Automated Extraction of Collocations from Intermediate English Textbooks: A Text Mining Study

Zafar Ullah[1]
Arshad Mahmood[2]
Muhammad Uzair[3]

Abstract

Many intermediate level students write wrong collocations because learners do not find standard collocations in their textbooks; hence, they are unable to extract and learn standard collocations which play their roles in fluency and language accuracy. Another problem is that learners, teachers and book writers do not have programming skills to extract collocations. This study aims to extract standard collocations from the generated corpus of intermediate English textbooks taught in Punjab, Pakistan. With the application of Knowledge Discovery Theory and Phrases tool, this study prepares a corpus of 82,487 words. Then meaningful standard collocations have been selected for educational purpose. The current research extracts 166 standard collocational patterns and 297 standard collocational examples belonging to 18 grammatical categories. It becomes a self-made mini collocational dictionary, and this study empowers language learners to generate such mini collocational dictionaries of ESL textbooks with Phrases tool. This study is potentially valuable for intermediate-level ESL students, teachers and textbook writers. Following this study, the learners will decrease collocational errors in their academic discourse and exams. The novelty of this study is evident that first-time, this corpus, and its collocations have been extracted from the research documents of intermediate English textbooks taught in Punjab, Pakistan.

Keywords: collocations, text mining, ESL textbooks, Phrases tool, educational

1. Introduction

Language employs co-occurring standardised linguistic patterns to convey its target semantic shades to its users. "The occurrence of two or more

[1] PhD Scholar NUML, Islamabad

[2] Professor/Director Publications, NUML, Islamabad

[3] Assistant Professor, NUML, Islamabad

words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening" (Sinclair, 1991, p. 170). Collocations consist of those words which frequently co-occur in any text (Firth, 1957) for example, "in terms of", "by my word of honour". So, the repeated collocations fix on the slate of memory until its violation seems absurd, wrong and comic, especially in the academic discourse. Collocations are also popular with other names: lexical phrases, fixed expressions, multiword expressions, bigrams, trigrams, quadgarms, formulaic language, popular co-references, prefabricated patterns and chunks.

This study addresses one main linguistic problem with its other interconnected aspects. Many non-native intermediate students speak and write wrong collocations, for instance, they write or utter "God's features", "search justice" and these linguistic usages annoy examiners and native language speakers because the correct and established phraseologies are "God's attributes" and "seek justice". Its root cause is that students learn unigrams and their semantic shades. However, they do not attempt to learn bigrams/ trigrams/quadgrams because standard collocations are not given separately in the textbooks. Extending this problem, school and college ESL textbook writers do not feel any need to include collocations in the textbooks along with its glossary, while TEFL textbooks are so significant that they are considered the most influential source of learning any foreign language (Littlejohn, 1998, p. 190); hence, there is a dire need to fill this gap.

ESL learners are unaware of the contributory role of collocations as building blocks of language and automated extraction of collocations. ESL teachers consider themselves incapable of extracting correct collocations on their own because they lag behind in technological knowledge of machine learning and Python programming language; consequently, they also neglect teaching collocations in their classrooms. In Pakistan, all textbook writers are senior teachers who are already weak in technological tools and text mining skills. So all stakeholders of ESL learning and teaching face problems about collocations.

The current corpus based study aims to extract collocations from intermediate English textbooks taught to more than 1 million students inPunjab, Pakistan. This technique can empower ESL learners, teachers and textbook writers to extract collocations without coding or programming skills. Thus, learners, ESL teachers, and textbook writers can generate a chain transition to include collocations in textbooks and ESL materials.

Learning collocations is significant in many ways. As a child learns any mother tongue (L1) from unigram to bigram or trigram, and later on, the child utters a full sentence. The current study focuses on the second phase of second language learners and fulfils their needs of learning collocations consisting of bigrams, trigrams and quadgrams. Sinclair (1991) emphasises the point that fluency in the language is not possible with new and creative linguistic expressions; rather it is possible with pre-structured phrases. Therefore, collocations enhance oral fluency because learners can produce linguistic chunks efficiently (Newell & Rosenbloom, 1981). So the extracted collocations empower learners in accelerating fluency in all language skills.

Presently English collocations dictionaries by Oxford, Longman, Macmillan are in the book form, and more than 15 collocations apps from Oxford, Longman etc. are available on PlayStore. Realising its significance and profusion, the current research motivates textbook writers to incorporate collocations in the textbooks. Initially, ESL learners and teachers can extract collocations with the Phrases tool in the Voyant suite. Furthermore, learning standard collocations can save a language learner from the embarrassment of the wrong usage of collocations. It will establish the accurate use of language.

Usually, collocations have lexical, antonymic and syntactic patterns. Six collocations patterns have become common for instance, Noun+Noun, Adjective+Noun, Verb+Adverb, Verb+Noun, Adverb+Adjective and Verb+Prepositional phrase, but this study aims to widen this academic boundary by adding some other standard collocation patterns for the benefit of intermediate students.

After the introduction section, this paper chalks out the plan to review the relevant literature. Then research methodology has been delineated according to Durant (2004)'s seven pointed typology. The next section presents results and discussion which encompasses extraction of standard collocation, its grammatical categories, development of criterion for segregating standard and substandard collocation. Afterwards, the use of collocations for teaching English has been elaborated with the selected examples. In the subsequent section, major findings and recommendations have been mentioned.

2. Literature Review

In this thematic literature review, several research projects emphasized previous findings, contributions and significance of collocations in language learning (Dechert, 1983; Nattinger & DeCarrico, 1992; Taguchi, 2008; Stengers, Boers, House, & Eyckmans, 2011). Several researchers suggested including lexical chunks in teaching material to become an integral part of the schema of learners (Conzett, 2000; Gitsaki, 1999; Harwood, 2002; Lewis, 1997). The following part discusses the usage of tools to extract standard collocations from textbooks corpus, their comparison, and criteria for standard collocations.

Initially, statistical formulae were employed to find collocations from the text. Therefore, some phrases were measured through T score tests, mutual information, and log-likelihood (Dunning, 1993). Again, technophobes and language teachers were unable to extract collocations with this method. To include the collocations from different domains of discourse and language learning, sample collocations were extracted on the basis of six motifs: 1. Theory: grammar, lexis 2. Clinical: linguistic disorders 3. Development: Mother tongue acquisition 4. Learning and teaching 5. Culture: cultural discrepancies and trends 6. Text: corpus or corpora (Wray, 2013). Therefore, collocations not only exhibited frequent language patterns but also expressed the central idea of discourse. Thus, collocations also manifested narratology of the text.

From eight disciplines of education, a 120 million word corpus was structured. Then 100 academic collocations having four words were extracted with WordSmith Tools. Further, these collocations were analysed with the Wilcoxon rank-sum test (a = 0.05) to find their recurrent usage in academic speech. Its major findings were the existence of the most frequent collocations for instance, ‘the case of’, ‘as a result of’, and ‘at the end of’ (Da Silva, Orenha-Ottaiano, & Babini, 2017). Usually, application of coding, t-score test and log-likelihood (Dunning, 1993) caused difficulty for social science researchers; hence, the Phrases tool of the current study functioned with embedded statistical formulae and automated extraction process for the ease of social science learners.

Empirically, collocations were employed to teach Business English in China, and the experimental study of teaching through COCA (Corpus of Contemporary of American English) and Wikipedia corpus was conducted on 23 undergraduate students for 23 days. Pre-test, post-test, questionnaires, and reflective journals were used to collect data. Primarily, this study enhanced collocations awareness through corpora-based teaching (Chen, 2017). It was revealed that the language of any particular domain could be learnt best from its core corpus and text mining tools. On the other hand, Cambridge Business English Corpus could have been more beneficial for research.

Hsu (2008) worked on lexical bundles of ELT textbooks printed from 2003 to 2005, and he found discrepancies in collocations. He emphasized that ESL teachers should realise the impact of ELT textbook-based collocations. The current paper suggested that learners, teachers, and textbook writers should be exposed to collocational extraction methods through Phrases tools. Thus, autonomous and ubiquitous learning would be promoted, and learning would be transformed into a research-oriented activity since textbooks were building blocks of the language. As researchers explore knowledge patterns from their datasets, similarly ESL students should explore collocations and other valuable linguistic features from their digitized books.

Native and non-native learners spoke any language differently because non-natives were not exposed to the real use of language. So, the word list of the Korean curriculum and native speakers' linguistic corpora were compared (Choi, & Chon, 2012). Non-native speakers were lagging behind in language learning because of their distance from real language use. This study necessitated the extraction and use of collocations for non-native language learners. In a later study, Boers, Demecheleer, Coxhead and Webb (2014) tested the effectiveness of phrase-focused exercises commonly found in English textbooks.

Another corpus-based research was conducted in Korea; therefore, a corpus of nine Middle Schools English textbooks was compared with nine other English textbooks published by three different publishers. In this study, the Concordance Program was applied to extract collocations. Their lexical items showed similarity of 3.6 times, but a few common collocations were found. Its results revealed a mismatch and lack of coordination between textbook writers for the same age group learners (Lee, 2015).

Following this research tendency of collocational extraction from textbooks, standard collocations were evaluated with a comparison of Hong Kong primary level English textbooks and the UK-based English textbooks. The former were non-native learners, while the latter were native learners. Corpora of both textbooks were compared; thus, Hong Kong books presented lesser collocations while the UK textbooks used more collocations. Moreover, lexical structures varied in form and use in both corpora (Russell, 2017). Mismatch of their linguistic abilities caused linguistic and cultural shocks for language users and created difficulties for language comprehension.

Cross comparisons of textbooks corpora, their collocations, and genres were conducted to compare their resemblance and variation. Biber, Conrad & Cortes (2004) extracted lexical bundles from university textbooks and conversational language from the classroom based on their frequency. The selected lexical bundles were compared, and it was found that classroom teaching discourse used more stance bundles than conversational language and academic prose of textbooks. It was a comparative study while the current study extracted collocations for language learning and teaching purposes.

The use of technological tools for text mining and corpus building increased as the volume of data increased over time. Cheng, Greaves, Sinclair, & Warren (2008) pointed out incongruities in collocations through Concgram software. It claimed that collocation extraction facilitated the comprehension of collocations (Sinclair, 1987). The first step was to extract collocations, and the second step was to generate new collocations on the existing patterns. The same notion was highlighted in Hermeneutica Theory thus: “They can be extended or built upon. They are generative and they lead to new things” (Rockwell, & Sinclair, 2016, p. 166). To conclude, collocations accelerated the comprehension process and constructed new bigrams and trigrams on the basis of the established collocations.

The current study developed the premise that collocations can be extracted from any document through corpus and text mining tools, then new collocations can be constructed on the pattern of already extracted collocations. Another question was raised about the inclusion of new words in new editions of dictionaries. When a word frequently occurs in media and academic writings, lexicographers include the word in new editions of dictionaries. When a sit-in of PAT (Pakistan Awami Tehreek) in Pakistan used the name of GULLU BUTT frequently in its speeches, Oxford Advanced Learners Dictionary included the word gulluism. By following the same linguistic trend in journalism, commercial book publishers have started to include collocations in new ELT books (Koprowski, 2005).

Contrary to it, school textbooks did not pay full heed to the inclusion of bigrams and trigrams (Wray, 2012). Wang and Good (2007) also notified collocations-based deficiencies in ESL textbooks. If textbooks were faulty, they would produce mismatched collocations. The current study highlighted the same research problem in the introduction section, consequently, strived to address this problem in the results and discussion section.

When collocations were evaluated, 33% figurative collocations and 33% idioms were found. Phonological multiword expressions (MWEs) were chosen on the basis of alliteration and assonance. This research explored that interdisciplinarity in language diverged into MWEs (Siyanova-Chanturia, 2017). This study discriminated between collocations and idioms, while the current study classified them into their grammatical categories.

The use of collocations went beyond language learning, and its application was started in neural machine translation by one-to-one mapping. Multilingual documents from newspapers were evaluated with a data visualization tool named Fips syntactic parser for French and English languages. Afterwards, a log-likelihood ratio statistical test was applied to them (Seretan, Nerima, & Wehrli, 2004).

Koprowski (2005) diagnosed the problem of textbook writers that they desired to incorporate collocations in books but were unaware of collocations' selection criteria. So being in this confusion, they avoided collocations. Here the point is machine-generated collocations had “noise” and less accuracy level. So the current study recommended the inclusion of meaningful and most content words (nouns, verbs, adjectives, adverbs) phrases as standard collocations while substandard collocations were eliminated. Thus, “and the” phrase was declared as a substandard collocation while “seek justice” was presented as a standard collocation.

Another study dealt with the issue of standard criterion for standard collocations; consequently, it suggested six parameters. Then an experiment was conducted by selecting forty terms, and 1000 most occurring collocations were shortlisted from BNC (British National Corpus), and lastly, 2000 standard collocations were finalised for the spoken English syllabus (Shin, & Nation, 2007). Thus, this study devised criteria, and its 2000 collocations were utilized for learning and teaching purposes.

The retention deficiency of collocations was found among university level students. Specific exercises had been introduced for teaching multi word expressions. This study used 20 verb noun collocations during eighth weeks of treatment through certain multiword expression exercises. The study found that repetition and exercises improved the knowledge level of multiword expressions (Ferguson, Siyanova-Chanturia, & Leeming, 2021).

Primarily, the current research introduced a simple and effective way of automated collocations which were extracted from educational textbooks, and its utility could be viewed from three aspects: ubiquitous quality of collocations, their correct use as a means of linguistic proficiency in all language skills and effect of collocations on ESL learning (Nesselhauf, 2003). Language learning evolved from the gradual process of learning from one word to multiword expressions. As their ability increased, the learning process expanded and reached the sentence level. Thus, after unigram learning, learning collocations (bigram/ trigram/ quadgram) became the next phase of language learning.

The current research raises three research queries to drive the current study: i. What are the standard collocation patterns and instances found in intermediate English textbooks in Pakistan? ii. What are the criteria to differentiate between standard and substandard collocations? iii. How do the extracted collocations teach correct English language patterns? Though Phrases tool extracts collocations, yet there is a challenge to differentiate between standard and substandard phrases, and there is also a need to utilise the extracted collocation for learning and teaching English language.

3. Research Methodology

Table 3.1 Durant (2004)'s Seven-Pointed Typology of Research

Features	Application
i. Framework for Research	Knowledge Discovery Theory
ii. Data Generation	Digitization of Intermediate English Textbooks
iii. Research Method Approach and Rationale	Mixed-Methods Approach
iv. Data Quality and Data Mining	DM Rules and Voyant Tools for DM
v. Data Handling and Data Analysis	Text, Visual and Quantitative Data Analysis
vi. Research Management and Application of Research	Ethical Considerations and implications
vii. Research Skills	Data Presentation Architecture

The typology of seven rules have been applied in the following paragraphs:

i. Framework for Research: Throughout the study, data were analysed through the lens of Rakesh Aggrawal's Knowledge Discovery Theory (KDD). "In active data mining paradigm,… rules are discovered, …the history of the statistical parameters associated with the rules is updated… we describe the constructs for defining shapes, and discuss how the shape predicates are used in a query construct to retrieve rules whose histories exhibit the desired trends" (Agrawal, & Psaila, 1995). KDD was also defined as "the extraction of implicit, previously unknown and potentially useful information from data" (Cabena, Hadjinian, Stadler, Verhees, Zanasi 1998, p. 9). Corpus building is a part of the knowledge discovery process, and its ubiquitous presence with just one-word hyperlink was another beneficial aspect that the whole corpus has been embedded in a hyperlink of one word.

ii. Data Generation: Research documents of the current study were intermediate English textbooks (Book I, II, III, and a novel) which consisted of 15 short stories, 20 poems, three one-act plays, ten literary essays, five biographical essays, and one novel named '*GoodBye Mr Chip*s'. They are taught to intermediate-level students as compulsory English textbooks. More than 1 million students read these textbooks every year and appear in the annual exam because intermediate is a gateway exam before selecting any career. In this way, research documents represented all literary genres. First of all, hard file textbooks were transformed into digitized files, and a corpus was prepared. It consists of an 82,487-word corpus which is accessible by CTL+clicking this hyperlink named said.

iii. Research Method Approach and Rationale: This study followed a mixed-methods approach. In the data collection phase, qualitative textual data were collected, however, in the analysis phase, quantitative data were shown as the occurrence of each collocation, the number of categories and qualitative data were shown as standard phrases.

iv. Data Quality and Data Mining: It is an empirical study of the corpus of Intermediate English textbooks to extract standard collocations for learning and teaching purposes. Phrases tool from Voyant suite was applied for the text mining process. Then generated collocations were segregated, and only standard collocations were selected for the current study.

v. Data Handling and Data Analysis: Corpus meant an electronically generated collection of texts to discover its token words, occurrence, collocations etc. So the use of "Collocations as instrumentation for meaning is a scientific fact" (Louw, 2010, p. 79). The key point is that collocations carry semantic shade and linguistic standardised norms for language users. Then each unit of 6 literary genres was uploaded separately on the Phrases tool in Voyant. Afterwards, the Phrases tool mined each textual unit (story, play, essay, poem, novel) and presented collocations. The First 15 most occurring phrases were shortlisted, and their snapshot was taken for this study. Some poems showed only a few phrases. One sample generated by the Phrases tool has been presented in Appendix A. To save space, those 54 snapshots have not been shown in the appendix. Then standard collocations (mostly content words) were kept in table 2 while substandard collocations (mostly function words e.g "and the") were ignored. Consequently, 166 grammatical collocation categories were found and their respective 297 sample phraseologies were written in their respective boxes. Afterwards, grammatical sequences of standard phrases were manually written in table 3. The list of phrases was classified as 18 grammatical categories.

vi. Research Management and Application of Research: This study is applicable for language learners and teachers. It is also useful for textbook writers so that they can compile a mini dictionary for learning the correct language.

vii. Research Skills: No certain computer programming skills are required for this study. Only linguistic skills are required for this study; hence, learners and practitioners from social humanities can do this text mining task for collocation extraction.

As validation parameters are concerned, collocations' accuracy and occurrence remain the same in every analysis. The current study draws a linguistic criterion between standard and substandard collocations because the Oxford collocation dictionary also leaves out substandard collocations. The research method is robust, and minor changes cannot distort its results. Collocation study is a combination of mixed methods, and data are produced in qualitative and quantitative forms. The current research measures and extracts collocation patterns that have pedagogical value; hence, it presents criteria for differentiating between standard and substandard collocations.

4. Results and Discussion

The current study mines textual data to extract collocation patterns without coding skills. For this purpose, the Phrases tool extracts collocation patterns along with their occurrence. The following tables 2, 3 show the results of this collocational study.

4.1 Instances of Collocation

Table 2 shows extracted instances of collocations.

Table 3.2 Collocations in Intermediate Textbooks

Collocation Patterns	Instances
1. Adj+Adj+N	blue suede shoes
2. Adj+N	noble deeds, Christmas time, a few minutes, night mail, best interest, Western Europe, national assembly, walnut cake, pink icing, acting head, a little money, fruit stalls, bright boy, many boys. summer holidays, early childhood, household use, good morning, eating habits, agricultural Commune, death rate, great mosque,
3. Adj+N+N	fifty thousand dollars
4. Adj+N+Prep+N	this sort of thing, sweet land of liberty
5. Adj+N+Adv	a few days later
6. Adj+N+Prep	total number of
7. Adj+Prep	afraid of, 2. full of,
8. Adj+Prep+Art	2. one of the , (one. Adj. Merriam Webster)
9. Adj+Prep+Prn	Most of them
10. Adj+Prep+N+Art+N+Adv	2. loveliest of trees the cherry now
11. Adv	a lot of , of course
12. Adv+Adj	too hot, not mere
13. Adv+Art+N	before the king, when the fact fails, as a result,
14. Adv+Prep	just to
15. Adv+Prep+Prn	ahead of him
16. Adv+Prn+Aux	before I had
17. Adv+Prn+V	as she left
18. Adv+Adj+N	when all god's children
19. Adv+Art+N+Prep+N	as a matter of fact
20. Art+N	the sun, the Greek, a man, a spaceship, the house, a bit, a dollar, the grain, an hour,
21. Art+N+V+Prep	the years went by
22. Art+N+Prep+N	a pair of shoes, a piece of string,
23. Art+N+Prep	a list of, a short of, a piece of, a sense of,
24. Art+N+Prep+Art+N	a quarter of a century
25. Art+N+Prep+Adj+N	the post of court acrobat

26. Art+N+Inf V+Prep	a time to cast away
27. Art+N+Conj+Art+N+Prep+Art+N	a book and a pearl in the oyster
28. Art+N+Prep+N+Conj+Adv+Prep+N	a man of words and not of deeds
29. Art+N+Inf V	a time to keep
30. Art+Adj	a little, a whole,
31. Art+Adj+N+Prep	a new kind of
32. Art+N+Prep+Adj	the sons of former
33. Art+N+Aux+V+Prep+Art+Adj+N	the package was lying by the front door
34. Aux	ought to, 2. may be
35. Aux+Adj+N+Aux	may be some eccentric millionaire is
36. Aux+V	have to look, did come,
37. Aux+V+Prep	have liked to
38. Aux+V+Prn	doesn't intrigue you
39. Aux+Adj+Inf V+N	will be able to join hands
40. Aux+Prn+Adj	aren't you ashamed
41. Conj	as if, 2. as well as, as soon as (Subordinate Conjunction)
42. Conj+Adv	2. and then
43. Conj+Prn	as she
44. Conj+Prn+V+Conj	and you wonder that
45. Conj+V+Prn+V	and let him go
46. Conj+Prn+Mod	if we can
47. Det+N	no time
48. Id	out of order , just as (Merriam Webster)
49. Inf V+Prn+N	to wash your face, to dig her grave,
50. Inf V+Art+N+Prep	to shed the blood of
51. Inf V	2. to stand
52. Inf V+N	to seek justice
53. Inf V+Inf V+Art	to try to burn a
54. Int+Adv	oh yes
55. Int+Adv+Aux+Prn+V	O where are you going
56. Int Prn+Aux+Prn,+N	what is it, son
57. Int Prn+Aux+Adj	who are stupid
58. Mod	used to

59. N	Carbolic acid, commander in chief
60. N+N	Mustafa Kamal, Abdal Rahaman,
61. N+N+N	melon, guava, mandarin
62. N+Apo+N	God’s attributes
63. N+Aux	faith is
64. N+Conj+N	gold and silver, disease and death, communication and transportation,
65. N+Prep	cooperation with, beak with, contribution of,
66. N+Prep+N	degrees of frost, grain of sand, sense of proportion, kinds of food, use of science, birth of Christ, cost of living,
67. N+Prep+Art+N	story of the string, cells of the body, jewel of the world, end of the week, culture of the mould, surface of the sun, clearing in the sky
68. N+Prep+Art+Adj+N	creation of a new world
69. N+V+Prep	applause broke out
70. N+Prep+Art+N+N	land of the pilgrims’ pride,
71. N+V+Prep+Prn	Arthur stared at her
72. N+Prep+Art	letters for the,
73. N+V	Leaves drinking, calculation shows,
74. N+V+Prn+N	friend breathing his last
75. N+V+Ref Prn	God calls himself
76. N+Prn+V	rain I hear
77. Nu+Adj+N	three hundred dollars
78. Prep+Prep	out of
79. Prep+Art+N	down the hall, at the hills, on the daybed, to the village, on the road, in the world, to the ground, for a moment, for a moment, in the presence, for a while, across the street, in a row, into the cold, on a Friday, in the world, in the world, across the Sahara, in the broiler
80. Prep+Art+N+Prep	in the street of, in the hands of, for the benefit of
81. Prep+Art+N+N	in the hills sir
82. Prep+Art+N+Prep+N	in the treatment of disease

83. Prep+Art+Adj+N	toward the deep valley
84. Prep+Adj	of two
85. Prep+Adj+N	in other words, in broken images
86. Prep+Adj+N+Apo	at St Mary’s
87. Prep+Det+N	in such cases
88. Prep+Art+Adj+N	to the next village, for a long time, in a low voice, on the culture plate, for the first time
89. Prep+Art+N+Conj+Prn+Mod	to the end that you may
90. Prep+Art+Adj	at the same, in the third
91. Prep+Art+Adj+N	on the lower side
92. Prep+Art+Adv	in the past
93. Prep+Art+N+Adj+Prep+N	like a garden full of weeds
94. Prep+N+Art	through love the
95. Prep+N+Prep+Art	on top of the
96. Prep+Aux+V	of being drowned
97. Prep+Adv	by now , at first,
98. Prep+Nu+N	per 1000 population
99. Prep+N	at last
100. Prep+N+Prep	in terms of
101. Prep+Prn	like that
102. Prep+Prn+Adj	in its most
103. Prep+Prn+N	in my experience , at my tongue, by his shirt front (shirt front-N)
104. Prep+Prn+N+Prep+N	by my word of honour
105. Prep+Prn+N+Prn+Aux+V+Prep	with this faith we will be able to
106. Prn	no one , 2. a few,
107. Prn+Adv+V	he almost whispered
108. Prn+Aux+V+Prep	I’m going to, you have learnt to
109. Prn+Aux+N+Prep+Art+N	it was 97 in the shade
110. Prn+Aux+V	you were seen ,
111. Prn+Aux+Art	I am the, I am a,
112. Prn+Aux+Prn+V+Inf V	what are you going to do
113. Prn+Aux+Prep+V	I had to smell,
114. Prn+V+Adv+Prn	he said as he

115. Prn+Mod	I might , I had to,
116. Prn+Mod+Mod	it would be,
117. Prn+Mod+V	I can cure , it may be,
118. Prn+Mod+V+Prn	I would whip them
119. Prn+Aux+Adj	it was like
120. Prn+V+Art+N+N+Prep+Prn+N	she took the card halves from her purse
121. Prn+V+Prn	he told her
122. Prn+V+Prep	She put on , he lived at,
123. Prn+V+Prep+Prep+Art+N	she went back into the kitchen
124. Prn+V+Prep+Art+N	she picked up the receiver
125. Prn+V+Prn	I followed him
126. Prn+N+Aux	her hair was
127. Prn+N	their relevance
128. Prn+N+Conj+N+Aux+V	my father and mother had cleared
129. Prn+Prep	none to
130. Prn+V+Prep	he went into, he came by
131. Prn+V+Prn+Aux+V+Inf V	you thought I was going to say
132. Prn+V+Inf V+Ref Prn	He came to know himself
133. Prn+Adj	6. my dear
134. Prn+Adj+N	his clear images
135. Prn+V	he assumes , I question,
136. Prn+Aux	he was
137. Prn+Aux+Prn+V+Prn	how will you have it?
138. Prn+V+Prep+Adj+N+Prep+Adv+Prn+Aux	I stand in good relation to all that is
139. Prn+V+Prep+Adj+N+Prep+Art	I stand in good relation to the
140. Prn+Prep+Prn+N	one of those mysteries
141. Prn+Prep+Art	one of the
142. Prn+Aux+Prep	I had to
143. Prn+Rel Prn	those who
144. Phr	even if
145. Prop Adj+N	Muslim Spain, Abbasid Caliph, Umayyad dynasty
146. PP	in front of
147. V+Adv	dried up

148. V+Art+N	open an account, make a book
149. V+N+Apo+N	got housemaid's knee
150. V+Prep+Prn	looked at him
151. V+Prep	got up, belonged to, comes in, hung with, begin to, getting into, work on, went up,
152. V+Prep+Prep+N	getting on in years
153. V+Prep+Art+N	pick up the pocket book, looked at the door, hesitated for a moment,
154. V+Prn+N	snatch my pocket book, cleared his throat, open your mouth, made his way, cleared this land
155. V+Prep+Prn	sticks to it
156. V+Prep+N	go to college,
157. V+Prep+Prep+Art	sat down on a
158. V+Prep+Art+N	sat by the fire, went into the living room
159. V+Prep+Art+Adj+N	bitten by a mad dog, made up his mind
160. V+Prep+Prn+N	look at your throat,
161. V+Prep+Prn+Inf V+Art	like for us to have a
162. V+Art+Adj+N	becoming the first customer
163. V+Prn+Adv	placed it before
164. V+Prn+Aux+Art+Adj+N	suppose it’s a genuine offer
165. V+Prn+Art+N+Prep+N	bring me a cup of tea
166. V+Adv+Prep	2.go back to

4.2 Repetition of Collocations

Among 297 standard phrases, nine phrases (full of, one of the, loveliest of trees the cherry now, may be, as well as, and then, to stand, a few, go back to) have been repeated twice, and one phrase (my dear) has been repeated six times.

4.3 Classification Chart of Collocations

Standard collocations of table 2 have been further classified into table 3 with respect to grammatical categories. All extracted collocations belong to the following 18 grammatical categories. If these repetitive phrases are subtracted from total collocations, 283 phrases have been declared as the final number of collocations.

Table 4.1 Collocations with Grammatical Category

Sr. No.	Sequence No.	Grammatical Category	Collocational Instances
1	1-10	Adjective	33
2	11-19	Adverb	13
3	20-33	Article	27
4	34-40	Auxiliary Verb	10
5	41-46	Conjunction	10
6	47	Determiner	1
7	48	Idiomatic expression	2
8	49-53	Infinite	7
9	54-57	Interjection	5
10	58	Modal	1
11	59-76	Noun	37
12	77	Number	1
13	78-105	Preposition	58
14	106-143	Pronoun	52
15	144	Phrase	1
16	145	Proper Adjective	3
17	146	Prepositional Phrase	1
18	147-166	Verb	37
	Total collocations= 166	**Total Grammatical Categories= 18**	**Total Instances= 297**

Discussion on the result includes discrepancies between standard and substandard collocations parameters.

4.4 Difference Between Standard and Substandard Collocations

“In active data mining paradigm,… rules are discovered, …the history of the statistical parameters associated with the rules is updated…” (Agrawal, & Psaila, 1995). As the debate of discrepancy between standard and substandard collocations is concerned, those phrases which consist of most of the function words (prepositions, articles, conjunctions, auxiliary) for example “on the”, “it the” in ‘*Hitch Hiking across Sahara Desert*’ are considered as substandard phrases. These two phrases do not convey any meaningful knowledge pattern because “Collocations as instrumentation for meaning is a scientific fact” (Louw, 2010, p. 79). So, two key criteria have been finalised: firstly, substandard phrases consist of all or most function words and become almost meaningless. Secondly, standard collocations consist of mostly content words, so they carry meaning. More occurrences in corpus prove established norms. More precisely, standard collocations consist of more ratio of content words than function words. Content words and their semantic shades supersede function words.

On technical grounds, machine learning algorithms extract repeated patterns whether they make sense or not. Therefore, the machine has done its work accurately, and next is the human task to recognize and separate standard and substandard collocations. Finding total sense in repeated phrases cannot be done 100% by the computer since 90% accuracy is considered a great success in machine learning models.

Table 2 shows 297 different phrases and their relevant examples from English intermediate textbooks with the help of the Phrases tool. The extraction of collocations is a beneficial knowledge discovery process for learners, teachers, and textbooks writers. The main advantage of super standard collocations is that collocations manifest the ideology and narratology of each lesson. By learning the 166 patterns and 297 standard phrases, intermediate-level learners can comprehend their four intermediate textbooks. Furthermore, they can utilise these phrases in their fluent written and oral exams by expressing their linguistic competence and performance.

4.5 Coinage of New Collocations on the Pattern of Existing Collocations

Keeping extracted phrases as a foundation stone, learners can build other standard phrases for their studies for example, laughter broke out, cholera broke out from sample collocations of “applause broke out”. In fact, language is a chain-like process that starts from unigram to bigram, trigram, quadgram/collocations and finally, sentences are spoken or written to produce a complete language code.

4.6 Extracted Collocations for Teaching Correct English

The extracted standard collocations serve the purpose of learning and teaching accurate language. Collocations emphasize the use of certain words with their adjacent words for example “to seek justice”, “shed the blood”. The correct use of prepositions should be learnt with the examples: “pick up”, “look at”, “in front of”. The accurate use of phrasal verbs should be learnt from the collocation “applause broke out”. The correct use of the auxiliary verb can be learnt from these instances: “he will”, “it was 97 in the shade”. The use of singular and plural agreement with the subject should be learnt from these examples: “three hundred dollars”, “hair was”. Collocations also teach about the correct use of verb form for instance “have liked”. They also teach the sequence of adjectives and nouns for example: “walnut cake”, “great mosque”, “blue suede shoes”.

4.7 Juxtaposing the Current Study with Other Studies

Previously, 37 collocation patterns were found (Shin & Nation, 2007), 33% figurative phrases, 33% idiom (Siyanova-Chanturia, 2017), and in the same continuation, the current research extracts 166 phrase patterns with their 18 grammatical categories. The current study extends previous collocational studies on ESL textbooks (Biber, Conrad, & Cortes, 2004; Hsu, 2008; Lee, 2015; Russell, 2017). The present study extracts lexical bundles as Biber et al’s works (1999, p.992, 993) extract lexical bundles from English textbooks. Learning English is the one purpose of extraction of 166 phrase patterns and the current study is in harmony with objectives of these previous studies (Nesselhauf, 2003; Seretan, Nerima, & Wehrli, 2004; Shin, & Nation, 2007), which were elucidated in the literature review.

5. Conclusion

One of the major findings is a compilation of a list of collocations from intermediate English textbooks. Secondly, it builds an 82,487-word corpus. Thirdly, it transforms the whole corpus into one word as a hyperlink. Fourthly, this study has presented 297 collocational instances, 166 standard collocation patterns, and 18 grammatical categories. Fifthly, the current study draws a dividing line between standard and substandard collocations. The phrase consisting of function words such as “on the” was substandard because it did not convey a strong meaning. Content words must be present in a standard collocation to express meanings. Sixthly, these collocations are useful for learning and teaching purposes as some practical examples have been given in results and discussion sections. Grammatical rules are best learnt with the help of textual instances, so this study provides examples for learning accurate language patterns. To conclude, this study empowers learners and teachers to extract collocations, then it differentiates between standard and substandard collocations, and finally it uses them for pedagogical purposes eliminating linguistic errors and enhancing linguistic fluency.

The application of this study is useful for more than 1 million intermediate students, ESL teachers and textbook writers. This research recommends the inclusion of a standard collocations list in all textbooks besides the unigram glossary. Learning collocations will enhance the fluency of language learners in all linguistic skills. Moreover, learners will memorize useful and repeated collocations since they represent semantics, grammatical knowledge patterns of prepositions, and idiomatic phrases.

As future implications are concerned, students should be empowered to extract collocations as their class projects, and their learning will enhance linguistic fluency in all language skills. ESL teachers should extract collocations from English textbooks, and they should exploit them within their classes. Textbook writers should declare collocations as the compulsory part of textbooks such as the inclusion of a glossary at the end of the book. Some futuristic collocational research can be conducted on IELTS, TOEFL, O Level, A Level English textbooks, classical and modern literature books. On the larger scale, cross-comparison of collocations and other machine learning methods of collocation extraction should be introduced in humanities. One futuristic computer science project should introduce algorithms that can differentiate between standard and substandard collocations.

References

Agrawal, R., & Psaila, G. (1995, August). Active Data Mining. In *KDD* (pp. 3-8). Menlo Park, California: American Association for Artificial Intelligence.

Biber, D. (1993). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, *19*(3), 531-538.

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

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. Language Teaching Research, 18, 54–74.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining. From Concept to Implementation*. Upper Saddle River, NJ: Prentice Hall.

Chen, L. (2017). Corpus-aided Business English collocations Pedagogy: An Empirical Study in Chinese EFL Learners. *English Language Teaching*, *10*(9), 181.

Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2008). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, *30*(2), 236-252.

Choi, H. Y., & Chon, Y. V. (2012). A corpus-based analysis of collocations in tenth-grade high school English textbooks. *Multimedia Assisted Language Learning, 15*(2), 41-73.

Conzett, J. (2000). Integrating collocations into a reading and writing course. In M. Lewis (Ed.), *Teaching collocations: Further developments in the lexical approach* (pp. 70-87). Hove, England: Language Teaching Publications.

Da Silva, E. B., Orenha-Ottaiano, A., & Babini, M. (2017). Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora. *Acta Scientiarum. Language and Culture*, *39*(4), 345.

Dechert, H. (1983). How a story is done in a second language. In C. Faerch, & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 175-95). London: Longman.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61-74.

Durant, G. B. (2004). A typology of research methods within the social sciences. NCRM Working Paper, 1-22. Retrieved from http://eprints.ncrm.ac.uk/115/

Ferguson, P., Siyanova-Chanturia, A., & Leeming, P. (2021). Impact of exercise format and repetition on learning verb–noun collocations. Language Teaching Research, 13621688211038091.

Firth, R. (1957). *Papers in linguistics 1934-1951*. Oxford: Oxford University Press.

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. San Francisco: International Scholars Publications.

Harwood, N. (2002). Taking a lexical approach to teaching: Principles and problems. *International Journal of Applied Linguistics, 12*(2), 139-155. http://dx.doi.org/ 10.1111/1473-4192.00028

Hsu, J. (2008). Role of the multi-word lexical units in current EFL/ESL textbooks. *US-China Foreign Language,6*(7), 27-39.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*(4), 322-32. http://dx.doi.org/ 10.1093/elt/cci061

Lee, J. (2015). The Repetition of Chunks in Korean Middle School English Textbooks. *English Language Teaching*, *8*(10), 60.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice*. London: Language Teaching Publications.

Littlejohn, A. (1998). The analysis of language teaching materials: Inside the Trojan Horse. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp.179-211), Cambridge: Cambridge University Press.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. http://dx.doi.org/10.1093/applin/24.2.223

Newell, A., & Rosenbloom, S. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.) *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: computer-assisted interpretation in the humanities*. MIT Press.

Russell, T. S. (2017). *Finding the formula: formulaic language use in Hong Kong primary school English textbooks* (Doctoral dissertation, University of Birmingham).

Seretan, V., Nerima, L., & Wehrli, E. (2004). A Tool for Multi-Word CoUocation Extraction and Visualization in MultUingual Corpora In *Proceedings of the 11th EURALEX International Congress*. Universite´ de Bretagne-Sud, Faculte´ des lettres et des sciences humaines, 2004. p. 755-766.

Shin, D., & Nation, P. (2007). Beyond single words: The most frequent collocations in spoken English. *ELT journal*, *62*(4), 339-348.

Sinclair, J. McH. (1987). Collocations: A progress report In R. Steele & T. Threadgold (eds): Language Topics: Essays in Honour of Michael Halliday. Amsterdam: John Benjamins. pp.319–31.

Sinclair, J. (1991). *Corpus, concordance, collocations*. Oxford: Oxford University Press.

Sinclair, J. M. (2007). Collocations reviewed. manuscript), Tuscan Word Centre, Italy.

Siyanova-Chanturia, A. (2017). Researching the teaching and learning of multi-word expressions.

Stengers, H., Boers, F., Housen, A. & Eyckman, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? IRAL, 49(4), 321-39. http://dx.doi.org/10.1515/iral.2011.017

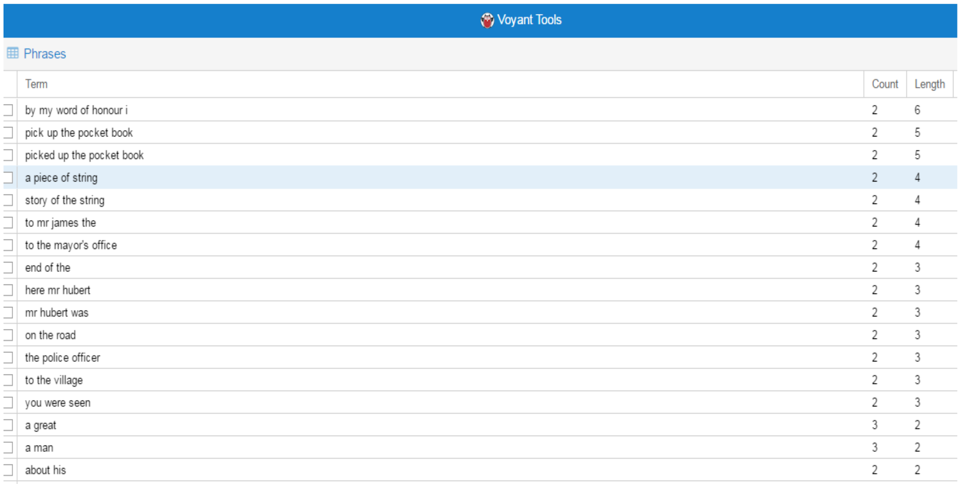
Taguchi, N. (2008). Building language blocks in L2 Japanese: Chunking learning and the development of complexity and fluency in spoken production. Foreign Language Annals, 41(1), 132-156. http://dx.doi.org/10.1111/j.1944-9720.2008.tb03283.x

Wang, J., & Good, R. (2007). The repetition of collocations in EFL textbooks: A corpus study. Paper presented at The Sixteenth International Symposium and Book Fair on English Teaching in the Republic of China, Taipei.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of the play. Annual Review of Applied Linguistics, 32, 231 254.http://dx.doi.org/10.1017/S026719051200013X

Wray, A. (2013). Formulaic language. Language Teaching, 46(3), 316-334.

7. Appendix A



	Term	Count	Length
□	by my word of honour i	2	6
□	pick up the pocket book	2	5
□	picked up the pocket book	2	5
□	a piece of string	2	4
□	story of the string	2	4
□	to mr james the	2	4
□	to the mayor's office	2	4
□	end of the	2	3
□	here mr hubert	2	3
□	mr hubert was	2	3
□	on the road	2	3
□	the police officer	2	3
□	to the village	2	3
□	you were seen	2	3
□	a great	3	2
□	a man	3	2
□	about his	2	2