners receive during their early years of studying English, but it might also be a direct result of the close and limited nature of the combinations which are allowed with function words, compared to the wide and maybe unlimited range of collocational possibilities of content words. The high-frequency nature of function words also makes them ready to be encountered so frequently earlier compared to the bigrams which are highlighted by the MI score.

The second research question examined the widely ignored moderately associated bigrams. Our results, however, indicate that there is a developmental trend which is similar to that of strongly associated bigrams. For moderately associated bigrams which are highlighted by MI score, the results suggest that after spending the first year at university (S3), learners show a significant difference from university freshmen (S1). Like what is observed for strong bigrams, the mean MI of moderate bigrams is also *increasing* which indicates that learners are moving towards the production of *more strongly* associated bigrams, as would be expected from the results of previous research (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger, 2018; Granger & Bestgen, 2014). However, the non-statistically significant difference between S3 and S5 learners may indicate that after the first year at university learners slow down their progress towards the use of more strongly associated bigrams. The increasing mean MI scores of moderate bigrams indicates that developing knowledge of rare word bigrams seems to adhere to a continuum pattern through which learners move from the use of moderately associated combinations before they manage to use strongly associated content word combinations which would amount to real collocations (e.g., *dominant role, daily life, young people, get married*, etc.).

Similar to our findings about strongly associated frequent bigrams, the category of moderately associated bigrams with a t-score did not show any significant difference between the learners of the three groups. The *decreasing* mean t-score which was observed for strongly associated bigrams continues with the same trend for moderately associated bigrams. Because the three groups did not show any statistically significant difference in the mean t-score both when we consider strongly associated or moderately associated bigrams, we may say that their knowledge of this type of bigrams is very similar. The *decreasing* means as we compare S1 to S3 and then S3 to S5 in both categories of bigrams suggests that learners have started to move away from strongly associated bigrams to the production of less strongly associated ones. In comparison to the production of learners in the category of rare bigrams, our results may suggest that for our three groups, knowledge of high-frequency bigrams (as highlighted

by the t-score) might start from the use of fixed combinations before learners move to the use of a variety of more flexible combinations. Previous studies reported a descending curve as learners enhance their knowledge of bigrams (Bestgen & Granger, 2014; Bestgen, 2016a; Durrant & Schmitt, 2009; Qian, 2019; Siyanova-Chanturia & Schmitt, 2008). It is worth mentioning that most bigrams which are highlighted by the t-score include function words. Hence, with more exposure to grammar, learners might be extending their knowledge of how to combine function words with other words. This is supported by the results of Durrant and Schmitt (2009) who concluded that non-native learners were quick to pick up highly frequent collocations (as attested by t-scores). For frequent word bigrams, the continuum pattern seems to be reversed. Because this kind of bigrams is mostly composed of function words, it appears that learners use these combinations as fixed units in the way they learnt them (*will not, is not, have to, going to, of the, have been, etc.*) before they move towards the use of less fixed units as they experiment with more uses of the function words.

Our analysis of inappropriate uses of bigrams (those which are associated with negative MI and t-scores) indicates that both when we look at bigrams which are highlighted by negative MI scores and those which are highlighted by negative t-scores, learners of the three groups improve their performance significantly. The statistical results show that negative MI score t-score means increased towards the positive side of the scores as we compare S1 to S3 and then S3 to S5. This suggests that each year spent studying EFL provides learners with better phraseological knowledge. Therefore, there seems to be a quasi-longitudinal progress towards better collocational knowledge as learners receive better language input.

Unlike the reversed effect of proficiency which we observed for the t-score means in the categories of strongly and moderately associated bigrams, for inappropriate word combinations, we can see that learners' means become closer to the positive side of the score scale as they gain better proficiency, with S1 having the lowest negative mean and S5 the highest. Although previous studies (e.g., Bestgen, 2014, 2017, 2019; Durrant & Schmitt, 2009; Granger, 2016; Granger & Bestgen, 2014; Qian, 2019) did not fully exploit the category of inaccurate bigrams, the present study showed that the produced bigrams with negative scores might be very useful in understanding the trajectory of learners' collocational development. In fact, the way bigram knowledge is proceeding may indicate that it moves along a continuum both for rare and frequent combinations. For rare combinations, with better knowledge of what makes an acceptable combination, learners use more moderately associated bigrams before they increase their scores towards the production of more strongly associated combinations. For frequent bigrams (i.e., the t-score), the continuum is reversed in the sense that learners first develop better knowledge of strongly associated

combinations first, probably because the form-focused instruction provides them with more opportunities to notice common function word combinations together. Being exposed to more input enhances learners' experience of how function words can be combined differently. Paquot and Granger (2012) mentioned that more advanced learners produce more deviant collocations than lower proficiency learners as an attempt to try a wider range of lexical phrases.

Our analysis of the produced erroneous combinations led us to the conclusion that the same errors are produced in the combinations irrespective of learners' proficiency level. The errors can be classified into the following seven categories:

1. Errors related to the use of the wrong preposition: Examples of this type of collocational errors are: *able in, at learning, basic of, capable to, depend of, ambitious about, aware about, forced for, focus in, interest into, involved about*, etc.
2. Errors related to the wrong use of verbs/modal verbs: Examples are: *are focus, are understanding, are vary, art are, should making, is are, is develop, considers as, behind go, can found, can having, choose go*, etc.
3. Errors related to modifiers and relative pronouns: Examples of erroneous combinations which involve modifiers or relative pronouns are: *factor who, reason who, mind who, teachers which, some student, some situation, some school, some subject, many information, many research*, etc.
4. Errors related the use of the third person singular: This type of erroneous combination is observed in the data of the learners of the three groups. Examples of this kind include: *education help, education offer, earth have, everybody have, everybody say, everyone put, everyone want, he have, he want, it have*, etc.
5. Errors related to the wrong use of articles: This type of errors is observed among the learners of the three proficiency groups. Examples of these errors include: *a best, a dreams, a first, a goals, a god, a knowledge, a health, a money, the communication, the good, the happiness, the humans, the illiteracy, the solve*, etc.
6. Errors related to adjective-noun word order: Examples of the wrong uses of the adjective-noun combinations are: *people capable, people productive, person successful, research scientific*, etc. This type of errors is most likely due to the interference of learners' L1 which is either Berber or Moroccan Arabic since in both languages the noun precedes its modifier in the combination.
7. Errors related to unusual/non-rule governed combinations: In learner data, a large number of errors results from combining two content words in a way which is not appropriate to native speaker usage. Examples of these combinations include: *be respect, be of, big mark, big value*,



*course learning, cover history, country talk, else bad, education study, greatest goals, everyday something, for job, goals they, etc.*

## 7. Conclusions

This study examined the productive bigram knowledge for three groups of EFL university undergraduates. The results of this study showed that learners' phraseological knowledge concerning bigrams does not develop at the same pace. We observed that the bigrams which are highlighted by the MI statistic reflect most significant differences between learners. This indicates that differences in collocational knowledge might exist in the category of rare word combinations which are mostly composed of content words. Learners' performance in the three categories of bigrams suggests that learners' phraseological knowledge is not only composed of strongly associated word combinations as most studies suggest. Exploring moderately associated and also inappropriate bigrams shows that there might be a continuum in the development of phraseological knowledge.

While picking up collocations seems to be a natural phenomenon, formal instruction can accelerate its pace. The results of the present study suggest that collocations which involve function words are mostly combined by grammatical relations. Consequently, learning this type of collocations can be sharpened by formal grammar lessons which mainly focus on relations between words such as *article + noun*, *adverb + verb*, *quantifiers + noun* or *preposition + noun*. While learning frequent bigrams might take place just by exposing learners to graded and rich input, we believe that formal grammar lessons which focus on how words are combined in language might accelerate this process.

Our analysis of errors which are produced by learners showed that the majority of errors involve combining a function and a content word. The analysis of learners' errors in forming combinations shows that most errors are of a grammatical nature. We noticed that learners at different proficiency levels produce very similar errors in combining words. We suggest that language teaching should also focus on the development of learners' phraseological competence through formal instruction. This would allow learners to notice how certain combinations which are not usually governed by rules such as *able to*, *depend of* are used by native speakers. This type of word combinations can also be learned through massive exposure to language input in content-based language classes such as extensive reading, literature, etc. Follow-up consciousness-raising classes are likely to sharpen learners' awareness of how native speakers use collocations.

Usage-based linguists who view language from a phraseological perspective (e.g., Gries, 2008a, 2008b; Hoey, 1991, 2004; Pawley & Syder, 1983; Sinclair, 1991, 1996, 2004) believe that language is constructed from an accumulation of

an extremely large number of possible word combinations. Therefore, we believe that teaching language combinations can be more effective if learners are exposed to combinations as fixed units. This can be done by dealing with the combinations in a thematic way. That is to say, breaking their teaching into modules or units which deal with different content topics (themes), and within each module, learners are exposed to a set of theme-related collocations.

The results of this study are nevertheless comprehensive. We stress the need of further studies about EFL learners in other contexts. Future studies should go beyond the first two years at university (semester 5) in order to test the 'continuum-fashion' claim of bigram development and to check whether by the end of the third year (semester 6) learners start to rely on rare bigrams which even more strongly associated and frequent bigrams which are even more loose. This shall further provide evidence for the continuum perspective which is adopted in this study. It would also be interesting to check if learner's mother tongue has any effect on learners' production of bigrams, mainly the type of errors which are produced by learners.

## References

- Bestgen, Y. (2016a). Using collocational features to improve automated scoring of EFL texts. In *Proceedings of the 12th Workshop on Multiword Expressions*, 84-90. <https://aclanthology.org/W16-1813>
- Bestgen, Y. (2016b). Évaluation automatique de textes: Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. *Traitement automatique des langues*, 57(3), 91-115.
- Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69, 65-78. <https://doi.org/10.1016/j.system.2017.08.004>
- Bestgen, Y. (2019). Évaluation de textes en anglais langue étrangère et séries phraséologiques : Comparaison de deux procédures automatiques librement accessibles. *Revue Française de Linguistique Appliquée*, 1(1), 81-94. <https://doi.org/10.3917/rfla.241.0081>
- Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28-41.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgrade & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181-189). Radopi.
- Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30-49.
- Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching (IRAL)*, 47, 157-177.
- Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, 19(4), 443-477. <https://doi.org/10.1075/ijcl.19.4.01dur>
- Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook*. (Vol. 2, pp. 1212-1248). De Gruyter Mouton. <https://doi.org/10.1515/9783110213881.2.1212>
- Granger, S. (2018). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228-247). Routledge.
- Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 52(3), 229-252.

- Gries, S. Th. (2008a). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13, 403-437.
- Gries, S. Th., Hampe, B., & Schonefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16, 635-676.
- Gries, S. Th., & Stefanowitsch, A. (2004). Extending collocation analysis: A corpus-based perspective on 'alternations.' *International Journal of Corpus Linguistics* 9(1), 97-129.
- Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21-33). Palgrave Macmillan.
- Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.
- Hoey, M. (2004). Lexical priming and the properties of text. In A. Partington, J. Morley, & L. Haarman (Eds.), *Corpora and discourse* (pp. 385-412). Peter Lang.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4-21.
- Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal*, 102(1), 120-141. <http://dx.doi.org/10.1111/modl.12447>
- Laufer, B. & Waldman, T. (2011). Verb-noun collocation in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647-672. <https://doi.org/10.1111/j.1467-9922.2010.00621.x>
- Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 2-46). Continuum.
- O'Donnell, M. B., 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.
- Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121-145. <https://doi.org/10.1177/0267658317694221>
- Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.
- Qian, Y. (2019). Dynamism of collocation in L2 English writing: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 0(0). <https://doi.org/10.1515/iral-2019-0012>
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1), 75-106.
- Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1-24). John Benjamins.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- Siyanova-Chanturia, A. & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review – Revue Canadienne Des Langues Vivantes*, 64. 429-458. <https://doi:10.3138/cmlr.64.3.429>
- Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, 16(2), 180-205.
- Wolk, K., Wolk, A., & Marasek, K. (2017). Unsupervised tool for quantification of progress in L2 English phraseological. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, 383-388.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics*, 21(4), 463-489.