Exploring the Use of Collocation in the Writing of Foundation-Year Students at King Abdulaziz University

Huda Yahya Y. Khoja

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Education

March 2019

Declaration

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Dedication

To my parents, my everlasting inspiration

To my husband and son, the sparkle of my life

Acknowledgments

I would first like to thank my supervisors, Alice and Richard, who supported me in all of my research stages. Beside the academic support, they have always been thoughtful and aware of any personal issues that may have disturbed or distracted me while I was here in the UK as a single parent, in the midst of all of my PhD burdens.

I would like to acknowledge the support and funding of the Saudi Cultural Bureau and the Saudi Ministry of Higher Education. I also would like to extend my gratitude to the Vice Dean of the English Language Institute at King Abdulaziz University, Badia Hakeem, the staff members and students for their help and assistance.

My family of which I am always proud of and feel so blessed to belong to. My parents, Yahya and Suaad, without whose continuous encouragement and valuable words, I would not have had the strength to go on. Their daily 'FaceTime' calls have meant the world to me, especially, when I was overwhelmed with work and felt had nothing to say. Seeing their faces and hearing their voices would comfort me. They had a dream and I had the commitment to make it true. My husband, Wesam, who sacrificed more than five years of our marriage and of our son's early years, tolerating this separation, has done no less than my parents in encouraging me to earn my PhD, and has played a significant role in this achievement. I know that I am lucky to have such an understanding, caring and loving husband as him. Lastly, and most importantly, Aous, my five-year-old son, my little hero, who deserves this degree as much as I do. He was the fuel that kept me waking up every morning looking forward to a better day, investing in every minute I had while he was in nursery, and now in school, and still had me able to enjoy a mother-son dinner and a bed-time story by the end of the day. He was the distraction and freedom I needed, away from my PhD. It hurts when I think of the time that I had to take him to nursery

when he was just a couple of months, yet it fills me with happiness and joy today as I am writing this page of his and my own life.

Talking about ‘collocation’ and the company it keeps, my PhD years would not have reached this stage without the company I kept. My three musketeers are Areen, Dina and Rawan, who filled my life with hope as much I filled their lives with complaints. They listened to all sorts of problems and were always a shoulder to lean on with a solution to carry on. They taught me that miles never count, and by lifting others we raise ourselves. My research would not have seen the light without my friend Alison’s continuous encouragement, positive vibes and endless support. My friends in Leeds, who appeared throughout my five years, Mona, Leena, Lujain, Amal and Ruaa: I always knew that God would never leave me and my son without company here. They were the family we had. The last and not the least, Nuha, who has been my constant support during my last year. I deeply appreciate her existence during this final stage, and a thank-you would never be enough. I will always remember the Nero window-seat times of sharing, complaining, crying, laughing, and hugging; those moments said a lot about our friendship even though we knew each other back home.

A final and sincere thank you to all of my brothers, sisters, nieces, nephews and in-laws for all of the support that you have given me, including visits, phone calls, messages or even secret prayers.

Abstract

English Foreign Language (EFL) learners face many challenges in the process of learning and using the language in a native-like way. One of these difficulties is the written production of collocations, where they know two words but fail to connect them accurately. Even though there has been a continuously growing interest in investigating EFL learners' written production of collocations, there has been little research carried out in the context of Arab EFL learners. This was especially the case when studying the different types of collocations investigated in this research project comparing two proficiency levels of learners, where collocation is not explicitly taught in this context. This research draws on a selection of 16 written samples produced by two levels of Saudi foundation-year students (pre-intermediate and intermediate) at King Abdulaziz University. The methodology follows three analytical methods to investigate learners' written production of collocations: the manual extraction of candidate combinations; then, the corpus-based approach to identify collocations by using BNC citations and an association measurement; and finally, the phraseological approach to identify acceptable collocations by referring to native speaker informants. The results based on this analysis support the research findings that Saudi learners produce a high number of acceptable collocations, and without much difference between pre-intermediate and intermediate level learners. The findings have also identified the types of acceptable collocations and their level of fixedness and the less idiomatic combinations indicating possible similarities and differences between the two levels of Saudi learners' productions. These findings contribute to EFL learners' research on collocations and language learning process and pedagogy, with a limitation to this learners' context. The study further contributes theoretically and methodologically to knowledge through the identification of collocations produced by learners using the association measurement, LogDice.

Table of Contents

List of Tables

List of Abbreviations

BNC	The British National Corpus
CIA	Contrastive Interlanguage Analysis
COCA	The Corpus of Contemporary American English
EFL	English Foreign Language
ESL	English Second Language
ICLE	International Corpus of Learner English
KAU	King Abdulaziz University
LLR	Language Learning Research
SLA	Second Language Acquisition
VAN	Verb-noun, Adjective-noun, Noun-verb and Noun-noun combinations

Chapter 1 Introduction

1.1 Purpose of the study

Collocation is the tendency of a word to attract, with a greater than expected chance, another specific word (Hunston, 2002:68). The effective production of collocations is an aspect of language proficiency, and development in L2 language learners. If they wish to achieve mastery or native like fluency, they need not only to learn vocabulary and grammar but also to acquire and use the language in an unmarked, native-like way. This acquisition includes having an adequate working knowledge of particular elements of a language, such as collocations.

Sinclair (1991:109) links the importance of collocations to the way any language speaker normally uses familiar and fixed word combinations, rather than just single words. Howarth and Cowie (1996a:82) describe the way a collocation exists in the brain as, "familiar, or institutionalized, stored, or memorized, word-combination with limited and arbitrary variation". Because of this natural processing of a language, native speakers are able to distinguish "non-collocational combinations" and feel that non-collocational combinations "are not fluent, not elegant, or just not the usual way" to say something (Heid, 1994:229). Adult native speakers produce collocations naturally and effortlessly because they are unconsciously sensitive to them. This sensitivity is most probably why collocations are frequent and natural to native speakers of any language while they do not appear to be transparent nor always predictable to non-native speakers. Thus, it could be challenging for language learners to fully comprehend and produce collocations effectively in English. Manning and Schutze (1999:184) claim that when a combination cannot be translated into another language, it indicates a collocation, as it is a native-speaker language feature that cannot easily be translated literally. However, this is not always true, as some studies such as Ibrahim et al. (2012) and Parkinson (2015) suggest that Iranian and Mandarin learners do have some similar L1 counterparts to English collocations, which rather results in a better production of collocations.

Since Sinclair (1991) highlighted the importance of collocation, there has been a growing interest in the importance of collocations used by language users and particularly L2 learners. Acquiring the ability to produce acceptable collocations is an important sign of proficiency as well as fluency, which may take L2 learners years to develop. Learning a language using word combinations may potentially assist the learning process, resulting in a higher proficiency level. Nation (2001:318) states that, “all fluent and appropriate language requires collocational knowledge”, to use the language appropriately rather than learn and use it as individual pieces. Schmitt (2010) also indicates that, in order to create a functional vocabulary size in a language, one needs to be familiar enough with a lexical unit, for example a collocation, to use it appropriately. Schmitt and Li (2010:16) further assert that this familiarity is achieved only when “the word becomes more established”, as a collocation, which “determine[s] the degree of higher-level mastery of a lexical item”. Furthermore, Wray (2000:474) argues that collocational use is important to ease the communication process by saving production and processing effort. For her, collocation use can shorten the process of delivering certain messages and express particular functional meaning, in some occasions, such as requests and commands. There is a benefit in producing meaningful collocations, and less word-by-word encoding process (Wray: 2009).

Therefore, an appropriate understanding of collocation is required in order to enable learners to acquire language comprehensively by employing reading, writing, speaking, and listening skills effectively. However, L2 learners face a challenge in producing collocations appropriately. Martyńska (2004:4) considers this challenging issue of L2 learners as an expected phenomenon being non-native speakers, and they should learn and memorize collocations as produced by native speakers. Thus, L2 learners need to be taught collocation and practice using them in a proper context. To investigate the challenges that face learners of English, research has been undertaken with learners of different L1s, by writers such as Granger (1998) with French learners, Siyanova and Schmitt (2008) with Russian learners, Kurosaki (2012) with Japanese learners, and Bahumaid (2006) and Brashi (2009) with Arab learners. Findings from those studies show, as will be discussed in depth in Chapter Two, that L2 learners tend to perform better in reception than production. Furthermore, this challenge in production varies among the different types of collocations. Shehata (2008) and Kuo (2009) have shown that some types of collocations such as verb-noun collocations are more problematic than other types such as adjective-noun collocations. Only a very few studies have addressed noun-noun collocations, including Parkinson (2015). There is a need to investigate the written production of these different types of collocations by Arab learners of English.

The production of these types of collocation is significant for language proficiency in the context of L2 learners. A number of studies suggest that even advanced-level EFL learners still struggle to produce accurate collocations (Huat, 2012; Ibrahim et al., 2012). A higher language proficiency can result in better production of collocations as indicated by Laufer and Waldman (2011). Nonetheless, because collocations can sometimes be found as "ready-made" combinations in the language of native speakers, where the appearance of a word is not entirely predictable to learners, this may be why collocations can still be problematic and not easily or accurately produced by L2 learners, even those at advanced levels. Henriksen (2013:32) writes that collocation differs according to its transparency to the learners: *take a book* is fully transparent i.e. a literal meaning, and *take a course* is semi-transparent i.e. a semi-literal meaning. While both of those types could be learnt and produced appropriately, the learners' proficiency level is still an issue which could affect their collocation production or not. Lack of collocational knowledge and inaccurate production of collocations could affect learners' performance in different communicative skills as suggested above (Wray, 2000). This study aims to investigate possible similarities and differences between two levels of Arab learners of English. It is valuable to conduct such an investigation because the EFL learners in this context are not given explicit teaching in collocation. Thus, it may be more challenging to them, and some aspects, types of collocations and the learners' level, need particular exploration.

The EFL learners who are the focus of this study are Saudi foundation-year students at King Abdulaziz University (KAU). They face a number of challenges in learning English. While there may be some differences in the students' language use due to their different proficiency levels, they still share some common language difficulties. As this study will show, these difficulties include grammatical mistakes, such as subject-verb agreement and grammatical references, and lexical mistakes, such as the inaccurate production of collocations. Subject-verb agreement and other grammatical problems, such as *the boys plays* and *by she* instead of *the boys play* and *by her,* are usually addressed in grammar classes. Unfortunately, in this context, as it is not a part of the teaching materials, teachers rarely address problems such as the use of *noisy voice* instead of *loud voice* or *loud sound.* These examples indicate that students may know two words e.g. adjective before a noun, but fail to connect them accurately to the appropriate collocates. In order to understand the Saudi learners' process of language learning at KAU, specifically, through evaluating their production of collocation in their writing, this study is needed. The Saudi university students under investigation are not taught collocations explicitly; thus, they sometimes succeed while others produce examples that do not sound native-like, as shown by the examples extracted from the written texts of Saudi students in this study. Nation and Shin (2008:340) indicate that the use of natural language for EFL learners is problematic, especially when language teaching focuses primarily on grammar. They illustrate this argument with examples of Korean students who inaccurately form collocations such as *thick tea* and *artificial teeth* instead of the collocations *strong tea,* and *false teeth*. Even though the former are grammatically acceptable, they are not like native speakers' usage.

This study will manually analyse 16 written samples from two levels of students, pre-intermediate and intermediate, to extract candidate collocations. These will then be identified according to fixedness using native-speaker corpus – the British National Corpus (BNC) – as a reference on the Sketch Engine tool. This corpus tool will further assist the judgment of identified collocations by allowing the use of the statistical association measurement, the LogDice in this case, to test the collocations' exclusivity. This use of corpus allows the examination of learners' production of the different types of collocations– verb-noun, noun-verb, adjective-noun and noun-noun in their written texts. As existing collocation research of Arab learners employed experimental methods that usually focus on particular lists and/or specific types of collocations (Shehata, 2008; Brashi, 2009), this study aims to widen the scope of this context by conducting writing analysis. Nevertheless, this corpus approach does have some limitations, especially by referring to one native corpus only. A further identification approach known as the phraseological approach will be used at a later stage of the analysis, which relies on native speaker informants'' judgement. It became essential to the study to combine the two approaches due to the learners' situation i.e. not being taught collocations, which may lead to them producing less idiomatic combinations creatively and/ or mistakenly. Such combinations may not occur in the corpus, but can be identified by native speaker informants through this phraseological approach. Granger (1998:59) suggests that studies of L2 learners need contrastive investigation to address the different needs of the learners' L1, "because not all learner problems are transfer-related". Different L1s can have similar problems in their collocational production, as well as problems unique to their L1, therefore, conducting this study in this context is important. Manning and Schutze (1999:186) indicate that *make a decision* is a problematic collocation for non-native speakers because the verb *make* does not indicate a specific meaning by itself, which is the case also for *do*, *have* and *take*; they need to be complemented. However, collocations including those verbs occurred in this study, and not all of them were found to be problematic for Arab learners as *make* and *take* are used in a similar way in Arabic.

Investigating learners' production of collocations by using this corpus approach following Sincalir (1991), Hunston (2002) and Brezina et al. (2015) attempts to contribute to EFL learner research. Much corpus-based research has been conducted to study the production of collocation by L2 learners to investigate just one type of collocation (Nesselhauf, 2003; Siyanova and Schmitt, 2008; Kuo, 2009) or particular contexts i.e. EFL vs. ESL (Parkinson, 2015), or EAP (Durrant and Schmitt, 2009). This study aims to investigate two levels of EFL learners i.e. pre-intermediate and intermediate, and compare their production to native speakers' language available in the BNC, and widen the importance of this investigation through using the LogDice association measurement to test collocation's fixedness.

By comparing the production of Saudi university students' collocations in written texts and the native language available in the BNC, this study has three objectives. Firstly, the study will investigate Saudi learners' production of acceptable collocations, and in what types. The study will also investigate the less idiomatic combinations produced by those learners, and their types. Finally, the study will investigate similarities and differences between the productions of acceptable collocations by the two levels of Saudi learners: pre-intermediate and intermediate levels.

1.2 Thesis structure and research questions

The thesis has seven chapters, including this introduction. The literature review is divided into two parts: Chapters Two and Three. Chapter Two discusses understandings and approaches to the identification and definitions of collocations. It begins with an introduction to collocation that shows the variety of understandings and definitions currently in use. It then discusses the different views of studying collocations in corpus linguistics, showing the approaches used to identify collocations. Chapter Three reviews the relevant studies of L2 learners and their production of collocations, describes research perspectives to identify collocations, and then discusses similar studies of L2 learners that have investigated issues related to the production of collocation such as learners' language proficiency and types of collocations.

Chapter Four describes the research design of the study before discussing its context and participants. It also illustrates the methods used for data collection and sampling. The final section of Chapter Four demonstrates the analytical procedure of the written texts, which consists of three steps: extracting manually candidate collocations from learners' texts; identifying these extracted collocations in the corpus; and referring unidentified collocations to native-speaker informants' judgement. Chapter Five details the data analysis, which starts by presenting case studies from the two levels of Saudi learners (pre-intermediate and intermediate) investigated, and then discusses the analysis of each level, and their production of collocations.

Chapter Six is the discussion chapter that starts by describing the main findings of the study by addressing the three research questions:

1. Do Saudi foundation-year students at university produce acceptable collocations in writing? If so, what are the types?
2. Which less idiomatic combinations do Saudi foundation-year students produce in writing?
3. What are the similarities and differences between the acceptable collocations produced by two levels of Saudi foundation-year students, studied in their written texts?

The remaining sections of Chapter Six discuss the issues raised by the findings described earlier in the chapter. Finally, Chapter Seven is the conclusion, reflecting on the implications, contributions and limitations of the study and suggestions for future research opportunities.

Chapter 2 Approaches to Identifying and Understanding Collocations

2.1 Introduction

As collocation has been widely studied, different understandings and approaches to definitions have emerged. Pawley and Syder (2014:212) write that a collocation could be a complete sentence, or less than a sentence, such as a clause or a phrase in the native spoken language. Mel'cuk (1998:23) says there is "no universally accepted formal definition of collocations nor a proposal for their uniform and systematic treatment in dictionaries". Wray (2000:264) also suggests that there is no easy way to differentiate between definitions of collocations. This chapter will initially give an overview of the understandings of collocations, what should be identified as a collocation and on what basis. This will lead to a discussion of the definitions of collocations given by different scholars according to collocation fixedness.

Secondly, this chapter will look at how corpora can be successfully used as a reference to identify collocations. Hyland et al. (2012:3) consider corpus linguistics a methodology in language studies because it enables researchers to refer directly to real language available in the corpora rather than consulting language specialists. However, it should not be the only appropriate approach to identify collocations in the current context. Some researchers have employed another approach to identify collocations, which is the phraseological approach. A detailed discussion of its advantages and limitations will be provided in the latter sections of this chapter before reflecting on the theoretical framework of the study by describing the definition of collocation under investigation through the use of the corpus and phraseological approaches to identify collocations produced by Saudi EFL learners.

2.2 Understandings of collocation

The term ‘collocation’ was originally developed from the Latin verb ‘collocare’, which means ‘to arrange’ (Martyńska, 2004:2). Early English/Latin bilingual dictionaries published in the sixteenth century included variations of the word ‘collocatus’ (Barnbrook et al., 2013:5). Different meanings of the Latin word came into the English language to form words, such as ‘to collocate’, meaning ‘to place, to set, to appoint to a place’, and ‘collocation’ showing ‘the act of placing; disposition’ or ‘the state of being placed’ (Barnbrook et al., 2013:5-6). Thus, the word ‘collocation’ linguistically reflects the act of joining words together to show a coherent and a particular functional meaning.

Palmer (1933) is the first researcher to refer to collocation as a vital phenomenon in the English language and argue that it should be given attention. In his, *Second Interim Report on English Collocations,* he defined collocation as “a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts” (cited in Granger and Meunier, 2008). Palmer’s initial definition reflects on the fundamental criterion of collocation, where some words or groups of word, tend to appear together and so need to be learned and recalled as lexical units. There was no indication for further features or characteristics until Firth developed a definition of collocation in the 1950s.

Firth (1957a:179) proposed that in a collocation, “You shall know a word by the company it keeps”, highlighting the connection between the parts of the collocation. He explains this by a number of examples, such as *cow* and *milk,* and *night* and *dark,* and stating that, “One of the meanings of *night* is its collocability with *dark*”, and when reversed, the words still form a collocation: “and of *dark*, of course, collocation with *night*” (1957b:196). Because words appear in each other’s company reflecting on their functional meaning together, a word in a collocation enables the language user to predict the other word and what will come next. Firth referred to this as the “mutual expectancy” of collocates where there is a possibility of expecting the complete collocation while reading or listening to part of it, which is also beneficial in learning common and central words such as “key-words, pivotal words, leading words” (Firth and Palmer, 1968:106). To them (1968:110), while grammar can be translated, it is not easy “to find parallels for collocations of pivotal words in any other language”. This is true for Arab learners of English as, for example, the adjective *strong* means powerful or tough قوي in Arabic. Thus, it could be problematic for Arab learners when translating it in a collocation such as *strong tea*, giving a very strange meaning in Arabic شاي قوي. Even though there is an Arabic collocation conveys the meaning of this collocation, *strong tea*, it literally translates in English into *heavy tea* شاي ثقيل, which is non-acceptable in English. Such words should be learnt in collocations as suggested by Firth and Palmer. It is still the case with grammar as Arabic grammar is different than English grammar, and cannot be translated word by word. Yet, grammar is given a lot of attention in the context of English teaching to Arab learners while collocation is usually not.

The feature of the mutual expectancy of collocations has led to the idea of studying collocation in contexts. Firth’s (1957b) contribution was not only theoretical, but also methodological. He was the first to use data from real produced language when researching collocations, focusing particularly on spoken language. As a result, he directed attention to the use of real language data to study and investigate collocations, which would fill a gap in L2 lexical use, both in comprehension and production, by showing words and meanings in contexts. Firth (1957b:181) additionally, introduced the "extended collocations", which refers to long-fixed expressions such as lexical bundles e.g. *from now on*. Given the focus of this study, the literature review will remain and develop only from his perspectives of collocation and its feature of mutual expectancy. Firth made a major contribution to the understanding of collocation by highlighting the fundamental feature of a collocation, which are words that have a larger chance to appear in the company of another word than any other random word. Corpus linguists took this notion into consideration and developed the understanding of collocation further in various ways to identify and study it as will be shown in the later sections in the chapter.

The most important contributions to modern research on collocation were made by Sinclair (1991) and Halliday (1995), as they converted the Firthian definition of collocation into theoretical and methodological practice, specifically, in generating the computer methods of corpus-based studies of collocations. Halliday's name is associated with 'neo-Firthian' 'scale and category grammar' (Palmer, 1968:9), and this part in the study of language is concerned with grammatical combinations that are known as colligation, which is not in this study's domain. However, Sinclair and his followers' work was mostly dedicated to collocation, as well as in relation to its studies in corpus linguistics, which is this study's focus. The following review will discuss the understandings and definitions developed after Firth by Sinclair and Halliday.

Sinclair (2004:25) explained the phenomenon of collocation as "words enter into meaningful relations with other words around them". Developing Firth's understanding that collocates co-occur in a company and one of the two collocates primes the other word, Sinclair defines collocations not only as word combinations that can occur close to one another, but also as combinations that are bound meaningfully. Sinclair gives a collocation three specific elements; the node, collocate and span. He describes the word under investigation as the "node", any other word occurring in its "specified environment" as the "collocate" (1991:115), and the "distance between two collocating words" as the "span" (2003:179). Sinclair (1991) argues that the appearance of part of the collocation makes the occurrence of the remainder of it predictable. This means there is a greater chance for a word to appear as a collocate to a certain word than other words that may occur randomly and arbitrarily. This is explained in a further description he gives for a collocation as "a frequent co-occurrence of words" that "doesn't have a profound effect on the individual meanings of the words, but there's usually at least a slight effect on the meaning, if only to select or confirm the meaning appropriate to the collocation, which may not be the most common meaning" (2004:28).

Collocation can work as "a good guide to meaning" (Sinclair, 2003:38), which is usually the case of nouns that are, by themselves, "ambiguous in meaning" most of the time. Appearing in collocations is a good indication of the meaning that is meant and associated with the nouns. To Sinclair (ibid.:36-37), nouns can indicate three notions: physical, such as *health* and *world*; ideas normally associated with the physical, such as *activity* and *attack*; and non-physical, such as *problems* and *attitude*. For example, the non-physical noun *problems* has a general meaning by itself. However, when it is associated with a physical noun such as *health*, the physical noun *health* gives the non-physical noun *problems* the relevant meaning. Collocation, to Sinclair, shapes a triangle of a node, a collocate, and a span, in which the node and the collocate form a semantic relationship determined by the span, and where the absence of one of the nodes or the collocates would leave the other meaningless. The meaning is not usually indicated by one word only nor through an arbitrary choice; it rather requires more than a word with an attractive "co-selection" (Sinclair, 1991:133).

This led Sinclair (1991:109) to discuss meaning interpretation in texts according to two principles: namely the open-choice principle, and the idiom principle. The open-choice principle, as he defines it (ibid.:109) is the usual way of describing the language, where "at each slot, virtually any word can occur" and often can be called "slot and filler". In this model, the node is the only choice and grammar is the only constraint. He (2004:29) also describes it as the terminological tendency because it reflects words in fixed meanings, where there is "little option but to use it" unlike the phraseological tendency which reflects on the idiom principle of a more natural reference. Sinclair (1991:110) believes that the open-choice principle is not enough to "produce normal text", which is how the world is organised in a particular way. The importance of the idiom principle and the phraseological tendency comes from explaining what is not explained by the open-choice principle, such as the phenomenon of collocation. He explains this by using Halliday's examples of "*strong tea* and *powerful engine"*, where the two adjectives share the same meaning, but they are not "interchangeable" when used with *tea* and *engine* reflecting strength (2004:29). The idiom principle explains this "meaning dependency", where the first words *tea* and *engine* that have the "core meaning" are chosen for their frequency and then the collocates *strong* and *powerful*, which are less frequent and meaning dependant.

To Sinclair (ibid.:115), collocation, "illustrates the idiom principle", in which "on some occasion, words appear to be in cohesion in pairs or groups and these are not necessarily adjacent". He explains this further that the language user usually has "semi-pre-constructed phrases", which assist the person's choice of words whether or not he or she is aware of this choosing process (Sinclair, ibid.:110). Thus, it is a natural way of processing a language and it carries as much significance for any text meaning and coherence as do grammar and structure (Sinclair, ibid.:112). The idiom principle has become an influential concept that has changed the direction of collocation research, from an implied lexical approach to a deeper investigation of language users' processes of comprehension and production. According to Hunston (2009:142), the idiom principle "proposes that most naturally-occurring language consists of a series of chunks rather than a series of independent words". People tend to switch, unknowingly, between the two: the idiom principle and the open-choice, but in analysing their language use, one principle only should be applied (Sinclair, 1991:114), which is the idiom principle in the case of collocation.

Following Sinclair (1991) are Stubbs (2002) and Hunston (2002), whose views of collocation are consistent with those of Firth (1957) and Sinclair (1991), who refer to the probability occurrence of one of the collocates due to the appearance of the other. Stubbs (1996:176) describes collocation as, "the habitual co-occurrence of two (or more) words", meaning that collocates have a greater chance of appearing in each other's environment than random words. He (2001:29) also refers to the three elements of a collocation assigned previously by Sinclair, which are the node as "the word-form or lemma being investigated", the collocate as the "word-form or lemma which co-occurs with a node in a corpus", and the span as "the number of word-forms, before and/or after the node". Stubbs (2002), further, is like Sinclair in indicating that a collocation should reflect a coherent meaning, and that is not only associated with different collocations of different lemmas, but also of different word forms. Stubbs (2009:120) suggests that the form of the word can change its meaning, the different collocations, and as a result the frequencies with which the collocations appear. He explains this concept in the example of the two lemmas *heavy/drink,* which co-occur frequently as collocations in their different forms and meanings as *heavy drinker* and *drink heavily,* but they rarely co-occur, as *heavily drunk,* although this may be due to semantic restrictions. Stubbs (ibid.:30) argues that collocations appear in a "linear string" or "syntagmatic relation", which is different to dictionary representations of words. Because collocations still require definitions beyond the scope of this given by dictionaries with the exact meaning of individual words, there is a need to refer to corpora and frequency in studying and identifying collocations. Information such as frequency, which assigns probability between collocates, and meanings of words in more natural contexts can be found through corpus searches.

Similarly, Hunston (2006) writes that the selection of the collocates as elements in a collocation depends on their relationship with one another. She describes collocation as “the statistical tendency of words to co-occur”, “or as the tendency of one word to attract another” (2002:12, 68). While one of the words is selected according to the meaning it is required to reflect, the other words selected are dependent on the first one. Hunston (2009:143) argues that the appearance of part of the collocation “alters the probability” of the rest of the words in a collocation. Whether these words have strong or weak relationships, it is possible to identify whether a collocation has been established in the language by referring to this probability through frequency in corpora. Hunston reports on studying collocation using corpora through identifying the relationship between the collocates according to their exclusivity. This means that collocates do not only need to co-occur together to indicate a specific functional or coherent meaning, but they also have to be said frequently. Furthermore, collocates have to appear more frequently showing a strong and exclusive bond. This exclusivity can be determined by using the statistical tests that are available in corpus software. While some researchers refer to the use of concordance lines in the corpus to find about language use and patterns, Hunston believes that the statistical tests can provide more accurate information about collocation than a human observer can (2002:12). Because of those tests different collocations can be generated and identified for a node. For example, the strongest collocates for the noun *tea* are *and, caddy* and *cup* respectively according to the three association measurements: t-score, MI score, and LogDice. It is not necessarily the adjective *strong* as cited in earlier discussions such as Halliday (2004) and Nation and Shin (2008). A detailed discussion of this corpus approach in identifying collocations and in relation to how it is applied in this study comes in a later section of the chapter.

By using Sinclair's understandings and descriptions of collocations as a starting point, some scholars modified the definition of collocation to different classifications to apply in language research and teaching. I discuss next those definitions of collocation, and then I will review the approaches to studying it.

2.3 Definitions of collocation

Pawley and Syder (2014:205) relate the linguistic knowledge of a language user to two elements: the "memorized", or morpheme items, and "lexicalized" sequences. Their classification reflects the general classification of collocations as grammatical and lexical, which was also suggested by Bahns (1993:57). Nesselhauf (2004:22) differentiates between the two types of classifications as follows. First, grammatical collocation is the co-occurrent relationship between a lexical word and a preposition: that is, open and closed word classes. These include verb-preposition, noun-preposition, and adjective-preposition, and can also occur in any combination. Second, lexical collocation is the co-occurrence relationship within a two-word combination. The two words are from the category of open-class words, which include verbs, nouns, adjectives and adverbs, and can occur in any of the following combinations: verb-noun, noun-verb, adjective-noun, noun-noun, verb-adverb and adjective-adverb. As this study looks at collocations that are two-word open-class combinations – specifically, verb, noun, and adjective combinations with a noun node – the review only discusses the lexical collocations.

Pawley and Syder (2014:212) also argue

> An expression may be more or less a standard designation for a concept, more or less clearly analysable into morphemes, more or less fixed in form, more or less capable of being transformed without change of meaning or status as a standard usage, and the concept denoted by the expression may be familiar and culturally recognized to varying degrees.

Pawley and Syder indicate, that despite the wide range and various scopes of collocations, there is no consensus amongst researchers as to what criteria should be used to identify them.

There are several elements involved in defining collocations such as learners' production of collocation, functions of collocations, and restrictedness. Henriksen (2013), classifies the concept according to learners' perspectives and how transparent collocations are: namely, fully-transparent, and semi-transparent collocations. Howarth (1998) and Schmitt (2006) classify collocations according to function: for example, slogans, proverbs, catchphrases, and engaging, sequence and technical combinations. McEnery and Wilson (1996) link collocation to the features and the types of words that constrain it and focus on the features in the patterns of collocation. Others such as Bahns and Eldaw (1993:57) define collocation simply as "the regular occurrence together of words", which is relevant to the understanding of Sinclair and followers discussed in the previous section of collocation, and will lead to collocations' definition according to their restrictedness or fixedness.

A number of scholars agree on two broad types of collocations, the 'free combinations' and 'idioms', with a cline or other groups between these two extremes. Handl (2008:50) states that though there are two basic categories of collocations: the 'free combinations' and the 'fixed combinations', there is still a need for "a stretch on the continuum" between the two. To start with, Benson et al. (1986:252) categorise collocations into five types: ‘free combinations’, ‘collocations’, ‘transitional combinations’, ‘idioms’, and ‘compounds’. Free combinations, as the term suggests, occur freely between two lexical items. Any word from the two-word combination has a number of choices to collocate with, e.g. *drink* can combine with words such as *water*, *coffee*, or *tea*. Collocation also involves the co-occurrence of two lexis, though the combinations are not as open to selection of collocates as free combinations. As the two types seem similar, Benson et al. (ibid.:253) distinguish between them with the example of *commit murder*, in which the verb is limited to the meaning reflected by the collocation, and the collocation of the two words is frequent. Frequency and restrictedness of collocates in a combination indicate a collocation, meaning that they appear together more than randomly. The third type is the transitional combinations, which are more restricted than collocations but less idiomatic. An example of a transitional combination is the expression *catch the bus*. While *the bus* reflects its usual meaning as a common noun, the verb *catch* does not indicate its usual meaning, as in *catch the ball*. Fourthly, idioms that are combinations reflecting a meaning other than the literal meaning of the words, for example, *raining cats and dogs*. Finally, compounds are combinations of adjective-noun, noun-noun, verb-preposition or verb-adverb which are “frozen” compared to the other combinations: free combinations, collocations, and transitional combinations. Examples of compounds include *fire escape, night owl, carry out* and *run across* (ibid.:254). Although idioms, like compounds, may also be seen as fixed to the other types, importantly, they are figurative, or have figurative origins. As collocations are usually used to refer to literal meaning, which is the focus of this study, the definition of collocations for the purposes of this research does not include idiomatic types, such as transitional combinations or idioms.

In Bahns' (1993) model, collocation types include 'free combinations', 'fixed combinations' and 'idioms'. Similar to Benson et al. (1986) and to Bahns (1993:57), the free combination is "the least cohesive", idioms are figurative, and the fixed combinations are between them. Fixed combinations, reflecting collocations, are considered to be on a scale between free combinations and idioms, literal most of the time unlike figurative idioms, but more restricted in collocate choices unlike free combinations. Howarth (1998:164) uses what he refers to as "the continuum model" to distinguish between those types: 'free combinations', 'idioms', 'restricted collocations', with a further division for 'idioms' as figurative and pure. As free combinations reflect the literal meaning of a collocation, their substitution does not affect the other collocate: for example, *write a comment, write a full stop,* and *write an address*. The same verb *write* can be used with a wide range of nouns, and substituting it with another relevant verb does not affect the meaning of the collocation: for example, *put a comment* vs. *write a comment*. On the other hand, restricted collocations require specific knowledge of the lexical items, and not just a surface co-occurrence. For example, the expressions *commit a crime* and *commit suicide* show strong and fixed relations between the node and the collocate, and the meaning cannot be conveyed using another verb. The last two types defined by Howarth are the idiomatic types: the 'figurative idiom' and 'pure idiom', which require "a semantic unity". Howarth (ibid.:168) gives an example of the figurative idiom as *put a premium,* in which part of the idiom *premium* is figurative while the other part *put* reflects the literal meaning. A good example of a pure idiom is *kick the bucket*, meaning 'died', as this meaning cannot be predicted from its components.

When Huang (2001) investigated the four types of lexical combinations: free combinations, restricted collocations, figurative idioms, and pure idioms, employing the classification developed by Howarth (1998), he found that learners were more likely to produce free combinations accurately than the other types. The least likely types produced by learners were pure idioms such as *had a whale of a time* while the restricted collocations such as *propose a toast* and figurative idioms such as a *paper tiger* were equally used. Huang suggested that there is a grey area in the middle between the free combinations, which learners do not find too challenging and idioms which they do not usually use. Relatively low-level learners, whose language proficiency level is not up to the idiomatic production of combinations and expressions, yet unchallenged for the production of the free combinations, are more likely found struggling in this grey area of fixed collocations. Fixed collocations are combinations of transparent meaning, but with restricted choice of words to convey this meaning naturally. The same issue was addressed by Nesselhauf (2003) in her investigation of the production of verb-noun collocation particularly. She suggested that the arbitrary or free combinations, which she names as ‘frequent combinations’, are obvious and easy to produce by learners. However, learners struggle in the ‘phraseological combinations’, meaning the restricted collocations and idioms, where words co-occur according to their restricted and fixed meanings unlike the case of the free combinations.

According to Howarth (1998:186), many learners, even at advanced levels, can only distinguish between free combinations and idioms and lack the sense of collocations that fit between these two types, which can be referred to as restricted collocations. The students’ texts I am investigating are relatively low level and they have not been taught collocation explicitly, so the fixed type of collocations, the most relevant for this study will be focus of my investigation.

In order to conduct this investigation according to aforementioned understandings and definitions of collocations, a combination of two approaches to identify learners’ production of acceptable collocations will be used, i.e. the corpus approach and the phraseological approach as will be discussed next.

2.4 Approaches to identifying collocation

Research on collocations includes both qualitative and quantitative approaches, which can be used jointly or separately to identify collocations. As the following discussion will describe, the quantitative approach considers the frequency of the collocation as given by corpora and statistical measurements of associations, while the qualitative approach, which is also known as the phraseological approach, refers to native speakers' judgement. McEnery and Wilson (1996:63) argue for a combination of both approaches as it combines the accuracy of the quantitative approach with the reliability of the qualitative approach. Granger (2018:228) adds that the quantitative measures should be chosen and applied with a special care as they contribute to the study findings on learners' production of collocations.

2.4.1 The corpus-based approach

Using the corpus-based approach does not imply that it is the only valid method to investigate language use. The most effective methodological approach depends more on recognising which approach is the most suitable in order to answer the research questions. There are a number of reasons that linguists view the use of corpus linguistics as an appropriate methodological tool. Johansson (1991:313) notes that corpus functions as an essential tool for linguists that complements other techniques and tests that are applied in language studies. McEnery and Wilson (1996:1-2) write that a corpus based approach can be applied to any area of a language, rather than functioning in only a single branch of linguistics. It builds a bridge between data and theory as suggested by Tognini (1996:65), "corpus validates and quantifies linguistics theory". Nesselhauf (2003) claims that a corpus-based approach is advantageous because it diverts the focus away from language errors to language features and patterns of users and learners. Teubert (2004:112) writes that it is valuable to provide evidence from and about language use by native speakers and language learners due to the long-term saved data, unlike short-term methods e.g. collected data from experimental and interventional methods. McEnery et al. (2006) add that a corpus-based approach is able to provide data that is objective and free from external influences as it is based on real examples, rather than other approaches which could be subjective and affected by factors such as the researchers' interests, which may lead to the invention of examples according to whether or not they wish to prove a hypothesis. Hyland et al. (2012:3) write on the influence of using corpus linguistics in applied linguistics as, "enormous, transforming both how we understand and how we study language across a range of different areas". Huat (2012:192) indicates that the corpus-based approach is useful in investigating the five main issues of L2 learners: namely, knowledge, language use, L1 role, instruction, and linguistic contribution. All these reasons make corpora use an appropriate approach to investigate and understand various phenomena in the language, including the use of collocations by native and non-native speakers, specifically L2 learners, as is the case here.

Stubbs (2007:163) states that the areas of collocation and corpus methodology are strongly related because the investigation of collocation indicates that language is dominated by patterns and norms. Stubbs (2002:238) suggests two methods of using corpora to study collocations. The first method is to study words in pairs, because word pairs create coherent meaning together, making a collocation. The second method is to study lexis within grammatical frames, which makes a colligation. The first method is used in the study as the phenomenon to be investigated are the two-word collocations of a lexical relationship which create a coherent meaning, not a grammatical one. On the approaches of studying collocation, Wray (2009:9) identifies three approaches. The first approach is used in research that focuses on the common words of a language and their usage, with relation to their statistical patterns. The second approach consists of studies that test theories using the corpora. The final approach is used in studies that investigate in detail a group of words using corpus tools. As will be seen in the studies discussed in the following chapter, it is important to note that classifying studies specifically under one or other of these approaches is not entirely accurate, as studies can have elements of two or more of these categories. In this study a combination of Wray's approaches is used as collocations produced by learners are tested against those produced by native speakers as available in corpora by using statistics provided by a corpus tool.

Although applying corpus methods does have some drawbacks (Granger, 2009), the use of corpora entails the use of natural language evidence to interpret results and thus assists in creating solutions to a number of linguistic issues such as the use of collocation by language learners. Hunston (2006:234) supports this as she argues against the criticism of using corpora in observing lexical units' frequency only, and instead appreciates the various applications of corpora that actually enrich linguistic research. The use of frequency of collocation available in corpora can provoke different interpretations, depending on research interests and questions. The frequency-based approach is principally concerned with the number of citations of collocations that appear in corpora within a span to determine whether a collocation is frequent or not. Both the span and threshold frequency score are set by researchers and can differ for each study. McEnery et al. (2006:52) describe this frequency as "the arithmetic count of the number of linguistic elements within a corpus that belong to each classification within a particular classification scheme". Granger (1998) defines this frequency-based perception of collocation by the high or low probability of the collocates' co-occurrence. Sinclair (2003:9) states that this approach is important because not all speakers use a language in the same way; language might be repeated and regulated differently by different speakers. Thus, the frequency would reflect the regulations of language users, showing what the most probable, accurate patterns are. These patterns of collocations can be difficult to notice, observe and investigate through concordance lines only, so the use of the corpus tools to generate the association measures based on collocates' frequency and co-occurrence assists the identification process of collocations. Teubert (2004:91) considers frequency an important measurement to determine collocations that are not available in dictionaries but are available in corpora, which was also suggested by Stubbs (Section 2.2).

Brezina et al. (2015:140) write that there are three important elements when searching for a collocation in corpora. The first element is frequency, which reflects the regularity of a collocation. The second element is distance, which is the span between the two words in a collocation. The third element is exclusivity, which is the strength of the collocation. By using this method, a researcher is able to interpret how relevant and co-occurring the two words of a collocation are, and whether they are frequent and exclusive enough to be considered an acceptable collocation. Sinclair (1991:116) discussed earlier this relationship between the frequency and collocation of two words through the collocation strength. For example, he looked at the collocations of the word *back*, as when the word *back* collocates with the verb *bring* as in *bring back*, it makes a downward collocation because it is collocating with a word less frequent than itself, the verb *bring* presenting the semantic relationship. However, when *back* collocates with the preposition *at*, it makes an upward collocation because it is collocating with a word used more frequently than itself, the preposition *at*. Most prepositions, pronouns, and verbs, such as *get* and *go*, are considered frequent, weak patterns (ibid.:116-117) whereas most verbs, nouns, and a few prepositions, such as *along* and *behind*, adverbs like *again* and *forth* and adjectives such as *normal*, are infrequent (ibid.:118). While upward collocations often include high frequency words such as prepositions, adverbs, conjunctions, and pronouns, which make the grammatical parts of a collocation, downward collocations are usually made of nouns and verbs that are of a lower frequency, indicating the semantic part of a collocation (Martyńska, 2004). Nelson (2000) describes this imbalanced relationship of strength in a collocation as "non-reciprocal", a relationship in which every collocation can be a downward collocation for one word and an upward collocation for another.

Subsequently, Stubbs (2001:29) also considers the "frequent co-occurrence" of collocations an indicator of significance. He (2007:127) affirms that frequency can be used as rational evidence because it shows the probability and regularity of collocations in a given language, but he (1995) further discusses this frequent co-occurrence of collocations in terms of collocation strength. The strength of a collocation can be measured using different association measurements in corpora, which the well-known measures and mostly used are: the mutual information (MI-score), *t*-score, and LogDice. According to Hunston (2002:73), the MI-score is "a measure of strength", which reveals the lexical behaviour of the collocation, and the *t*-score is "a measure of certainty" and conveys the grammatical features of a collocation. McEnery and Xiao (2006) write that the MI-score indicates a strong link between collocates if they are given a high score, 3 or more, and the *t* score, which depends on corpus size, is usually significant with a score of 2 or higher. The corpus size is important to the t-score, but it is not significant for the MI-score, as Hunston (2002:73) indicates: "the larger the corpus is, the more significant a large number of co-occurrences is". Thus, the MI-score can be compared across different size corpora, while the t-score cannot. McEnery et al. (2006:57) suggest that an MI high-frequency collocation includes a low-frequency collocate, while collocations with a high *t*-score usually belong to high frequency collocates. As a result, the strength measurement of a collocation is strongly linked to the statistical stance of frequency numbers that are available in the corpora to indicate fixed collocations. The third measurement, which was created to overcome this differences problem among the different measurements, is the LogDice. Rychly (2008:9) argues that, “The LogDice score has a reasonable interpretation, scales well on a different corpus size, is stable on sub-corpora, and the values are in reasonable range”. He (ibid.:9) states that a negative is an insignificant value and each positive point is as twice often for a collocation as the preceding score. The LogDice is different than the other measures in having “a fixed maximum” value which is 14, unlike the rest of the association measurements (Gablasova et al., 2017:164).

Even though the three association measurements refer to the frequency of collocation in corpora, they do not produce the same list of collocations. As collocations are presented in many different combinations of strength and frequency in the corpus, their strength measurement indicates their fixedness and restrictedness. To identify collocations produced by Saudi learners, this study will refer not only to the raw frequency of collocations, but will also apply one of the association measurements; the LogDice. The t-score mainly focuses on high frequency collocates (e.g. grammatical words), which is not the focus of this study, and tests certainty in a way that is very similar to raw frequency, which is not enough to judge the fixedness of collocations. For example, the three top collocates (without counting punctuations) of the noun *food* in the BNC using the t-score are: *and*, *the*, and the verb *be*. The MI score tests the strength of the collocation of low frequency collocates (e.g. lexical words), which is suitable to study collocations’ fixedness. The top three collocates (adjectives) for *food* in the same corpus using the MI score, are *fibre-rich*, *uneaten*, and *high-fat*, they are very technical and academic language, which Gablasova et al. (2017:164) refer to such collocates as “rare exclusivity”. This is not either an adequate measurement to test collocations and their fixedness produced by the learners of this study due to their proficiency level and nature of the context as will be described in the methodology chapter. The LogDice overcomes the problems raised by those two association measurements. Kilgarriff and Kosem (2012:13-14), furthermore, indicate that because the MI score stresses on “rare words”, or “sophistication” of the collocations as suggested by Paquot (2017:6), other statistical measures were designed such as the MI3, log likelihood and Dice coefficient, which resembled the *t-score* in highlighting high frequency and functional words. Thus, the LogDice was presented. As the LogDice is similar to the MI score in being fixed and comparable across different corpora of different sizes, it is similar to the t-score in identifying high frequency collocates too. It is still testing collocation strength and “exclusivity”, which is more suitable in language learning research (LLR) and the level of language learners as indicated by Gablasova et al. (2017a). The three top collocates (adjectives) for the noun *food* in the BNC using the LogDice, are *fast*, *fresh*, and *healthy*, which score 7.7, 7.5, and 7.2 respectively. They are exclusive and make strong collocations and at the level of the learners in this study, especially that L2 learners usually depend on high frequency words (as lexical not grammatical) they learnt already or been exposed to. Thus, the LogDice is the most appropriate association measure for this study investigating collocations and their fixedness produced by low-level learners. A further explanation of the application of this measurement will be discussed in the relevant sections in the methodology chapter.

2.4.2 The phraseological approach

The second approach used to identify collocations is the phraseological approach. While the frequency-based approach deals with the frequency of collocations available in corpora and statistics produced by corpus tools, the phraseological approach refers to the meanings of collocation as identified by native speakers. Nesselhauf (2003) makes the distinction between the two approaches in identifying collocations, that the frequency-based approach is associated with the randomness of collocation, while the phraseological approach is associated with the semantics of collocation. There are occasions where collocations cannot be identified only by referring to their frequency in corpora; therefore, they are referred to native speakers' judgement. Corpora may not include all native speakers' language for different reasons (Hunston, 2002): for example, the corpus size, or its last update. Also, some collocations may not appear in corpora frequently enough to score over a threshold, in studies such as the current one, though, they are still acceptable and sound natural when heard by native speakers. Furthermore, there are the different topics learners may write about and corpora may not include exactly the same words or collocations produced by learners, which is the case in this study. The different contexts, level of language proficiency, and taught curriculum or the language learners have been exposed to, all form a unique language experience for each learner (Gablasova et al., 2017a). Because of these limitations that can occur with the first approach applied i.e. the corpus-based approach, the phraseological approach is also applied as a further procedure for collocations that cannot be identified by referring to corpora. Nesselhauf (2005) argues that seeking native-speaker informants' judgement on collocations would be logical because corpora are a collection of native speakers' language usages.

As the phraseological approach relies on native speakers' intuition rather than on a collocation's frequency in corpora, their judgment about collocations depends on the familiarity and meaning of the collocations to them. Teubert (2004:93) claims that in any language, native speakers are sensitive to meaning and frequent use of words in their language, and so native-speaker judgement is a reliable procedure. However, Hunston (2002:20) writes that "intuition is a poor guide to at least four aspects of language: collocation, frequency, prosody and phraseology." She (ibid.:21) explains that it might be easy sometimes to judge collocations like *play game*, but on other occasions, the task can be very complicated when it comes to collocations including adverbs, or can be impossible when it requires frequency evidence. This complication can be the case even with apparently straightforward collocations such as *high building* and *tall building* as identified by Tsui (2004:45), where she writes that *high building* is acceptable while *tall building* is not. However, by referring to statistics i.e. the LogDice since it is the association measurement used here (available in Sketch Engine), *high building* scores 7.3 and *tall building* scores 6.0, which mean that both are fixed collocations. This shows that human intuition should not be the only valid measurement for a linguistic phenomenon like collocation. Statistics usually give a solid indication, and a combination of the two approaches can be more appropriate for studies like this one. Despite criticism for its purely qualitative stance with no reference to statistics, Hunston (ibid.:22) considers the phraseological approach an important tool in addressing limitations found in corpora.

Researchers such as Nesselhauf (2003) and Parkinson (2015) applied the phraseological approach in their investigation of learners' production of collocations. Nesselhauf (2005) referred to native-speaker informants being able to judge collocations as to whether they think those given collocations are acceptable or not. Parkinson (2015) used the same method of native-speaker informants to determine whether collocations were appropriate or inappropriate. However, this approach does not show a link to collocation strength, as no concrete numbers or frequency can be given by native speakers, as the processes of familiarity are intuitive. As a result, this study only refers to this approach when collocations cannot be identified in corpora. It is considered a process of collocation recognition with no reference to collocation exclusivity.

With an acknowledgement to the limitations of each approach, the use of the two approaches in the way described are suitable for identifying collocations under investigation. Barnbrook (2013:164) claims that collocation does not have a specific rule or approach, because there are “no restrictions on the word-class, or the position of the collocates relative to the node, or even the relationship between the node and the collocate”. Barnbrook (2013) suggests that because of the nearly arbitrary way that words combine in natural contexts, there is no borderline when studying them. Each type of research and the collocations under investigation is different and unique, and has its own rule to approach.

2.5 Collocation in this study

The idiom principle, which describes collocation, as proposed by Sinclair (1991), is how natural language occurs and functions, and should be given attention. The “co-selection” of words is linked to the meaning (Sinclair 2004:133). As a major contributor to the field of collocation and corpus linguistics, Sinclair’s (1991) understanding of collocation shapes the theoretical framework of this study. Here, collocation is defined as: *the lexical co-occurrence of two words in a specific window where they convey a coherent meaning together*. The span window varies in research, but the five word span between two open word classes (i.e., verb-noun, noun-verb, adjective-noun, and noun-noun) limited in a clause has been chosen for this study (see Section 4.5.1). Sinclair (2003) believes that nouns by themselves are most of the time ambiguous in meaning and their collocates show the required meaning, thus nouns as the nodes to collocations which are being investigated in this study. Due to Saudi learners’ low-level of language proficiency, a study exploring collocations of two word combinations is appropriate. In addition, the three types of words under investigation i.e. verbs, adjectives and nouns, are basic elements for learners at their level when forming sentences and coherent meanings (See Section 4.5.1), especially given that adverbs rarely occurred in the data. The reason for choosing these types of collocates will be discussed further in the following chapters (Section 3.3) in relation to L2 learners' problematic production of collocations.

To identify a word combination as a collocation, or not, Sinclair (1991) and his followers such as Hunston (2002), affirm that collocates should have a greater than random chance of appearing within each other's window. Because of this, Stubbs (2007) adds that the study of collocation is correlated to corpora. An investigation of learners' production of collocation in this study will look into the exclusivity and strength of collocates, which will be achieved through the statistical association measurements available in corpus tools. In addition, while native speakers are sensitive to collocation in their L1, L2 learners face problems in using collocations, and as given above, the fixed collocation in particular. This is a collocation type which comes between the free combinations and the idiomatic combinations, where language users would put words together conveying a literal meaning, yet the choice of the words would make a difference. The collocations under investigation are those produced by Saudi learners in written texts reflecting collocations' literal meanings, indicating semantic relationships between the collocates to make a coherent meaning. These collocations will be identified and judged through the corpus-approach because the statistical measures in the corpus tools are able to provide evidence for the strength of collocations. By testing collocations produced by learners through using the LogDice association measurement and native speaker informants, this study contributes to the language learning research (LLR) literature. It was noted by Gablasova et al. (2017a) that most of the language learning research (LLR) on collocations relies on using the MI score, which reflects the "rare exclusivity", while the LogDice is not given such attention as the MI, and they argued in favour of the LogDice as it can be considered to be suitable as it addresses the "exclusivity" feature of collocation (ibid.:164). In case collocations do not exist in the corpora because of the limitations mentioned earlier, native speakers’ judgement will be used as in Nesselhauf (2003) and Parkinson (2015). Nonetheless, these judgements are purely intuitive and cannot measure association between collocates.

The next chapter will review the relevant studies on L2 learners and their production of collocation showing the gap this study aims to fill.

Chapter 3 L2 Learners and Collocation Research

3.1 Introduction

The problematic production of collocations by L2 learners has been investigated in various studies, using corpus approaches in a similar or different way to this study, depending on the scope of the research. This study applies the corpus-based approach to analyse learners’ written texts, and this part of the literature review will discuss a selection of relevant studies. I will look at a variety of applications, for example as in elicitation approaches, because it is important to show consistency or differences in findings. Furthermore, there is very little research using corpus approaches in analysing Arab learners' use of collocations. Therefore, this study aims to fill this gap of the application of corpus approaches to investigate the written production of English collocations by Arab learners from two proficiency levels.

In this chapter, studies on L2 learners’ production of collocations are presented chronologically in three groups. The first group of studies looks at how collocations produced by L2 learners are identified. The second group of studies explores problems in L2 learners’ production of collocations. These studies will also link the problematic production to different types of collocations. The third group reviews studies that link the process of language learning of L2 learners to their collocational production in relation to factors such as language proficiency level, language extension, L1 interference, and L2 exposure. The chapter then ends with an overview of the research findings on L2 learners’ uses of collocation and the gaps that remain to be covered in this and future research which has led to the research questions.

3.2 Research on identifying collocations produced by L2 learners

Research on L2 learners' production of collocations differ in how collocations are named, identified and judged. As described above, this study will focus on 'acceptable collocations' i.e. exclusive co-occurrence of collocates as identified in the corpus and the LogDice score, or occasionally where appropriate, by native speaker informants' judgement according to phraseology. Other studies employ terms such as 'true collocation' (Evert, 2008), 'appropriate collocation' (Siyanova and Schmitt, 2008), and 'acceptable collocation' (Kuo, 2009) using similar or different identification approaches as will be discussed. In this study non-acceptable or inappropriate collocations will be considered non-collocations and named as less idiomatic combinations.

In one of the earlier studies, Nesselhauf (2003) analysed German learners' written texts for their acceptable production of verb-noun collocations using dictionaries, the BNC and three native speakers' judgements. At the first stage, she distinguished between the combinations to identify whether they are free combinations, restricted collocations or idioms. To do this, she referred to two dictionaries: The Oxford Advanced Learner's Dictionary (OALD 2000) and the Collins COBUILD English Dictionary (CCED 1995) to decide whether a combination is restricted or not. When the verb is described in the dictionaries with more than a noun of a specific meaning, the combination is a free combination such as *need something/ someone,* whereas *fail a test/ an exam* is restricted by those nouns to indicate this specific meaning (ibid., 2003:228). However, sometimes the case is not clear and cannot be determined through the use of the dictionaries. In these cases, Nesselhauf, then combines the corpus approach with the phraseological approach by using a threshold of two or more occurrences in the BNC and two or three native speakers to judge whether or not this combination should be extracted for examination

While dictionaries include a considerable quantity of English vocabulary, the ability to identify restriction is not a very clear nor straightforward procedure. It depends on the researcher's judgement as well as the native speakers'. As for the acceptability criteria, she used the same dictionaries and in addition the Oxford Dictionary of Current Idiomatic English (Benson et al., 1997) or the occurrence of five-citations as the threshold score in the BNC. When a collocation could not be identified as acceptable, it would be referred to two native speakers (one American and one British) to judge it as either correct, wrong, unsure, or acceptable, this last term being used when it is less certain than correct, but it cannot be classified as wrong or unsure. The informants were requested to make a correction if the collocation was judged as unsure or wrong. If the judgement by the two informants was different, a further two native speaker informants were asked for their opinions. Eventually, a classification scale was created based on the informants' responses to identify collocations as clearly acceptable, largely acceptable, unclear, largely unacceptable and clearly unacceptable collocations.

According to current definitions of type of methodology, Nesselhauf's could be considered almost purely qualitative, starting from the manual analysis by observing the syntactic patterns of combinations extracted, then classifying them into free combinations and restricted collocations, and finally identifying them to as whether they are acceptable or not. Nesselhauf (2003:229) admits that the procedure she followed to distinguish between the free combinations and restricted collocations was "subjective to some degree". Even though she described the three-step procedure systematically, it remains hard to follow and challenging to apply, especially when employing the phraseological approach to identify and then classify collocations. When she used the BNC, she referred to the number of citations for a collocation with no statistical measurement for the association of the collocates. Therefore, she can be criticised for this omission as this methodology does not provide enough evidence to judge the restrictedness of collocations. Her 2003 study is also problematic because it uses native-speaker judgements not only for acceptable/ non-acceptable collocations, but also to interpret their responses to judge restricted collocations according to a continuum scale. Even though this study follows Nesselhauf's in some aspects such as using the corpus and phraseological approaches, it acknowledges the weaknesses in her studies and thus aims to address the gaps and limitations in her methods as will be described in the methodology chapter.

Siyanova and Schmitt (2008) investigated the production of another type of collocation the adjacent adjective-noun collocations produced by Russian learners. They analysed essays from the International Corpus of Learner English (ICLE) for learners and the Louvain Corpus of Native English Essays (LOCNESS) to compare learners' use to native speakers'. After the manual extraction of the collocations, they referred to the BNC to obtain the frequency and MI score with which to judge collocation acceptability. Siyanova and Schmitt's methods were both detailed and rigorous; firstly, they used the bands of frequencies as: 0 (meaning failed to appear), 1-5, 6-20, 21-100, and >100 occurrences (ibid., 2008:435). When a collocation fails to appear, it is considered atypical, and when it is in the band of 1-5, it is considered infrequent. Frequent collocations, or typical collocations, would fall in the bands above 6. Secondly, they used the MI score of 3 or more with the criterion of frequent collocations of 6 or more occurrences to judge their acceptability. The procedure used by Siyanova and Schmitt (2008) in identifying acceptable collocations appears to be more reliable than Nesselhauf's methodology as it combines the attestation in the corpus of a certain number of citations to show conventional use of the language and statistics that reflect on restrictedness quantitatively. However, they did not feel it necessary for a further reference to identify the collocations produced by the learners qualitatively such as the judgements of native speakers. However, the phraseological approach as discussed in Chapter Two, Section 2.4.2 has a number of advantages in the study of collocation, particularly in identifying their acceptability not fixedness, as will be described in the following research.

Kuo (2009), who also analysed learners' writing using the BNC and native speaker informants, investigated the production of verb-noun and adjective-noun collocations by Taiwanese EFL learners. In the initial analysis, Kuo referred to the BNC, but the study is limited by not including the number of citations or frequency used to identify collocations and consider them acceptable. She just referred to a 'high frequency' as a well-formed collocation, which is not specific enough for this level of research. In order to check learners' problematic production of collocations, she further employed an online website, http://candle.cs.nthu.edu.tw/vntango/, based on Taiwanese collocational errors. In the first stage, she referred to the BNC and the collocation checker to identify collocations. In the second stage, she followed Nesselhauf (2003) by including the phraseological approach by assigning two native-speaker informants in order to check the collocations identified in the first stage. The native speaker informants would categorise a collocation as correct or erroneous. They, for example, agreed on collocations such as *black list* and *answer a question* and disagreed on others such as *pay time* meaning *spend time*, and *middle exam* meaning *midterm*. Kuo employs the phraseological approach to judge a collocation's acceptability without indicating its restrictedness. Unlike Nesselhauf, Kuo asked native speaker informants to judge collocations produced by learners and which appeared in the BNC, to decide whether they are acceptable or not. No further interpretations for fixedness were requested as native speakers' intuition is a poor indication of collocation strength. Even though this application of the phraseological approach was advantageous in Kuo's study, the reference to the number of citations or collocation frequency in corpus can be seen as a drawback. This is because frequency alone does not provide enough information about collocation for further interpretations, such as in the use of association measurements, which can give an in depth reflection of learners' production of collocation.

Laufer and Waldman (2011) focused on the accurate production of verb-noun collocations by Hebrew learners from three proficiency levels. The analysed essays were taken from the Israeli Learner Corpus of Written English (ILCoWE) in comparison to essays from Louvain Corpus of Native English Essays (LOCNESS). First, they referred to one of two dictionaries i.e. the BBI Dictionary of Word Combinations (Benson et al., 1997), or the LTP Dictionary of Selected Collocations (Hill and Morgan, 1997). In addition to one of those dictionaries, they referred to the BNC, and a native speaker to identify any incorrect collocations such as *use a chance* and *inflict arguments* (Laufer and Waldman, 2011:658). Laufer and Waldman's methods are similar to those used previously by Nesselhauf and Kuo through combining the corpus and phraseological approaches. However, Laufer and Waldman were different in that they used only one native speaker informant. A single native speaker informant is unlikely to be enough to satisfactorily judge collocations produced by learners, especially when considering acceptability. The informant's judgment can be subjective depending on their knowledge, use and experience of the language. Moreover, they referred to one dictionary only, unlike Nesselhauf, which limited the scope of this research. Different studies have used and employed the same approaches differently according to the research objectives of each study even though some limitations still exist and can be identified.

Parkinson (2015) followed the same combined approaches; however, she further included the Corpus of Contemporary American English (COCA). Her study was also different because of the type of collocations she investigated. She studied the use of noun-

noun collocations produced by EFL; Mandarin and Spanish, and ESL; Tswana, learners. She followed Nesselhauf in using the threshold score in the BNC of five citations, and used 25 citations for COCA as it is five times larger than the BNC. When collocations reached the threshold scores in both the BNC and COCA with an MI score higher than 3, they were considered frequent collocations. If they scored less than five and 25 citations in the BNC and COCA respectively, they were considered infrequent collocations even if they had an MI score higher than 3. When they scored an MI score less than 3, they were considered phrases. If collocations could not be judged according to their citations in corpora, they were referred to 11 native-speaker informants. When eight agreed on a collocation, it was considered appropriate e.g. *cartoon language* and *tourism student*, and when four or more disagreed, it was considered inappropriate. Parkinson's methods combined the corpus and phraseological approaches as in the previous studies described; however, she did not employ the use of dictionaries at any stage. There was no need in her study because she identified collocations through using two steps in corpora i.e. the number of citations and the association measures. She overcame a number of the limitations of earlier studies in identifying collocations produced by learners, especially qualitative limitations, by employing statistical measurements such as the MI score, that are more reliable than subjective judgements. Parkinson further referred to native speaker informants, in a similar manner to Nesselhauf, but with a greater number of informants and, as discussed previously only in judging acceptability as native speaker informants' judgement for determining collocation restrictedness is unreliable.

Fernández and Schmitt's (2015) study mainly discussed the relation of association measurements to collocations produced by Spanish learners, and which measurement could be the most appropriate and reflective of their collocation knowledge. They tested learners' production of collocation in a gap filling task, where target collocations e.g.

exploit resources and *clockwise direction*, were chosen from COCA according to the following criteria i.e. target collocations should be frequent lexical collocations of 2-gram only, agreed on as natural English by native speaker informants, dispersed in raw frequency, *t-score* and MI score ordering, and not having direct Spanish equivalents. The majority of the participants, about 70%, had scores between 21- 40 out of 50 in the given task, whereas half of the remaining participants had less than 14, and the other half had more than 40. Fernández and Schmitt further showed that learners' production of correct collocation correlated in COCA with raw frequency and *t-score*, respectively, by 20% and 17%, while it did not show any significant correlation with the MI score. Even though there is a difference between the percentages of raw frequency and *t-score* results, it is slight, which indicates the similarity in the two measurements in identifying high frequency collocations. However, to understand better this relationship between the three tests and learners' production of collocation, Fernández and Schmitt divided learners' scores into three groups: low level (1-20, *N*=21), medium level (25-30, *N*=40) and high level (40+, *N*=14). They found that high frequency collocations are not produced in greater numbers in any of the three groups, meaning that, different scores did not show a direct relation to more or less production of high frequency collocations. They suggested that learners' production of collocation lies between frequent and infrequent collocations, meaning that most of the collocations produced by learners were in the midrange of frequency. A further association measurement such as the LogDice, that tests these two criteria of collocations i.e. frequency and fixedness, together should be applied to test learners' production of collocation. Learners do not only need to learn functional language i.e. high frequency collocations, but they also have to acquire language needed for communication such as exclusive collocations.

Each of the aforementioned studies had its own method of identifying collocations produced by learners. However, each study is similar in that they have two or more ways of identifying collocations. In this research, which mainly uses the corpus approach, the BNC will be referred to because the participants are taught British English. Similarly to Nesselhauf and Parkinson, this study will consider collocations with five or more co-occurrences in the BNC as a first step of identification. In order to attend to exclusivity and identify acceptable collocations and their fixedness, the second step will use one of the association measurements relying on the rationale explained previously in Chapter Two, Section 2.4.1. Evidence arguing in favour of this decision comes from Fernández and Schmitt's study, which is to employ the LogDice measurement in this study (See Section 2.4.1). Finally, as the phraseological approach is viewed as the least appropriate for reliability to identify collocations' fixedness, it is only applied as a final approach in this study when a collocation is not found in the BNC.

3.3 L2 learners' problematic production of collocations

The relevant studies in this section will review the problematic types of collocations in the uses of L2 learners. The review will include studies using corpora to analyse learners' written texts, such as this study, as well as task-based studies. In the studies discussed here L2 learners generally showed similar problems in producing collocations though they were always better in recognition. Thus, the review will be able to reflect on the gap in the research into L2 learners' production of collocations, especially as this issue was most of the time associated with certain types of collocations learners produced.

Using corpus approaches to investigate the production of collocation, Granger (1998) conducted a comparative study between the native and non-native speakers' production of adjectives and amplifiers i.e. adverbs. She conducted the study by examining essays from different corpora: the International Corpus of Learners English (ICLE) for learners, the Louvain corpus, the International Corpus of English (ICE), and Belles Letters, a category of the Lancaster-Oslo/Bergen (LOB) corpus for native speakers. She used the TACT software to extract collocations, and referred to their raw frequencies in comparing learners' overuse of amplifiers such as *totally*, and underuse of amplifiers such as *highly* to native speakers' use. To Granger, this overuse and underuse of those amplifiers in particular can be explained by the similarity of their uses between English and French (the learners' L1). In a subsequent task, where she wanted to test significant collocations, she asked two groups – L2 learners and native speakers - to match amplifiers i.e. adverbs with adjectives, to create accurate collocations. She found that learners formed more combinations than native speakers, and repeatedly. Yet, not all of the combinations learners formed were acceptable or, using her terminology, significant collocations, such as *irretrievably different*. Learners produced creative combinations such as *shamelessly exploited*. Granger did not indicate what instrument she used to evaluate collocations produced by learners such as a corpus or a dictionary. She only referred to the comparison of learners' use to those of native speakers. Granger notes that native speakers used collocations such as *acutely aware*, and sometimes they created combinations that learners did not produce at all such as *astonishingly short*. Granger's study suggests that learners can produce and maybe create a large number of combinations, but they often fail to produce them accurately as in the task. Granger argued that learners have a tendency to rely on what was described by Sinclair (1991) as the open choice principle, with grammatical restraints rather than coherent meaning, opposing the idiom principle that mostly reflects the use of native speakers. However, Granger's findings highlight the problematic production of collocation by L2 learners qualitatively by merely comparing their use to native speakers' even though the study lacks a statistical reference. This is especially noteworthy as she investigates the production of significant adverb-adjective collocations in the task, the production of which may be poorly judged by the subjectivity of the native speaker.

Nesselhauf (2004) examined learners' production of verb-noun collocations by analysing German university students' essays taken from the ICLE. Although she referred to dictionaries and the BNC, as described in the previous section, her investigation was purely qualitative like Granger's. Nesselhauf investigated free combinations and restricted collocations, and found that this type of collocation i.e. verb-noun in its restricted form, is challenging and an area in which learners tend to make errors. One-third of their total production comprised non-acceptable collocations. She reported *reach an aim* as an example of these erroneous collocations that had no occurrence in the BNC, whereas the collocation should be *reach a goal.* When *reach a goal* is measured using the LogDice, it has a score of 6.8 showing a restricted collocation. Even though, by using her qualitative interpretations, Nesselhauf was able to give the correct collocation for learners' faulty production of restricted collocations by referring to statistical association measurements. This is an advantage that would create a more reliable and systematic procedure for researching collocations.

In addition to the verb-noun collocations, Martyńska (2004) examined other types of collocations produced by Polish learners. He investigated noun-verb, adjective-noun, and adverb-adjective collocations in a task-based study. The overall results of the tasks examining these four types of collocations were not significantly different showing that they were able to produce successfully almost half of the collocations from each of the four types. However, the learners tended to complete recognition tasks, such as multiple-choice tasks, better and with greater accuracy than production tasks, such as giving and completing collocations. Martyńska indicated that verb-noun collocations e.g. *make an appointment* and adjective-noun collocations e.g. *high-heeled shoes* were easier for Polish

EFL-learners than adverb-adjective collocations e.g. *totally exhausted* and noun-verb collocations e.g. *car breaks down*. It is still the case indicated by other researchers of learners' difficulty collocational production more than recognition, yet Martyńska highlights the other types with which learners experienced difficulty with. This finding is in contrast to Chiu and Hsu's (2008) study, which will be discussed in detail in Section 3.4.1 below, who found that only collocations including adverbs are problematic for Taiwanese learners, and that noun-verb collocation production was similar to producing verb-noun and adjective-noun collocations.

Although these task-based studies are effective in investigating various types of collocations produced by learners, they are limited by the kind of collocation chosen by the researcher. Analysis of learners' written texts can still cover the investigation of the different types of collocations; however, with a greater range of collocations possibly produced by learners.

Siyanova and Schmitt (2008) tested the acceptable production and recognition of the adjective-noun type of collocation. After examining learners' production in comparison with the BNC, they found that less than half – about 45 percent - of the produced collocations were acceptable whereas the other half were between atypical and infrequent collocations. These findings from a study described in Section 3.2 above which were based on frequency and a statistical measurement, suggest that L2 learners' production of adjective-noun collocations are problematic. The EFL and ESL learners under investigation were further found to have difficulty in perceiving this same type of collocation. Siyanova and Schmitt compared their learners to native speakers by asking them to rate a list of adjective-noun collocations based on acceptability on a six-point scale between very uncommon, uncommon, fairly uncommon, fairly common, common, and very common. The list included frequent and infrequent collocations, where the frequent collocations should have 20 or more occurrences in the BNC with an association measure on MI score of 3 or more. They also had to appear in the BBI Combinatory Dictionary of English, and the Oxford Collocations Dictionary (OCD). The infrequent collocations did not appear in any of the references used. The constituent parts of the infrequent collocations should be meaningful and grammatically correct. They give as an example, the collocation *law-obedient people*. As *law* and *obedient* separately are frequent, they are infrequent as a collocation *law obedient*. However, as a collocation *law obedient*, is grammatically correct and meaningful. Siyanova and Schmitt's results indicate that both EFL and ESL learners had less reliable intuitions when rating collocations, especially about infrequent collocations, and that native speakers' judgements were consistent with the BNC and their intuitions toward infrequent collocations were successful. This finding is not surprising when comparing the rating of collocations by learners to those of native speakers. The same results were attained by Granger (1998) when investigating the adverb-adjective collocations.

In an Arabic context, Shehata (2008) investigated the production of verb-noun and adjective-noun collocations produced by university students using a task-based approach. She compared their production of collocations by employing three tasks, two of which tested learners' collocational production i.e. gap-filling, whereas the third tested learners' collocational recognition i.e. appropriateness judgement. Learners' collocational recognition was better than their production, specifically, with adjective-noun collocations which they found less difficult to recognise than verb-noun collocations. In contrast, the results of Shehata's quantitative analysis show that learners scored higher when producing verb-noun collocations e.g. *have an effect* than when producing adjective-noun collocations e.g. *golden age*. This suggests that the problem is not only related to whether collocations are produced or recognised, but also it corresponds to the type of collocations, something that was observed by Martyńska (2004). When looking at the two examples of the collocations in the BNC, *have an effect* appears 7,782 times scoring 7.5 on the LogDice while *golden age* appears 203 times scoring 8.0 on the LogDice. These results in addition to the work done by Shehata suggest that there is a problem related to types of collocations and that this problem may lay between the focus on high frequency collocations and exclusive collocations. Thus, examining learners' production of collocations requires the application of association measurements.

Kuo (2009), who studied verb-noun and adjective-noun collocations like Shehata, but mainly focused on production, had different results. In analysing Taiwanese EFL learners' written production of those two types of collocations, and by referring to the BNC according to high frequency, Kuo found that the learners produced a relatively high number of acceptable collocations. Around 82 percent of the collocations produced were acceptable and only around 18 percent were not. She further indicated that her learners produced adjective-noun collocations more accurately than verb-noun collocations. Her learners produced accurate collocations like *keep healthy* and *deaf ear* and erroneous collocations included *promote appetite* and *serious promise*, which Kuo interpreted to mean *increase appetite* and *firm promise*. These findings of her learners' ability to produce a high number of appropriate collocations were different to studies that analysed learners' written texts using corpora such as Nesselhauf and Siyanova and Schmitt, which can be related to the investigation of high frequency collocations in Kuo's compared to restricted collocations in the other studies.

Brashi's (2009) study is another task-based research project that investigated Arab learners' production and recognition of collocations, yet focused only on the verb-noun type. He gave learners multiple-choice and gap-filling tasks, where they were asked to choose a suitable verb for given nouns. In the first task, learners had to choose from a given list while in the second task, learners had to provide their own verbs. His findings were consistent with Shehata's, highlighting the problematic issue of collocational production. Learners gave more correct responses in the multiple-choice task, which tested recognition with a 79% success rate, than in the gap-fill task, which tested production, with only a 38% success rate. Learners took twice as long on the second task than the first, which may suggest difficulty in producing collocations. The example collocations that Brashi used for this task were taken from the Collins COBUILD English Collocation Dictionary (Sinclair et al., 1995), and results were examined using the same reference plus three other dictionaries i.e. The BBI Combinatory Dictionary of English (Benson et al., 1986), the Dictionary of Selected Collocations (Hill and Lewis, 1997), A Dictionary of English Collocations Based on the Brown Corpus (Kjellmer, 1994), and software i.e. WordPilot 2000 (Milton, 2000). Brashi's study is similar to Martyńska, in that they both limit the investigation to a controlled context rather than examining learners' free production of acceptable collocations.

Fan (2009) investigated Chinese ESL learners' production of adjective-noun collocations that were extracted from a compiled corpus using ConcApp. This corpus included a writing task produced by Chinese and native-speaker learners of 300-word essays written to describe one picture. Fan compared them with each other to test collocation overuse and underuse, according to frequency, produced in the two groups' writing. She found that learners did not only produce fewer accurate collocations, which is expected when compared to native speakers' use, but also lacked the ability to produce collocations with adjectives already familiar to learners, such as *large*, *big,* and *deep*. Fan's findings with ESL learners are consistent with those of Siyanova and Schmitt, who also noted that native speakers are likely to produce types of informal collocations that are unfamiliar to L2 learners. Fan's native-speaker subjects used informal collocations, which had not been acquired by learners, such as *fat round face*, *chubby face*, and *roundish face*. Learners instead produced collocations such as *round face*, *circle face* and *baby face*. Nevertheless, Fan's method has some limitations as there was not a valid reference by which to judge the collocations produced by her subjects. She merely compared the two groups' production of collocations identified within a 24 character span before and after the node in a sentence. Even though one of the two groups were native speakers, it would be more reliable to use established reference tools to identify the collocations produced. For example, by referring collocations produced by learners to the BNC in Sketch Engine, there are no citations for *circle face,* while *baby face* and *round face* score 5.0 and 5.3 on LogDice respectively. However, the native speakers' collocations *chubby face* scores 3.6 on LogDice whereas *roundish face* is cited once only in the BNC. The collocation *fat round face* did not co-occur as this whole creative combination, but it appears as *fat face* alone scoring 5.6 on LogDice. Other creative combinations like *fat cheery face, fat cheerful face* and *fat baby face* also did not occur as a whole. An additional means of identification would improve the validity of the methodology used.

Durrant and Schmitt (2009) also studied adjective-noun collocations, where adjectives can include noun modifiers e.g. *human rights*, in EAP academic and argumentative essay samples produced by both learners (from the ICLE) and native speakers (from their MA assignments and Prospect magazine). Durrant and Schmitt used two tests for strong collocations; the *t-score* for high frequency collocations and the MI score for exclusive collocations. After the manual analysis, they compared the two groups using the BNC, and found that learners produced as many collocations of high *t-score* frequency as native speakers. However, the learners tended to produce the same collocations already produced by them regularly, which may reflect their limited range of collocational knowledge. This suggests that learners tend to produce collocations and language in the way that is familiar to them, which makes them feel secure, while native speakers are confident to show creativity and uniqueness in their combinations. Durrant and Schmitt also found that the ICLE learners relied on familiar high-frequency collocations, while native speakers tended to produce low-frequency collocations such as *densely populated*, which learners underused. This supports the argument raised in the previous chapter (see Section 2.4.1) for using the LogDice association measurement in this study because of the frequency of collocations learners are most likely rely on. Thus, using LogDice solves the 'rare exclusivity' of low-frequency collocations as initially proposed by Gablasova et al. (2017a) and overcomes the lack of exclusivity of the t-score, which merely calculates the high frequency combinations.

Besides these two types of collocations i.e. verb-noun and adjective-noun, this study also examines learners' production of the noun-noun collocations, which is not as widely addressed in the collocation literature as the other two types. Besides Durrant and Schmitt (2009) who investigated noun modifiers in adjective-noun collocations, Parkinson's (2015) research is one of the few studies that has investigated the production of noun-noun collocations. She focused on those collocations produced by EFL and ESL learners in argumentative essays taken from three sub-corpora of the ICLE corpus. She compared three groups of learners and found that their production seemed to be related to their L2 environment. After identifying collocations, as described in the previous section (3.2), Parkinson found that Tswana learners, who were ESL learners, produced more appropriate collocations than Spanish learners, who were EFL. This suggests that collocation is more easily acquired in naturalistic learning environments. She also found that learners whose L1 allowed the noun-noun phrase system, such as in Mandarin, were better able to produce this type of collocation than the other two groups (Parkinson, 2015:111). Learners produced some inappropriate collocations such as *crime doings* and *window closet* rather than *criminal acts* and *shop window*. Therefore, this type of non-noun collocation is also found to be challenging and problematic for learners, especially when using pre-modifying nouns, and needs further exploration.

L2 learners can face problems with the different types of acceptable lexical and fixed collocations (Martyńska, 2004), but this study will focus on those types of collocations reviewed in the relevant studies i.e. verb-noun, noun-verb, adjective-noun and noun-noun collocations. Regardless of L2 learners' contexts and the approach used for the investigation, relevant research shows that the learners do face difficulties in collocational production more than their recognition. This was also highlighted in studies concerning Arab learners such as Brashi (2009), who focused only on the production of verb-noun collocation, and there has been a continuous and growing interest in this type of collocation. Through analysing learners' written texts by referring to corpora, studies like Nesselhauf (2004) have been able to highlight the problematic production of verb-noun collocations, specifically the restricted combinations. Other researchers such as Siyanova and Schmitt, (2008); Fan, (2009) and Durrant and Schmitt, (2009) studied the frequent and infrequent use of adjective-noun collocations in comparison to those of native speakers, and they found that this type of collocation is problematic. Some researchers have studied a combination of the two types i.e. verb-noun and adjective-noun collocations to investigate if one is more challenging than the other such as Shehata (2008) and Kuo (2009). However, Shehata's results examining Arab learners showed a better production of the verb-noun collocations whereas Kuo's, which analysed EFL learners' texts using corpora, showed that adjective-noun collocations were more accurately used. Noun-verb collocations were mainly investigated in task-based studies such as in Martyńska (2004) with Polish learners and Chiu and Hsu (2008) with Taiwanese learners, whose results were not similar. Whether this difference in results was due to learners' different L1s or the nature of the investigation i.e. written vs. spoken and tested collocations, noun-verb collocations have not been given much attention. Finally, the use of noun-noun collocations was addressed by Parkinson, (2015), who suggested that its problematic production is linked to learners' L1 and L2 exposure. Even though this current study is not investigating in depth the issues of L1 interference and L2 exposure, it still relates to Parkinson's investigation into noun-noun collocations.

Thus, there is a research gap in analysing the production of these types of acceptable collocations and their fixedness (verb-noun, noun-verb, adjective-noun and noun-noun) produced in Arab learners' written texts, and specifically by applying corpus approaches. While previous research has shown that the written production of certain types of collocations is problematic, none was concerned with EFL learners whose L1 is Arabic. It is still important to investigate the written production of those learners as it is considered less controlled than the task-based studies, especially as previous research has also shown that different results, findings and interpretations can be achieved when using different approaches. A further importance of this study is that it examines collocation in the written production of two levels of EFL learners. This investigation refers to corpus tools to examine the production of acceptable collocations according to their level of fixedness. There is a need to study these different types of collocations including verbs, adjectives and nouns all together, as they work as fundamental structure of a meaningful sentence. Moreover, a comparison of this type i.e. production of collocation between two different levels of learners, is also important, rather than comparing learners' use to that of native speakers'. There have been many research studies conducted to analyse learners' production of collocation from different L1s or contexts, but very few studies have been undertaken on learners from the same L1 but with different proficiency levels as the next section will show.

3.4 L2 learners' production of collocations and language learning

While collocations can be learnt and acquired by learners, they are still difficult and challenging to produce whatever their level of proficiency, similarity with their L1, or experience of the L2 environment. This may cause collocational errors which can lead to communication problems. Bahns and Eldaw (1993) argue that many students use paraphrasing instead of recalling correct collocations. Hill (1999) gives the example of *his disability is forever,* which a learner might write instead of *he has a permanent disability*, as illustrating how a learner fails to use words he/she knows to make a coherent sentence, and therefore in trying to communicate the meaning, produces errors instead. Even though the learner may know the two words *permanent* and *disability*, a lack of collocational knowledge of adjective-noun collocations can cause such an error.

Some studies such as Li (2005) link errors in the production of collocations to lack of grammatical knowledge as well. In her study of EFL learners' collocational recognition and production, she found that Taiwanese learners' collocational errors were due to a lack of knowledge of grammar rules. She reports that most errors occurred in collocations which included prepositions, such as *sympathy on them* and *at the summer vacation*. This is similar to the case reported by Ibrahim et al. (2012), which suggested that Persian learners made more grammatical errors as in colligation, than lexical errors as in collocation, due to the difficulty of acquiring and using prepositions. However, this could be a special case because the Persian equivalents of collocations include collocations such as *break promise* and *do homework*. However, Fan's (2009) study, mentioned in the previous section (3.3), suggests that the inaccurate production of collocations by learners is caused by both the lack of L2 grammatical and lexical knowledge. This implies that sometimes there is no borderline between grammatical and lexical errors, and that it is rather a combined process and results as learners unintentionally switch between the open choice and idiom principles. EFL learners' knowledge of collocations is not always the reason behind their accurate or inaccurate collocation performance but could also be related to their overall language knowledge. It is important to refer to these studies because they show that this lack of grammar can affect the collocation as a whole, and as a result, a whole coherent meaning, which will be discussed later in the Discussion Chapter.

Other studies, which concern purely lexical issues, like this study, address the problems of learners' collocational production with no reference to the role of L2 grammatical lack. Such studies have mainly investigated four issues related to learners' language proficiency; the use of analogy or extension; the interference of L1; and the ignorance of and/or exposure to L2. Because the use of extension is linked to the language level of learners, these two issues will be considered in the following section, especially, as the learners' language proficiency is a significant contributing factor investigated in this study. L1 interference and L2 exposure are also relevant topics in language learning and the collocational production, but are not covered in depth in this study. Still, reviewing them is essential in discussing the study findings.

3.4.1 Learners' Proficiency Level

While learners of lower and higher levels may face problems in collocational usage, many studies argue that errors in collocational production are fewer in higher-level learners. Howarth (1998) and Kuo (2009) suggest that higher-level learners may produce a different type of error, which happens when learners use extensions of language to produce collocations, but they fail to do so accurately. This section discusses the effect of learners' proficiency levels on their collocational production, and whether it leads to a better performance as their level improves or to a further form of collocational errors, such as in the use of extension.

Abdul-Fattah and Zughoul (2001) investigated the production of collocations between two levels of learners i.e. postgraduate and undergraduate students. In their task-based study, which was primarily concerned with the use of the verb *broke* and its Arabic equivalent *Kasara* – كسر, they found differences in learners' responses related to their proficiency levels. In the multiple-choice task, lower-level learners averaged 42.77 percent correct, compared to higher-level learners, who answered 57.57 percent correctly. Even though some of the collocations were not totally transparent and would have an idiomatic expression such as *he broke the prevailing silence* and *some workers broke the strike*, still learners chose correctly. Most of the incorrect items were actually irrelevant items due to learners' lack of concentration while taking the tasks as Abdul-Fattah and Zughoul mention. The scores of the two groups of learners were also different in the translation task. The higher-level group scored better than the lower-level, with scores of 18.9 percent and 14.5 percent, respectively although the percentage of correct responses dropped drastically across the two levels when compared with the previous task results. Many of the correct responses were due to L1 equivalents of English collocations such as: *he broke his will*. Their findings suggest that there is a difference in the production of collocations by Arab learners of different proficiency levels, and that their L1 does not have a positive impact on their collocation production. Even though the difference in the two groups' results is not significant, their language proficiency levels showed a greater impact than their shared L1. This further suggests that collocation is not just a problem of vocabulary that might be considered to be advanced but is relevant for all levels of learners. Still, Abdul-Fattah and Zughoul's findings are limited to not only one type of collocation, but also to only collocations associated with the verb *broke*.

Chiu and Hsu (2008) examined the relationship firstly between learners' language speaking proficiency and their collocational production, and secondly learners' language speaking proficiency and their collocational knowledge. They found no relationship between production of collocations and language speaking proficiency whereas there is a correlated relationship between language speaking proficiency and learners' collocational knowledge. They tested Taiwanese EFL learners in three tasks examining the following types of collocations: verb-noun, noun-verb, adjective-noun, adverb-adjective, and verb-adverb collocations. Learners were tested for their collocational knowledge by a gap fill task, for their collocational production by answering questions about a film after reading a list of collocations, and finally for their proficiency level by measuring the average of two speaking tests. To rate collocations produced by learners, they referred to The BBI Dictionary of English Word Combinations (Benson et al., 1997). If the collocation was not found, they referred to either of the two online corpora: Simple Search of British National Corpus and VLC Web Concordancer, or one of the two native speaker informants in case of English variety occurrences. The quantitative analysis showed that there was no significant relationship between learners' collocational production and collocational knowledge nor with language proficiency, while there was a significant relationship between their collocational knowledge and language proficiency. This suggests that learners' language proficiency relates to knowledge about collocation, i.e. recognition, but does not necessarily positively affect their production of collocation. It is still the case as was found by Abdul-Fattah and Zughoul's study (2001), which shows that the learners' proficiency level does not play a major role in the successful production of collocations, even in the production of different types of collocations as found by Chiu and Hsu (2008). However, this result was influenced by the fact that Chiu and Hsu's learners were tested on their collocational production by being given a list of collocations, which is interfering with their performance. Also, their study only correlated the collocational production to learners' spoken English. Their study is one of the very few that investigated noun-verb collocations. While Martyńska (2004), as discussed in Section 3.3, noted that noun-verb collocations were more challenging than verb-noun and adjective-noun collocations in learners' production, Chiu and Hsu found this type of collocation is almost equally challenging for learners' to produce as verb-noun and adjective-noun collocations.

The Laufer and Waldman (2011) study, mentioned earlier in Section 3.2, also investigated the relationship between the learners' proficiency level and their production of verb-noun collocations. In a study similar to this one, using corpus approaches to analyse learners' production of collocations, Laufer and Waldman investigated Hebrew learners at three language proficiency levels (basic, intermediate, and advanced). They found that the production of verb-noun collocations is problematic for Hebrew learners from all the three levels. However, by comparing the frequency of collocations produced by Hebrew learners, they found progressive differences in the production of acceptable collocations between the three levels. Advanced learners accurately produced around 6.2 percent of verb-noun collocations, while intermediate and basic learners produced respectively around 5.3 percent and 4.3 percent. Yet, collocational errors were found at all three levels; basic and intermediate learners had a similar percentage of erroneous collocations, with 33.3 percent and 33.6 percent respectively, while advanced learners had a smaller percentage of errors, with 31.9 percent. Laufer and Waldman found that Hebrew learners produced inaccurate collocations, such as *use a chance, learn children* and *do a decision* which were caused by L1 interference as well as because of the different proficiency levels. Although Laufer and Waldman's study showed differences in the results of learners' production of collocations across different levels, the gap between these levels is slight and insignificant. The results are also similar to previously discussed findings (see Section 3.2). Even though Laufer and Waldman's study contributed by investigating the production of collocations by different levels of learners, it addressed only one type of collocations which was judged according to frequency.

Huat (2012) also investigated the production of verb-noun collocations among three different proficiency levels like Laufer and Waldman (2011); however, Huat focused on the qualitative development of this collocational production between the levels. He investigated the production of verb-noun collocations by Malaysian EFL learners in essays based on pictures he gave them. After analysing the written samples of the three levels: beginner, intermediate, and upper intermediate, he found that there is a developmental sequence in the production of verb-noun collocations. Additionally, the two higher-level groups' production of collocations is similar in frequency and pattern. From a proposed eight classifications of verb-noun collocations, intermediate and upper intermediate levels shared six of them i.e. collocations including *hear, help, pick, pluck, save* and *thank* whereas beginners included collocations with *fish* and *shout,* and did not include *hear* and *pluck.* Beginners tended to use verbs with more general meanings, such as *pick some flowers,* intermediate and upper intermediate tended to produce collocations, such as *pluck some flowers,* which has a very specific meaning. According to the association measurement scores, the first collocation, *pick some flowers,* scores 6.4 and 6.9 on MI score and LogDice respectively while, *pluck some flowers,* scores 7.6 and 5.1 on MI score and LogDice respectively. While both collocations are considered fixed and restricted, the MI score indicates rare and very exclusive collocations with precise or specialized meanings, the LogDice score shows frequency as well as exclusivity. Huat's findings suggest that lower level learners relied more on frequent and strong collocations, as the LogDice indicates, whereas the higher-level learners were able to produce low-frequency and strong collocations, as the MI score indicates. Nonetheless, Huat did not indicate whether the higher-level learners produced the collocation *pluck some flowers* accurately in the context.

Similar to Huat, Ebrahimi-Bazzaz et al. (2015) explored the production of verb-noun collocations in a story-writing task of four different levels of university students: freshman, sophomore, junior and senior. Iranian learners were asked to write six stories based on six pictures in 60 minutes. Each picture came with a set of three nouns which learners could use to produce the verb-noun collocations. The quantitative results showed a progression between the four levels, and that the production of the verb-noun collocations developed systematically with the learners' levels from one year to the next. One of the examples they give that lower-level learners produced is the collocation *shoot the ball*, which developed into the accurate collocation produced by higher-level learners, *kick the ball*. Ebrahimi-Bazzaz et al.’s study was limited to the verb-noun collocations produced with the nouns given, which were mainly chosen for their differences between English and learners’ L1. Based on the discussion in the previous section (3.4), the problematic production of collocation can be different not just according to the different levels of language proficiency but also according to the types of collocations.

Paquot (2017) conducted a study on French EFL learners by investigating a compilation of learners’ essays produced as requirements in their university courses. The essays were written between 2009 and 2013 by upper-intermediate (B1+), advanced (C1), and very advanced (C2) learners according to CEFR. She extracted collocations according to their grammatical patterns: adjectival modifiers i.e. adjective-noun, adverbial modifiers i.e. adverb-adjective/ adverb-adverb/ adverb-verb, and verb-object i.e. verb-noun. Unlike most research including this current study, Paquot did not refer to the BNC or COCA, he referred to a more specialized corpus L2 Research Corpus (L2RC), which was more suitable to his learners' written texts. His findings when examining the written production of those three levels for the aforementioned types of collocations showed that differences were not statistically significant. The mean of MI scores across the three levels increased systematically as 11%, 14% and 15% with no big differences among them. However, differences were more obvious in learners' production of the different types of collocations investigated across the three levels. Adjective-noun collocations showed a significant difference between B2 and C2, but not between adjacent levels such as B2 and C1 or C1 and C2. Conversely, challenges in the verb-noun type across the three levels of learners were more obvious between them, and especially between the very advanced learners (C2) compared to the other two lower levels (B2 and C1). Paquot's results suggest that the issue with learners' production of collocations is more qualitative rather than quantitative. Therefore, this study aims to investigate this issue of learners' written production of different types of collocation in relation to their proficiency levels closely through a qualitative analysis of their written texts.

Farooqui's (2016) finding suggests that advanced level EAP learners can produce collocations similar to native speakers. She explored the production of collocations by advanced EAP learners in three UK universities. Using frequency, she compared the writing samples of learners and native speakers to various academic articles and journals of expert writers from a computer science corpus. The results showed similarities in the production of the two groups in noun collocations when compared to noun collocations produced by the expert writers. However, both groups of native speakers and learners produced a similar number of verb collocations as the expert writers used. She conducted follow-up interviews with participants and found that sub-discipline, genre, and topic of the written texts were important factors in the accurate production of collocations. This suggests that advanced level learners can reach the level of native speakers when they have the same knowledge and expertise in a certain discipline enabling them to produce appropriate academic or specialized collocations. Farooqui's results indicate the possible effect of the material given to those EAP groups. She mainly focused on the differences between the two groups according to the underuse and overuse of collocations depending on their frequency.

This positive effect of learners' language proficiency level on their production of collocations proved to be true even when studying languages other than English. For example, Forsberg (2010) investigated the effect of language proficiency in Swedish learners of L2 French by comparing the oral collocation production from different learners' proficiency levels (beginner, high school students, and advanced university students). After analysing learners' interviews, he found that among those groups of learners, only advanced learners' production resembled the production of native speakers. Highly advanced leaners of French showed very little difference in their production when compared to native speakers. Although Forsberg's findings imply that there is a chance for advanced-level learners to produce collocations in a native-like manner, there is still a possibility that these results were due to the relative similarity between the learners' L1 and L2. Siyanova and Martinez (2015) also studied the relationship of learner proficiency levels in Italian. They investigated Chinese learners' production of noun-adjective collocations over three levels: low, intermediate, and higher levels, all of whom were participating in an intensive Italian language course. They analysed learners' essays on various topics and found that at a higher level, they produced higher quality L2 collocations. Although there was almost no difference in the number of the collocations between the three chosen samplings over the five-month course, learners were eventually able to produce more high frequency collocations than they did at lower levels and acquired stronger associated collocations. This suggests that a collocational performance development is possible, enabling learners to reach native speaker level when learning L2 whether it was a written or oral production, which might also suggest a relationship between learners' L1 and target language. However, this comparison was mainly quantitative.

While some studies did not indicate any relationship between learners' language proficiency and collocational production such as Chiu and Hsu (2008), the majority of the studies argued in favour of a positive relationship with more accuracy in producing L2 collocations. Whether L2 is English or not, the aforementioned studies support the argument that affirms the positive correlation between learners' language proficiency and their collocational production. This is true with studies that indicated a slight progression such as Abdul-Fattah and Zughoul (2001), Laufer and Waldman (2011), Paquot (2017), and with studies that show a similar native-speaker use by advanced level learners such as Farooqui (2016). Furthermore, some studies show a positive relationship between learners' language proficiency and production of certain types of collocations, such as Huat, (2012), Ebrahimi-Bazzaz et al. (2015). However, most of them limited their investigation to one type of collocation, or controlled the context of the study. Additionally, many studies were concerned with comparing the number of collocations or the high frequency collocations produced by learner with that produced by native speakers. There is a need for a study that fills the gap of investigating the written production of different types of acceptable collocations in relation to learners' language proficiency, especially given that fixed collocations have been shown to be more challenging for learners. Such a study could evaluate how learners produce L2 and these findings could then assist the process of teaching and learning and whether language proficiency can lead to a better performance as suggested in this section or an erroneous production as discussed in the following section.

3.4.2 Language extension

Language extension can be described as the tendency to extend the use of language to reflect a meaning, where its accurate expression is unknown to the learner (Ellis, 1994). Such a phenomenon can occur in the production of collocation by L2 learners in relation to their language proficiency level. Researchers, mentioned in this section, have argued that this issue is a result of learners not knowing a word or an accurate collocate, and therefore they rely on different techniques to extend their knowledge and use of the language to fill these gaps.

Cowie and Howarth (1996) compared essays written by native speakers and learners. They found that learners' written production lacked accurate use of collocations due to two reasons: extension by analogy, and experimentation. Cowie and Howarth described the process of extension by analogy as an attempt by the learner to use one word, which had been learnt previously, with numerous collocates. For example, a learner may use *win a discussion* rather than *win an argument* as verb-noun collocations. Learners may also produce *handsome girl* instead of *handsome boy* as an adjective-noun collocation. The verb *win* is restricted to the noun *argument,* as is the adjective *handsome* to the noun *boy*. Nevertheless, learners may think of extending their use according to their semantic references to produce other collocations. In addition, Cowie and Howarth defined experimentation as when a learner becomes confident and begins creating new combinations. For example, learning the adjective *long* and its use in the collocation *long road* may lead to an extension use, such as *long street,* thinking of *road* and *street* as synonyms can be used with the same collocates. Learners may learn a word in a certain context and then think that it applies to other contexts. While experimentation may result in some successful collocations, it is still not a sure method of creating them.

In the context of Arab learners of English, Abdul-Fattah and Zughoul (2003: 71) also considered the issue of extension as a possible reason for collocational errors. They reported that learners use different techniques trying to imitate the language of native speakers or convey a desired meaning, especially the higher-level postgraduate learners. Among the techniques are those that are related to the use of language extension: substitution, overgeneralization and analogy, assumed synonymity, quasi-morphological similarity, and derivativeness. A substitution is when a substitute, often a synonym, is used that has some of the semantic properties of the correct lexical item such as *he cracked her heart* instead of *broke*. Overgeneralization and analogy happen when learners try to extend the use of certain words into a new context such as *the police ashamed the law* rather than *broke the law*. As the name of the strategy indicates, assumed synonymity, occurs when learners fail in using a synonym correctly such as in *some workers interrupted the strike* rather than *broke the strike*. Quasi-morphological similarity is when learners rely on using words that have a similar morphological structure such as the two verbs *retreated* and *retracted*. Finally, derivativeness is when learners attempt to create a word from a previously learnt and known words. For example, Abdul-Fattah and Zughoul's learners derived the verb *lighted* from the noun *light*, the opposite of heavy. However, they noted that those strategies were not exclusive to one of the two levels of learners but occurred in both groups' data in differing amounts. Learners may sometimes feel more confident experimenting with the language they have previously acquired whether they succeed or not. They may also lack the necessary guidance in order to learn how to use what they have learnt effectively, even if they were of a low-level of proficiency. These findings are noteworthy in assisting language learners to use their language knowledge effectively to produce native-like collocations.

Kuo's (2009) study, which was discussed earlier (see Sections 3.2 and 3.3), also employed the term 'extension' to refer to learners' extended use of collocations through synonymy and approximation. By synonymy, Kuo means the failure to use a synonym in a collocation instead of the correct collocate, such as *broaden your eyesight* instead of *broaden your vision*. While *vision* has a metaphorical meaning reflecting the desired meaning in *broaden your vision*, *eyesight* does not. Kuo's learners mistakenly used *eyesight* as a synonym to *vision*. Approximation is the use of incorrect collocates when a collocation is semantically still acceptable. For example, a learner may say *giant car* instead of *big car*. Although using *giant* in the collocation does not sound accurate, it still expresses the desired meaning.

Even though research that studied the erroneous production of collocation by learners because of the language extension employed different terminologies, the descriptions of this process overlap. Ultimately, all terms refer to the incorrect use of collocates, caused by learners' lack of L2 knowledge. That is why a learner would look for a synonym or try and imitate previously known words in collocations. Therefore, it is a phenomenon that is linked to learners' language proficiency and will be discussed along with this study's findings, specifically, when referring to the problematic types of collocations produced by each level of learners. While the textbooks Saudi learners are taught from (Soars and Soars, 2011), like many other teaching materials, lack an explicit focus on collocation (Brown, 2011), Saudi as well as EFL learners may tend to use the vocabulary they have already learnt to produce collocations, something which is not always successful. Additionally, this lack of teaching collocation may lead them to use a direct translation from their L1. These are issues need further exploration, particularly in terms of the types of collocations produced by learners, and are discussed in the light of this study's findings in Chapter Six.

3.4.3 L1 interference

For Marton (1977)**,** learners may not be able to produce the same collocations as native speakers accurately because L1 interference may affect their use and cause them to rely on L1 translation. Many researchers identify L1 interference and direct translation as one of the major issues affecting collocation production across L1s of EFL learners. Many of the research studies mentioned in previous sections, such as Abdul-Fattah and Zughoul (2001) and Laufer and Waldman (2011), suggest this possible effect of learners' L1 on their collocational use, mainly through examining the use of L1 translation.

Abdul-Fattah and Zughoul (2001), who investigated Arab learners' use of the verb *broke* in Arabic, *Kasara* – كسر, found that learners achieved lower scores in the translation task than in the multiple-choice tasks. Respectively, from the higher level to the lower level, only 18.90 and 14.47 percent of the responses were correct in translation whereas 57.57 and 42.77 percent of the responses were correct in the multiple-choice tasks. The results showed different interpretations of the use of L1 translation. Abdul-Fattah and Zughoul investigated Arab learners' recognition of the collocates of the verb *broke,* where learners succeeded in responding to collocations, such as *broke the table* and *broke his opponent's nose,* by translating from their L1. However, some collocations of the verb *broke* were not successful when translated: for example, *shattered her heart* instead of *broke her heart* and*flouted her husband's oath* instead of *broke her husband's oath.* Even though the meaning of the latter is not transparent and could be a reason for it not being correct, the earlier *broke her heart* is said in the same way in Arabic. Their findings also show that learners with high levels of proficiency still face difficulty in producing English collocations, even when using L1 translation. It is true even when most of the English collocations investigated in their study have L1 equivalents.

Kim (2003) found a strong influence of L1 and learner’s proficiency levels when investigating the production of different types of collocations by Korean learners. In a task-based study, he tested learners on four types of collocations: verb-noun, adjective-noun, verb-preposition, and preposition-noun. His learners had not been taught collocations; however, when responding to a post-task questionnaire, the learners said that they relied on taught vocabulary to answer the collocation test. They also used bilingual dictionaries to help them with their translation from L1 to L2. Kim found, through their responses to the questionnaire, that learners’ production of collocations in the test indicated their knowledge of words and meanings in isolation rather than combined with other words. He also observed that L1 interference influenced learners’ use, and correlated more with two types of collocations: adjective-noun and verb-noun collocations. Kim, further, suggested that learners relied on three strategies in their production of collocations. The first, he named, the “full comprehension”, by using taught chunks of prep-noun like *by mistake*. The second is the “educated guess”, which learners used words they have learnt to combine in a collocation such as verb-prep, *look at*. The third is the “random guess”, which principally was encouraged by the L1 matches, such as in *do the shopping*. His learners were not taught collocations explicitly, and their use of bilingual dictionaries may have interfered the process of L1 translation, or combining vocabulary they learnt to form collocations.

Shehata’s (2008) findings are similar to those of Kim’s, but investigating Arab ESL and EFL learners’ production of collocation. In her task-based investigation, she noted that learners produced inaccurate collocations, such as *wide imagination* and *wide public.* These inaccurate collocations imitated learners’ L1 i.e. Arabic, equivalents as in *khayal wase'e* – خيال واسع- and *jumohoor a'areedh* – جمهور عريض-, where the English adjective *wide* was used as a literal translation of the two Arabic words, *wase'e* – واسع- and *a'areedh* – عريض-. All three adjectives share the meaning of 'spacious' or 'huge'. Her results were also conclusive with the use of two types of collocations: verb-noun and adjective-noun. However, Shehata's study is different to Kim's in not giving the learners a further instrument to use while translating such as dictionaries.

The influence of L1 on types of collocations was also found across a number of different L1 studies e.g. Polish EFL learners. Martyńska (2004), (see Sections 3.3 and 3.4.1), found that while all instructions were given in the learners' native language to make sure they did not misunderstand any task, it did not help learners to reach a higher performance. As an example of a verb-noun collocation is *ride a car* instead of *drive a car,* and *make shopping* instead of *do shopping* or *go shopping*, which were in a matching words task to create collocations. A further example is the Polish verb *robić,* which is equivalent to the English delexical verbs *do* and *make*. While English has collocations like *make breakfast* and *go shopping*, because of a literal L1 translation of this verb *robić,* a Polish learner may produce a collocation like *make shopping*. In addition, he points to this finding of L1 interference affecting learners' collocational production more than recognition. These results may be because textbooks and teaching materials focus on teaching verbs and their use more than other words, such as adjectives and nouns. Prepositions also are usually taught with examples, specifically, to show their use and meaning. Martyńska's findings are similar to previous studies indicating that verb-noun and adjective-noun collocations among other types are those most influenced by the learners' L1. However, they all studied this L1 interference through tasks which were controlled in some aspects.

Bahumaid (2006) also investigated the effect of L1 translation on the production of collocations by university-level translation teachers, who can be considered high proficiency language users. They either taught translation or were involved in translation professionally. They were given a list of 30 sentences to translate, including Arabic and English collocations that have similar grammatical forms, composed of nouns, verbs, adjectives, and prepositions, and which have various literal or restricted meanings. The findings suggest that using L1 translation did not help participants even when they are of a very advanced level, especially in collocations such as *broken English* and *virtually dead*. The participants even found that translating from Arabic into English is more difficult than from L2 to L1, which suggests their relatively lower competence in L2 than their own mother tongue. Therefore, they needed to refer to bilingual dictionaries frequently in collocations equivalent to *run a discussion* and *growing industries* before giving either a literal translation or a synonym. However, Bahumaid's study was limited to those 30 sentences he chose to test participants on.

Some researchers such as Yamashita and Jiang (2010) have examined the influence of L1 on learners' production of collocations with Japanese EFL and ESL learners by comparing their performance to native speakers in a task-based study. The task included a collocation list of 24 congruent (with equivalence in L1) such as *make lunch*, and 24 incongruent (with no equivalence in L1) collocations such as *kill time*. Participants had to rate collocations as acceptable or not using yes or no, and they were timed. Both ESL and EFL learners were slower with incongruent collocations than with congruent collocations, though ESL learners scored higher than EFL learners. This finding suggests that L1 has an influence on the performance of ESL and EFL learners even though L2 environment still plays a significant role in producing accurate collocations. Even though Laufer and Waldman's (2011) study was less controlled than

Yamashita and Jiang's, they were consistent in highlighting that advanced learners performed no better than lower-level when producing erroneous collocations caused by their L1. This was because the Hebrew learners in Laufer and Waldman's study tended to rely on their knowledge of words as individuals rather than developing them as collocations, and on word-to-word L1 translation of verb-noun collocations, such as *solve the disease* and *get the aim*. Thus, L1 translation is not always advantageous in producing collocations by learners, and it can sometimes be a cause for collocational errors.

Kurosaki (2012) used another approach to test the effect of L1 in the production of collocation by learners. He compared the collocational recognition and production by EFL learners with different L1s, namely, Japanese and French. In multiple-choice and translation tasks, he tested them on four types of collocations: verb-noun, delexicalised verb-noun, adjective-noun and adverb-adjective. The results show that although different L1s learners produced collocations differently, the different origin of those L1s; for example, French is an Indo-European language like English, while Japanese is not, was not significant. Their performance was more correlated to the types of collocations. In the recognition task, French learners performed better with verb-noun and delexicalised verb-noun collocations, while Japanese learners performed better at adjective-noun collocations. However, in the translation task, both French and Japanese learners were influenced by their L1 although this is correlated to types of collocations. The number of errors in collocation with verbs were nearly the same in both learners' production; nonetheless, French learners' L1 seemed to cause more errors with prepositions whereas the Japanese learners' L1 seemed to cause more errors with nouns. This suggests that L1 interference, no matter what their L1 is, does not necessarily correspond consistently and positively with learners' production of collocation but most probably depends on other factors, such as the type of collocation.

In consequence, L1 interference could only be a minor issue in learners' production of collocation, and sometimes with a positive impact. From the studies discussed above, Kuo (2009), who investigated the collocational errors made by Taiwanese EFL learners, found that L1 interference was not a dominant cause of collocational errors. Ibrahim et al. (2012) argued in support of the positive influence of L1 on learners' production of collocations when investigating postgraduate Iranian EFL learners. They suggest that L1, along with cultural influences such as topics and types of food, could affect the learners' accurate production of collocations because of the Persian equivalents. This observation is also consistent with those of Parkinson (2015), who found that Mandarin learners' L1 had a positive effect on their noun-noun collocation use. This positive effect is because they have L1 counterparts for the noun-noun phrase system, unlike Spanish and Tswana learners, who do not. This feature allowed for the production of L2 noun-noun collocations by the Mandarin learners, who produced more accurate collocations than the other two groups, even though Tswana learners were in an L2 environment.

Throughout the different contexts, most of the aforementioned studies suggest that L1 does influence the production of collocation, whether in recognition or production, and that L1 translation is not always productive. This is true even for advanced learners. Given the concerns of this study, the primary three objectives are to investigate the accurate written production of collocation by Saudi learners in relation to the types of collocations and their different language proficiency levels. While this study does not investigate the issue of L1 interference or the use of L1 translation in great detail, exploring the problematic types of collocations can lead to this L1 interference as a possible cause for the difficulties encountered. Therefore, my study will argue in favour of studies such as Kim, (2003) whose learners are not taught collocations, and Shehata, (2008) whose learners are Arabs. Also, both of them have investigated verb-noun and adjective-noun collocations, which are two of the types under investigation in this study, and suggested that learners' literal L1 translation can cause collocational errors. Laufer and Waldman's (2011) study can further be relevant as it linked the collocational errors caused by L1 interference to learners' language proficiency level, which this study does as well.

3.4.4 L2 exposure

A number of research studies, discussed in earlier sections, have suggested that L2 exposure – specifically, the learners' L2 environments – can affect learners' production of collocations. In Siyanova and Schmitt's (2008) study, they divided their learners into three groups, according to how long they had lived in the L2 environment, to rate the acceptable collocations. Their finding suggests that learners who had lived for longer in the L2 environment performed better than learners who had not, and developed a better sense of frequent and infrequent collocations, but not to the level of the native speakers' sensitivity. This is consistent with the Shehata's (2008) study of Arab learners. When she tested ESL learners at an American university and EFL learners at an Arabic university for their collocational recognition and production, she found that the ESL learners performed better than the EFL learners. Interestingly, the L2 environment had a further impact for ESL learners in that they were not influenced as much by L1 as EFL learners. Meechai and Chumworathayee's (2015) study findings of Taiwanese learners is also consistent with these results. Learners who attended the English program were better and less influenced by L1 in their collocational use, than those who attended the L1 regular program. Even though the two groups made numerous errors in producing, particularly verb-noun, collocations in translation, and gap-filling tasks, still the L2 environment showed its impact in elevating learners' production of collocation and minimizing the L1 influence.

In contrast, Yamashita and Jiang's (2010) investigation of Japanese EFL and ESL learners' production of collocations suggests the opposite. ESL learners scored higher than EFL learners in collocations with L1 equivalents whereas they did not perform better in collocations with no L1 equivalents. Although an L2 environment played a significant role in learners' production of collocations, it was not enough for collocations that do not have L1 equivalents. Fan (2009) writes that the production of EFL and ESL learners of the same L1 overlaps between L1 interference and L2 confusion. The subjects of her study were weak in the use of grammatical, as well as lexical, collocations despite being in L2 environment.

Some other researchers investigated this effect of L2 exposure explicitly by exposing learners to an L2 environment through elicitation tests. Webb, Newton, and Chang (2013) studied the recognition and production of 18 verb-noun collocations of high *t-score* and low congruency with learners' L1 by Taiwanese EFL learners. They were divided into five groups, one of which was a control group that was not exposed to the recordings. A native speaker read recordings that included the 18 target collocations. Recordings were repeated once for Group 1, five times for Group 2, 10 times for Group 3, and 15 times for Group 4. This method aimed to test the effects of repetition and L2 exposure on learners and their production of collocations. One pre-test and four post-tests were given immediately after the task, which were conducted without telling learners about the tasks in advance. Results of this study suggest that learners can incidentally produce collocations with meanings when exposed to them repeatedly, which implies that when learners are intentionally and intensely exposed to the L2 environment, their collocational production can improve dramatically.

Fernández and Schmitt (2015), mentioned earlier in Section 3.2, studied in a subsequent stage of their research the influence of L2 exposure on Spanish learners' production of collocation through their use of everyday activities such as the social media. They asked the participants in a given questionnaire about the time they spend weekly on three activities: reading, films/ TV and social media. They fell into three groups: 0-1 hour, 1-2 hours, and more than two hours per week. Fernández and Schmitt found that this informal exposure of L2 had a positive impact on participants' learning of collocation, and that some activities i.e. reading had a greater influence than the other two activities i.e. films/ TV and social media. These results are indicative of the correlation between the informal exposure of L2 and learners' production of collocation as Fernández and Schmitt's study was not explicit about the choices of those three activities in specific. Furthermore, participants may vary in their interests, and as a result, their input of language from this informal L2 exposure will be different.

This role of L2 exposure was illustrated in the use and acquisition of language variations, namely British and American English in relation to how learners' language production could be further influenced by their teachers' input. In a corpus-based study, Larsson (2012) investigated Swedish EFL learners' production of American and British English spelling and vocabulary. All Swedish learners showed a dominance of British over American spellings and vocabulary, despite having stated their preference for American over British English in a questionnaire. However, Larsson's study suggested that these learners' results were affected by their teachers' preferences and the British English learning materials, and that their preference was due to media exposure. On the other hand, Hameed and Fatima (2016) found a gap between the taught British English and the American English surrounding learners. They investigated Saudi university students' awareness and preferences towards the English varieties by testing them on pronunciation, spellings, and attitude. Their results show that students' written and spoken use of American English was more accurate than that of British English. Hameed and Fatima (2016), thus, suggest that the students' oral production could be due to language exposure via movies or online TV programmes.

L2 exposure is not of direct relevance to this study investigation. However, there is a link between the lack of L2 exposure, as described in Chapter One, and that the Saudi learners are not taught collocations. While they are not ESL learners, they are not either given an explicit training nor focused classes on how to identify and use collocations. Thus, their language learning process lacks a proper exposure to and practice of English collocations. It is unlike the case of learning grammar and sentence construction which are taught to develop writing skills, and acquiring vocabulary in relation to basic listening, speaking and reading topics and skills. This lack may result in producing either erroneous or inappropriate collocations, especially as studies in this section suggest that there is an effect of L2 exposure on learners' production of collocations. This is either through living in L2 environment, involving them in English medium programs, or simply through implicit tasks and interventions. This was especially noticeable in the two types of collocations investigated in this study i.e. verb-noun and adjective-noun.

As this study is not task-based such as some research discussed in this section (Webb, Newton, and Chang, 2013), which investigated thoroughly and precisely the effect of any L2 exposure, yet the findings of this study could be indicative of how much this lack of L2 collocational instruction may affect Saudi learners' written production of collocations. This study is concerned with other factors such as learners' language proficiency levels and the different types of collocations, which some of the studies mentioned above have highlighted in addition to L2 exposure such as Shehata (2008).

3.5 Summary

The main objective of this research is to examine Saudi learners' production of acceptable collocations in their writing according to the following research questions:

1. Do Saudi foundation-year students at university produce acceptable collocations in writing? If so, what are the types?
2. Which less idiomatic combinations do Saudi foundation-year students produce in writing?
3. What are the similarities and differences between the acceptable collocations produced by two levels of Saudi foundation-year students, studied in their written texts?

To achieve this, a corpus approach to identify those collocations produced will be employed. Selected research, discussed in this chapter, applied corpus approaches through combining the reference to the number of citations of collocations in the corpora and to the association measurements, to judge acceptable collocations and their level of fixedness. When Nesselhauf (2003) and Parkinson (2015) referred to the BNC, five citations were required for a collocation to be considered acceptable, whereas in COCA, it is 25 citations. However, the case is different with the use of the association measurements. Each researcher would refer to the test i.e. *t-score*, MI score and LogDice, that is more suitable for the study's purpose. In this study, which is investigating relatively low-level learners, the LogDice will be used. This measurement combines the benefits of the *t-score* by recognizing high frequency collocations while still considering the exclusivity that is a feature of the MI score. The reason for not using the t-score is due to the nature of this research in which the focus is the lexical combinations i.e. collocations, rather than grammatical i.e. colligations, which is usually the focus. The MI score can measure the lexical combinations and was commonly used by previous researchers such as Siyanova and Schmitt (2008), Parkinson (2015) and Fernández and Schmitt (2015), but it is not used in this study because it mainly identifies very exclusive combinations of rarely or low-frequency collocations, which is not suitable for the level of the participants in this study. Thus, by using the LogDice to identify collocations produced by Saudi learners, I aim to contribute to knowledge and literature of language learning.

The production of collocations by L2 learners is not perfect, and is often said to be challenging if not problematic across different levels of learners. This is due to a number of reasons; this study aims to address two of them. The first one comes from the learners' problematic production of different types of collocations, and less idiomatic combinations. Learners always have more difficulty in production than in recognition (Martyńska, 2004; Shehata, 2008; Brashi, 2009), which was suggested to be sometimes linked to particular types of collocations such as in Nesselhauf (2004), Siyanova and Schmitt (2008), Fan (2009), Kuo (2009), Durrant and Schmitt (2009) and Parkinson (2015). In other research, it is linked to the learners' language proficiency level, which is the second reason addressed in this study, such as in Abdul-Fattah and Zughoul (2001), Laufer and Waldman (2011), Huat (2012) and Farooqui (2016). There is a need to combine the investigation of learners' written production of collocations of different types along with the effect of their proficiency level. This study considers whether similarities and differences in leaners' production of collocations are due to the different types of collocations or their proficiency levels. It will also explore what sorts of similarities and differences could be found, especially when looking at the less idiomatic combinations as they would possibly suggest problematic areas in learners' production.

As being discussed that aforementioned studies were concerned with analysing learners' written texts, other than Arabic L1 learners. However, it is important to conduct a similar study in the context of Arab learners considering a less controlled and led investigation than examining them in given tasks, such as gap fill and matching activities. In addition, some studies such as Kuo (2009) that have analysed learners' written production – other than Arab learners - were basically concerned with one or two types of collocations i.e. verb-noun and / or adjective-noun, especially, in the case of investigating the production of two or more levels of learners. There is a need to widen the scope, and study those types of collocations given above along with noun-noun collocations, and compare the different uses among two or more levels of learners. Indeed, these types can indicate the fundamental blocks in learners' writing skills as they try to create a coherent meaning in their texts. As has been said (see Sections 3.4.2 and 3.4.3), most teaching materials concentrate on teaching verbs, nouns and adjectives, and separately, not in combinations or collocations such as the case in this study context.

Besides these two major issues investigated in this study, which may affect the use of learners' written production of collocations, there are two other factors suggested by previous research. They are learners' L1 interference and L2 exposure. As noted previously, this study is not investigating these two issues in great detail. The main focus of the causes will be the problematic types of collocations and the learners' language proficiency as the research questions indicate. Nonetheless, there might be some occurrences of collocational errors produced by the Saudi learners to be reported in the findings along with possible causes such as the ones proposed by early research. The effect of L1, when it is Arabic, was indicative in some of the studies such as Shehata (2008), and Brashi (2009), but only in the production of verb-noun and adjective-noun collocations. At this point, L2 exposure does not relate to Saudi learners' context; however, their written production of collocations may be affected by a lack of explicit L2 instruction; in addition, to taught materials and topics. Studies discussed such as Webb, Newton and Chang (2013) and Meechai and Chumworathayee (2015) were in favour of exposing learners to L2 environments whereas others such as Larsson (2012) and Hameed and Fatima (2016) were different highlighting more on the role of teachers and taught materials.

Chapter 4 Methodology

4.1 Introduction

The methodology chapter of this study falls into two main parts: the first is the research design, and the second is the research method including the analytical procedure. To answer the research questions (given below in Section 4.2), a strategy of triangulation has been used as a main framework for the study. It addresses the cross-linguistic comparison of Saudi learners' written production of collocation with those of native speakers' available in the BNC through the different research instruments used to identify collocations, as will be described. This strategy also enables the cross-sectional comparison between the two levels of Saudi learners; pre-intermediate and intermediate using this same methodological approaches.

The remaining sections of the chapter will discuss the three primary dimensions required in a corpus-based research project such as this, which are: authentic data, specialised software to analyse and compare data, and a researcher to interpret the outcomes in relation to proposed research questions. The research methods are illustrated by describing the participants' contexts and how data was collected and sampled. The three-step analytical procedure applied to analyse the learners' written texts, is then described. The first is extracting candidate collocations under investigation from learners' texts, and classifying them. The second is identifying those extracted collocations in the BNC available in Sketch Engine software by referring to collocations' citations, frequency and LogDice score. The last step is concerned with those combinations, that were not identified using the BNC, and that is to refer them to the judgement of native-speaker informants to whether or not they can be considered as acceptable collocations. After these analytical procedures, the two levels of Saudi learners' production of collocations can be compared with each other, and in relation to the different types of collocations under investigation.

4.2 Research design

It is a complex task to investigate whether collocations produced by Saudi students in their writing are acceptable or not, and to identify the similarities and differences between pre-intermediate and intermediate learner levels in their uses of the different types of collocations. Therefore, the research design is framed primarily through triangulation. The triangulation methodology allows the use of different methodological approaches to study a certain phenomenon and assures the systematic analytical process and improves the validity of the research (Jick, 1979:602; Dörnyei, 2007:61). Three approaches have been used in the study to present a qualitative and a quantitative approach to the manual analysis of the learners' texts. The approaches used are contrastive interlanguage analysis (CIA); a frequency-based approach; and a cross-sectional approach. The three approaches together aim to answer the three research questions:

1. Do Saudi foundation-year students at university produce acceptable collocations in writing? If so, what are the types?
2. Which less idiomatic combinations do Saudi foundation-year students produce in writing?
3. What are the similarities and differences between the acceptable collocations produced by two levels of Saudi foundation-year students, studied in their written texts?

This study uses contrastive interlanguage analysis (CIA) as a general methodological framework to enable the linguistic comparison of the learners', under investigation, use of collocations with native speakers' use of collocations available in the BNC. Granger

(2009) employs the CIA approach to analyse learner corpora because of its ability to investigate the special features of learner language in comparison to native speakers' language. She further states that CIA helps study the language of two language learners in different contexts by showing similarities, differences, and/or generalizations. Huat (2012:192) also agrees on the usefulness of applying the CIA approach by using corpora as a tool to compare the language use of learners and native speakers as well as that of several groups of learners from different L1 backgrounds. Hence, framing the study using the CIA approach allows the proposed investigation of comparing Saudi learners' written production of collocation with that of native speakers as evidenced in the BNC. In addition, the CIA allows the comparison between the two different levels, pre-intermediate and intermediate, of Saudi learners' uses of collocation.

In order to conduct these comparisons by referring to corpora, the study applies the frequency-based approach, as discussed in the literature review, Chapter Two, Section 4.2. This second approach allows collocations produced by the learners to be checked in relation to their frequency and association measurements in the BNC, available in Sketch Engine software, to judge their acceptability. This approach also coincides with Wray's (2009) classifications of studying collocation, as indicated in Chapter Two. Although Wray's three methods of studying collocations are concerned with examining common words, a group of words, and specific theories, something which is not precisely this study's concern, they do use corpora as a reference. This study, in contrast, investigates four types of collocations, i.e. verb-noun, noun-verb, adjective-noun, and noun-noun by referring to the BNC. By doing so, the extracted collocations from Saudi learners' written texts can be evaluated for their fixedness and then examined for other issues, such as the most or least problematic types. This frequency-based approach was applied by several studies as described in earlier sections (2.4.1 and 3.2) in the literature review.

Furthermore, interpreting the frequency numbers and scores of strength of collocations in the BNC using the corpus software validates the comparison between the two levels of Saudi learners to indicate possible similarities and differences in their production. Further details are given below.

The third approach this study applies is the cross-sectional approach, which is used to explore similarities and/or differences between two groups of learners. Even though Huat (2012) suggests that the CIA approach is mostly and exclusively used with longitudinal studies, Granger (2009) notes that learner corpora research applying the CIA approach can also use the cross-sectional approach. Ellis and Barkhuizen (2005:97) state that cross-sectional studies resemble longitudinal studies. Namely, much as cross-sectional studies allow the researcher to record the differences between levels at the same period of time, longitudinal studies record the differences among learners with the same proficiency level over an extended time period. Granger (2012) calls this cross-sectional approach ‘quasi-longitudinal’, in which the sample of learners is from the same type, i.e. the same L1, but from different proficiency levels. In this study, the Saudi learners are from two proficiency levels (pre-intermediate and intermediate), but they still belong to the same context and are investigated during the same time period. The writing samples of the two levels of learners were collected at the same time using the same criteria and then were analysed by applying the same analytical procedure.

The three methodological approaches of this triangulated research design overlap in examining the three research questions with each approach reflecting on a specific angle. The three research questions focus on pre-intermediate and intermediate Saudi learners’ use of collocation by comparing their written production with that of native speakers.

4.3 Participants and Context

Participants in this study are female Saudi university students attending the foundation-year program at King Abdulaziz University (KAU) in Jeddah, Saudi Arabia. Participants are between the ages of 18 and 19 and are attending one of four general English courses provided by the English Language Institute (ELI) at KAU in preparation for their university programmes.

The educational system in Saudi Arabia separates male and female students. For the purposes of this study, only the written production of female participants was collected. In general, students in Saudi Arabia can be grouped according to two kinds of language learning experiences they have prior to university. The first group of students have attended government schools that teach English to students from the age of nine with five hours of English lessons per week. The second group have attended private schools, where they usually start to learn English at the age of six and with a more intensive English learning experience. If any students are bilingual or have a very high level of English, their fluency is most likely due to family circumstances or further individual English learning lessons. For these reasons, the ELI requires all students at KAU to take the Oxford Online Placement Test (OOPT) (https://eli.kau.edu.sa/Pages-en-st-resources.aspx) in order to place them at the appropriate language learning level. These levels coincide with the Common European Framework of Reference for Languages (CEFR) levels A1 to B1+.

The students at the ELI are taught using the New Headway-Plus books (Soars and Soars, 2011) at the following four levels: beginners, elementary, pre-intermediate, and intermediate. The textbook lessons are organized with different topics that vary according to each level. Each unit targets different skills, including reading, grammar, listening, speaking, vocabulary, and everyday English expressions. Only writing is taught separately using a writing booklet, which has been developed by the ELI academic staff. The tasks in the writing booklet for each of the four courses are extensions of topics in the corresponding textbooks. Each lesson in the writing booklet begins with exercises that integrate the vocabulary and grammar from the taught materials into practical writing tasks. The lessons then develop those tasks into more advanced tasks in order to brainstorm ideas on the topic before the actual writing task begins. The lessons end with a task that requires students to write a paragraph (beginners and elementary levels) of 120 words, or an essay (pre-intermediate and intermediate levels) of 200 to 300 words. The students write up to three paragraphs or essays per course in topics assigned by the teacher or the course coordinator. The writing texts used in this research project were taken from the writing assessments students have to take at the end of the attended level. They were also asked to produce the same number of words they used to practice in class; however, the topics may or may not be the same. Yet, topics generally correspond to what students have learnt in the grammar and vocabulary classes.

4.4 Methods for data collection and analysis

This section first describes the procedure for collecting data and sampling. It then explains the three-step analytical procedure to extract and identify collocation in the both Saudi learners' levels' writing.

4.4.1 Data collection and sampling

Collecting and sampling commenced after ethical approval from the University of Leeds Ethics Committee was received. The data collection was carried out from January 11, 2015 to March 5, 2015 at the ELI of KAU.

As mentioned in the previous section, first-year students at KAU attend four levels of general English courses as part of the foundation-year programme. The language teaching levels are beginner, elementary, pre-intermediate, and intermediate. Students attending the four courses are taught writing skills, and those attending the pre-intermediate and intermediate levels write 200-to-300-word essays as part of their coursework. To investigate learners' collocation production, essays were collected, which the students produced without intervention and prior knowledge that their essays could be used for research purposes. Kuo (2009) and Fan (2009) also employed learners' written work for similar investigations. These researchers contributed to a corpus-based study by analysing 200-to-300-word essays written by learners from Taiwan (Kuo, 2009) and Hong Kong (Fan, 2009). Some written samples exceeded the word limit even though teachers asked their students to write only 200-300 words; samples exceeding the word limit are also noted and included in the current study.

Howarth (1998:31) indicates that the significance of the study of collocation is that it shows the L2 learners' points of view i.e. their production. This study investigates Saudi learners' written samples which in this context represent freer written than studies which use elicitation methods. This sort of data collection reflects the actual language the learners' produce without any intervention by a researcher. It should be emphasised that students did not write essays for the purpose of the study but as part of their coursework. Even though writing is part of their coursework, and they are influenced by the topic, timing, and purpose of writing (e.g. course grades), the students did not intend to demonstrate their collocational use in their writing. As described earlier, the given topics and time limits were both determined by the university, but students were not asked to use or produce certain types of collocations in their essays. In addition, students were not aware at the time of writing the essays that their texts would be used for research purposes. They were informed afterwards in order to obtain their consent for participation.

A random selection sampling procedure was followed at all stages of the data collection. Fraenkel et al. (1993) note that random sampling selection uses a casual selection of samples, which avoids subjectivity in selecting, yet is applied systematically. Dörnyei (2007:97) describes it as every *n*th number in a group of samples. There were nine groups of learners from each of the: pre-intermediate and intermediate levels, including between 25 to 30 students in each group, enrolled according to their university ID numbers. By applying the random sampling proposed by Dörnyei, every third of these nine groups was selected. Thus, the initial written samples were collected from three groups of learners in each level, resulting in 90 samples from the pre-intermediate level and another 90 from the intermediate level. To narrow the sample selection further, every third sample of the 90 texts were selected, resulting in 30 written samples from the pre-intermediate-level texts and 30 from the intermediate-level texts.

After this initial selection process of students’ written samples, and with the help of the two levels’ teachers, the students were contacted to obtain their permission and have them sign consent forms allowing their written texts to be used for this study and to ensure the privacy of their information (Appendix Three). The consent form was provided in both English and Arabic and was signed in both languages. The form stated that these essays would only be used for research purposes and the authors would remain anonymous, with only the research supervisors and the author having access to the texts. The consent form also included brief information about the study and the researcher’s contact information in case students wished to withdraw their permission at a later date. However, the form indicated that withdrawal should be within two months of the signatory. All of the students who provided the selected written samples agreed to let me use their written texts and signed the consent form. The final stage after students' approval and been obtained was to photocopy their written samples as originals are not allowed to be removed from the KAU campus.

Since the data analysis is manual, including in-depth and individual text analysis as well as time consuming for the current research time frame, this number of samples was still more than required (Dörnyei: 2007:38). Thus, the number samples had to be reduced further as in the initial procedure, every third text was selected to reduce the number down to ten written samples from the pre-intermediate-level learners and ten from the intermediate-level learners. After initial examination, two of the pre-intermediate level texts were not legible to read and were excluded. For consistency two samples from the intermediate level texts were also excluded by selecting the third and sixth texts. Thus, there were 16 written samples for this study, eight from the pre-intermediate level and eight from the intermediate level. Two months after the data collection, all data were anonymously entered and stored on the university hard drive. There was no further contact with the students after the data collection period.

4.5 The analytical procedure

As described above (Section 4.2), this study uses a three-step approach in analysing the students' written texts to investigate their collocational use. The first is extracting candidate collocations from the texts manually, the second is identifying those extracted collocations according to their fixedness by referring to the BNC, using the Sketch Engine tool. The third step concerns only collocations that were not identified from the corpus and entails consulting native-speaker informants for their judgement.

4.5.1 Extracting candidate collocations

Various research projects have used different procedures to extract candidate collocations from learners' written texts. Some of the studies used extraction software programmes such as TACT (Granger, 1998) and ConcApp (Fan, 2009). Other studies relied on manual extraction such as Siyanova and Schmitt (2008) and Durrant and Schmitt (2009). Due to the relatively small number of texts and the low-level of learners, as well as the interest in observing the learners' writing closely, manual extraction was the procedure used for candidate collocations. However, this procedure is challenging and should be employed systematically with care and awareness of its limitations. Even though it was conducted by previous researchers such as Siyanova and Schmitt (2008) and Durrant and Schmitt (2009), they addressed adjacent collocations and of one type i.e. adjective-noun. This study is different as it is investigating more than one type of collocation, and within a wider word span than those of the adjacent collocations. This manual extraction is time-consuming and would need several revisions to assure the systematic application, especially as this study was carried on only by a single author with no other co-researchers. Although extraction software can minimize any mistakes which may occur because of the manual extraction, it was still felt that manual extraction is more suitable to the students' level of writing and sometimes the less idiomatic use of the language as well as being more effective for handwritten samples. As stated previously, these Saudi EFL learners are of pre-intermediate and intermediate proficiency levels, and their written production is likely to contain grammatical errors and spelling mistakes that can better be explored and dealt with through manual investigation. Before finalizing the extraction procedure and criterion of candidate verb-noun, noun-verb, adjective-noun and noun-noun (VAN) combinations from learners' written texts and to ensure its validity, the analysis was tested with samples from both a pre-intermediate level and an intermediate level learner. This initial testing has further supported the decisions made and followed systematically throughout the extraction.

The extracting procedure firstly started by proofreading learners' written texts for spelling mistakes, which were corrected, and in the case of American English spellings changed to British English. This is to ease the process of extracting the candidate collocations but was done without influencing the learners' lexical use. For example, a pre-intermediate level student wrote "*my dad tryed to call a taxi*", which was corrected to "*my dad tryed [tried] to call a taxi*". It is apparently a spelling mistake where the student misspelled the verb *try* in its past form. Square brackets were used to indicate a spelling correction. Examples which included grammatical errors were left as they are and were not corrected as the main concern of the study is the production of collocation; the lexical production of fixed combinations. In an example like, *"we just falled in deep sleep",* the learner misused the past tense of the verb writing *falled* instead of *fell*, and the preposition *in* rather than *into*. This was not relevant as the study refers to the lemma of the words under investigation i.e. verbs, noun and adjectives. There is no specific or important reference to verb tense such as in the example mentioned, or to noun plurality e.g. letter vs. letters, or adjective conjugations e.g. good, better and best. In such examples, what is of interest to this study is the co-occurrences of the two collocates in a fixed and exclusive combination. *Fall sleep*, as an example, is to be considered a candidate collocation, whether the verb is *fall, falls, falling, fell* or *fallen*. The decision to not correct grammatical errors as well was made because it was preferable to keep the data as close as possible to its original form with not much interference. I made no changes to students' writing even in case of ellipsis such as in the example *I was very tired and my brother also*, where the full form should be *I was very tired and my brother* *was tired also [too]*. All written texts in Appendices Five and Six are after corrections, and as used in the study.

This study focuses on the production of the following types of collocations: verb-noun, noun-verb collocations, adjective-noun collocations, and noun-noun collocations. These were named as VAN combinations and identified as candidate collocations at this extraction stage. V refers to verbs, A to adjectives, and N to nouns. The extraction procedure followed has specific criteria in choosing these verbs, adjectives, and nouns collocates from the learners' written texts. First, those types of collocations are all with noun nodes: for instance, common nouns such as *city* and *life.* According to Bloor and Bloor (1995), common nouns are names used to address specific meanings and take countable or uncountable forms, such as *system*. Individual proper nouns such as *Ahmad* and *Jeddah* were excluded because they are, in this study, mostly related to the Arabic context, which is not common in native English corpora. Thus, they would not make a difference in the use of English collocations. Proper nouns such as *Google* were also excluded. In extracting candidate adjective-noun collocations, two types of adjectives: noun modifiers such as *beautiful photos* and verb complements such as *restaurant is small* were included.

When extracting candidate verb-noun collocations, lexical verbs, the open set verbs, and the primary auxiliaries *do* and *have* were included. Bloor and Bloor (1995) employ Halliday's classification of verbs into three major classes: lexical verbs, auxiliary verbs, and modal auxiliary verbs. For Halliday, lexical verbs are open sets like *write* and *read,* while auxiliary verbs are closed sets such as the primary *do, have* and *be,* and the modal auxiliary verbs *can* and *could*. Unlike the primary *be*, the other two primary verbs *do* and *have* are more problematic for Arab learners. This is because Arabic does not have an equivalent verb for *do* in collocations such as *do homework* and *do exercise* nor for the primary verb *have* when used in collocations such as *have fun* and *have lunch*. Thus, Arab learners are more likely to find them challenging when used, or misused, in collocations by the Saudi learners. The primary verb *be* such as in هو سعودي meaning *he is Saudi* and إنه 18 عاما meaning *he is 18 years old*, has an equivalent in pronouns in Arabic not particular verbs. It is similar to the case of the auxiliaries *do, have* and *be* when they are used as helping verbs in English. As a result, and as the data used in this study showed, learners tend to omit this primary *be* in their English writing due to its non-existence in Arabic. Additionally, the verb *be* is a frequent one in English, and language learners, in general, would have less problems with frequent lexical words. Therefore, it was excluded in the analysis. It should also be noted that no articles, pronouns, prepositions, conjunctions, determiners, quantifiers, or punctuation were included in the extraction criteria because this study is interested in lexical combinations exclusively, i.e. collocation, not in grammatical relationships such as in colligation. McEnery et al. (2006:82) indicate that grammatical words – for example, *the* and articles – occur in such high frequency that they can be excluded from the collocation investigation. However, they are included in the span word-count. It should be noted also that prepositional verbs such as *fall in, sit in*, and *live in* were not included as this study focus is the lexical combinations only.

The last and most important step in the extraction procedure was whether the collocation should be investigated as adjacent; for example, *small house*, or in a wider span; for example, *the house around the corner is small*. According to Stubbs (2002:215-216), considering lexical boundaries when extracting collocations can be problematic, and it would be difficult to determine the span. The problematic nature of such a consideration was further noted by Halliday, who raised this challenging issue of boundaries as occurring even in individual words. Halliday (2004:1) highlights the difficulty of identifying a clear-cut point for where to begin or end words such as "*English-speaking*", and for deciding whether such words should be hyphenated or presented as two separate words. Authors differ in their choices of word spans according to their aims and methodologies. For example, while Sinclair (1991) and Biber et al. (2004) consider a four-word window as a suitable word span for collocations, a five-word window within a sentence is still common and applicable (Arazy and Woo, 2007: 530). Thus, this word count span of a collocation, and whether to limit this span to a sentence, or beyond a sentence boundary, or less than a sentence, for instance clauses and phrases, remains the researcher's decision.

In the early stages of analysis, a five-word span limit within the sentences was used. However, it proved to be ineffective, resulting in many collocations that Farooqui (2016:102) dubs "linguistically uninteresting combinations" such as *children-toy* that can be frequent combinations but not necessarily attracted, meaning making strong collocations. Because of the learners' language proficiency levels, and the fact that run-on sentences and poor use of punctuation occurred continuously, analysing sentences was especially difficult. For example, a pre-intermediate student wrote the sentence, "*I felt so happy because I finished my exams so I needed to relax in somewhere"*. The noun identified in the sentence is *exams* and according to the five-word span it would collocate with the verbs *finished* and the adjective *happy,* which appear before it, and with the verbs *needed* and *relax,* which appear after. This led to awkward combinations or, using Farooqui terminology, uninteresting combinations. It was the same result when extracting using a smaller window such as a four-word span. Still, the verb *needed* appeared in the span of the noun *exams*. Smaller windows as two-word or three-word were not considered because they resulted in the omission of what seemed like relevant combinations. Thus, the application of a five-word window limit in sentences was not ideal in this study context. While using a five-word span has the advantage of including a variety of collocations produced in learners' written texts, the risk of producing uninteresting combinations makes it problematic for selecting collocations. However, if the span is limited to less than a sentence, i.e. a clause, then the probability of finding suitable collocations is significantly increased.

Even though adjacent collocation is important, it was decided that collocations in a wider span needed to be explored to gain a clear understanding of the students' collocation production. Durrant and Schmitt (2009) extracted collocations manually and limited their extraction to adjacent collocations i.e. pairs of words appearing next to each other, such as *green eyes*. In cases where the collocation had two adjectives, they extracted only the closest adjective to the modified noun. For example, in *beautiful green eyes*, they extracted only *green eyes*. When this system was applied in the early stages of the analysis, it created some limitations, which had to be modified. For example, if a student wrote *clean place*, the collocation is adjacent and would be considered as an adjective-noun collocation; but if she wrote *clean and nice place*, the collocation would not be adjacent and would not be included. This also shows that working on lexical groupings smaller than clauses, such as phrases, would result in losing some of the collocations that are not adjacent yet have a coherent meaning with the noun node i.e. in cases of verb-object collocations such as *the weather was fantastic*. Therefore, the five-word span was applied within a clause rather than a sentence. Employing the boundary of a clause allows for a larger number of lexical combinations in student texts, and more interesting collocations that would not have been included had adjacent collocations only been included.

According to Huddleston and Pullum (2002:20), clauses typically comprise two parts: the subject and the predicate, which is the definition used in this study but with the inclusion of conjunctions. Sentences are broken into clauses, and then candidate VAN combinations are extracted. Starting with the noun node and looking for the collocates i.e. verbs, adjectives and nouns, five words before and after. To illustrate this process, a student's sentence, "*when the experience was good, we would feel happy*", was divided into two clauses, "*when the experience was good*" and "*we would feel happy*". The first clause includes the noun *experiences*, which was extracted. The second clause, however, does not include any common nouns; therefore, no VAN combinations were extracted. As candidate VAN combinations should co-occur in a five-word span count before and after the noun node, only one collocate co-occurs with the noun *experience*. In the example, "*when the experience was good, we would feel happy*", *experience good* is extracted as a candidate adjective-noun collocation. There are no other collocates i.e. verb, or another noun or adjective for the indicated noun node.

This procedure for extracting candidate VAN combinations is subject to a limitation, which was addressed by Evert (2008). He claims (2008:12) that collocation co-occurrence can be measured by three criteria: surface co-occurrence, textual co-occurrence, and syntactic co-occurrence. He defines the surface co-occurrence as the word span between the collocates, the textual co-occurrence as the unit the collocates appear in, and the syntactic co-occurrence as the syntax the collocates share. As described throughout this section, candidate VAN combinations appear in a span of five words within clause boundaries, which address the surface and textual co-occurrence defined by Evert. For example, in a clause written by one of the students, "*Before 10 years ago, I visited my uncle with my mother"*, the nouns *uncle* and *mother* would collocate with the verb *visited* according to the five-word span and a clause boundary. However, *mother* is part of a prepositional phrase. Even though they do have a coherent meaning as a collocation, the student's language does not reflect the bond of the two items in a straightforward way. This is especially important, as this study is investigating collocations of two words, identified using a corpus approach discussed below. Therefore, this extraction process was modified by looking further at the syntactic relation of the two collocates in the candidate VAN combinations. This means that verb-noun, noun-verb collocations can include subject-verb and verb-object but prepositional phrases such as described in the aforementioned example were excluded. However, they can appear within noun phrases, adjectival phrases, or adverbial phrases such as *tell my family, visit sick person* and *became very good student*. The syntactic relation in the adjective-noun collocations, as described earlier, will be when adjectives are noun modifiers and verb complements. Then in the noun-noun collocations will be when the noun, the collocate, is a modifier such as in *room service* and *fruit garden*, and the two nouns co-occur in a list such as in *schools and hospitals* and *cars and buses*. It is a further advantage to limit the extraction like this in order to have a systematic replicable procedure as well as a much more focused investigation.

It is also to be noted that some candidate VAN combinations with three words such as *checked Google map, saw dancing fountain* and *asked security guard* were treated as special cases. Most, if not all, similar examples, did not appear frequently enough to reach the threshold score as well as the fact that this study is mainly concerned with two-word collocations. In such cases, *dancing fountain* and *security guard* were extracted as noun-noun collocations and *checked map, saw fountain* and *asked guard* as verb-noun collocations. When appropriate to meaning, as in the example of *asked security guard,* but not in *saw dancing,* a third candidate collocation was extracted also; the verb-noun *asked security*.

After extracting candidate VAN combinations - verb-noun, adjective-noun, and noun-noun collocations - from all the texts, they were entered into tables according to their types (see Appendices Seven and Eight), and then identified in the corpus according to the procedure in the following section.

4.5.2 Identifying collocations in the corpora

To identify whether or not candidate VAN combinations extracted from learners' written texts make acceptable collocations, they were checked against the BNC. The BNC was chosen as representative of British English users' language production. The BNC is also widely used and considered a reliable reference corpus in research. The BNC, used on the Sketch Engine, contains 96,134,547 words of modern British English and was completed in 1994 and revised in 2001 and 2007. It includes data from both spoken and written English texts produced by different ages and genders and in various topics and domains. This corpus is readily accessible in different online corpus tools, but preferably used for linguistic studies on software such as the Sketch Engine (https://old.sketchengine.co.uk/auth/corpora/) and Brigham Young University (BYU) (http://corpus.byu.edu).

Sketch Engine software was chosen because it includes all association measurements to test collocation strength i.e. *t*-score, MI score and LogDice. As noted earlier in the literature review (see Chapter Two and Three, Sections 2.4.1 and 3.2), the LogDice measure will be used in this study though other measurements may occasionally be referred to. However, the disadvantage in using this software is that it does not have an American English language users' corpus, such as the Corpus of Contemporary American English (COCA). COCA is the most commonly used corpus in research as a counterpart to the BNC in investigating English American as a corpus-related tool (Kilgarriff and Kosem, 2012:6). It would have been beneficial to refer to the two corpora using the same software for reliability and consistency and this was my original plan.

However, COCA is only available in the BYU website. Sketch Engine only has an American National corpus (spoken) with 3,202,026 words and another American National corpus (written) with 11,048,137 words. Even though this study is investigating learners' production of collocations generally, compared to those of native speakers, without specifying the mode (spoken or written), the compilation of the two modes available in Sketch Engine will still be much smaller in the COCA. Additionally, it was thought that a focus on the BNC, which includes British English, would suit the purpose of the study better as learners under investigation are taught British English textbooks, and would be expected to produce the same variety. Relevant findings to this corpus choice will be addressed accordingly with the study limitations. Another reason for not using the COCA on the BYU is that this software does not feature all the association measurements. It relies on the use of the MI score and the raw frequency. This is in addition to the user-friendly interface of the Sketch Engine compared to the BYU, especially when featuring the collocation lists. Sketch Engine is "a leading tool" in corpus research and lexicography as indicated by Kilgarriff and Kosem (ibid.:8). For these reasons, and as which English variety Saudi learners may use is not one of the study's research questions, it was decided to focus only on the BNC as a reference. Therefore, Sketch Engine software works best for this study.

In this study as the frequency-based approach is used in order to identify collocations produced by Saudi learners, it is applied in two steps. The first step is to check those candidate VAN combinations in the BNC according to their number of citations and conventionality following researchers such as Nesselhauf (2003), Siynaova and Schmitt (2008), and Parkinson (2015). This means that a collocation should be found in the corpus co-occurring five times or more as a threshold score and conveying the same meaning as produced by the learners. The single occurrence does not mean that it should be ignored, but the repeated events make the collocation more noticeable and significant. Therefore, the “minimum” language pattern is two occurrences according to Sinclair (2004:28).

When starting the search using the Sketch Engine, the corpus (the BNC) is first chosen. Then the node, that is the noun of the extracted candidate VAN combination, is entered in the ‘lemma’ option, and the part of speech (PoS) option is noun, to make the concordance. Subsequently, the positive filter entering the collocate lemma in the (query types) option, selecting the collocate (PoS) i.e. verb, adjective or noun, is chosen along with the span number (-5/ +5) to make concordance lines for the extracted candidate VAN combination. If a more sophisticated search needs to be built and conducted, a corps query language search (CQL) option can be used (Kilgarriff and Kosem, 2012:4). However, this was not necessary here because the single options available on the Sketch Engine interface is satisfactory for this study. Studies concerned with adjacent collocations and Multiwords Units (MWU) can start the search through using the n-gram option, as a bi-gram or tri-gram. The n-gram is also useful in lexical bundles and clusters studies, where attached sequences is an apparent feature. It is not applicable in this study because the collocations under investigation are not necessarily continuous sequences. They rather appear within a window of five words whether next to each other, nearer or further. Even though Cheng et al. (2006:412) identified the concgram to overcome this n-gram default in searching non-contiguous collocations, still this study considers filtering the search in the way mentioned above is the most suitable. As this study is not investigating specific collocational patterns such as contiguous, however, it is analysing the collocation production of learners in their free writing and in different parts of speech.

After the first step, and before searching manually through the lines of the concordance for the collocation availability and conventionality, a random sample was created. This is to ensure that the five co-occurrences are random and from different texts in the corpus as well as to lessen the number of the concordance lines. For example, in the extracted candidate VAN combination *buy coffee*, the lemma *coffee* was entered and 'noun' as PoS was chosen, and then the results were filtered by entering the verb *buy* and the number of the word-span (-5/ +5). After generating the concordance lines, the sample option was selected to create 1000 lines of random sampling. The resulting lines were searched for manually for five citations with a similar use and meaning of the candidate collocation *buy coffee* such as in "*buy some coffee"*, such as *buy you a coffee, bought us a coffee, buy coffee, buy him a cup of coffee,* and *buy a cup of coffee*. Examples like *buy you a paper at coffee time* and *You never drank this coffee! You're always buying things and then wanting Oh!* were not counted because they are not semantically related to the student's collocation. This example shows the importance of this step of verifying candidate VAN combinations extracted from learners' written texts, and to find whether or not produced combinations by learners are conventional and truly exist in the language of native speakers before looking at their fixedness or idiomaticity according to strength measurements. It is further a beneficial tool to be used to identify collocations produced by learners along with the association measurements because there could be some occurrences such as the candidate VAN combination *take shots*. This combination produced by a learner can be identified straight away when looking at its LogDice score that is 5.7, and we can say it is a fixed collocation. However, when looking at the context the learner wrote it in, she meant by *take shots*, having a medical injection, whereas *take shots* in the concordance lines belong to meanings such as gun shots, game shots and shots of drinks. The collocation occurs frequently in the BNC 237 times; nonetheless, there are less than five co-occurrences for the meaning indicated by the learner. Therefore, searching through the concordance lines for the meaning conveyed by learners' production of those combinations first and before looking at their LogDice score is essential. Finding five citations with similar use and meaning of the collocates is an appropriate threshold as it is in line with previous studies such as Nesselhauf (2003) and Parkinson (2015). The same procedure was followed for all candidate VAN combinations extracted from learners' texts in these types; verb-noun, noun-verb, adjective-noun and noun-noun collocations. Those that appeared five times or more indicating sufficient conventionality, were tested for their strength using the LogDice measure, whereas the candidate VAN combinations that appeared less than five times or did not score in the LogDice, were referred to the native-speaker informants.

In order to judge collocations' fixedness, the Sketch Engine software was also used. For example, the lemma *coffee* as 'noun' in the PoS option was entered as in the first step above. After the concordance lines are made, the 'collocations' option was chosen. A collocation candidates table appeared, where the lempos (lowercase) in the range -5/ +5 with a minimum frequency of five in the corpus, sorted by the LogDice was selected. As lempos refers to lemma and part of speech, it was chosen because this study is investigating the appearance of lemmas of words within syntactic frames, and part of speech option is relevant, especially, when words from two different categories share the same form e.g. *place* has the same form as a verb and a noun. Lowercase is chosen to exclude any proper nouns such as *house* in *the White House*, or any particular and specific reference. When the results appear, the collocate was searched for manually through the collocation list, and in this case, it is the verb *buy*. It scores 5.8 on the LogDice, which makes it a fixed collocation. Candidate VAN combinations extracted from learners' written texts need to achieve a score over a zero (positive value) to be considered a fixed collocation (Rychly: 2008).

Although LogDice has many advantages and is preferred in this study rather than other measures (Chapter Two, Section 2.4.1, and Chapter Three, Section 3.2), it has a limitation. Whilst the MI score and *t*-score do have a cut-off point to judge a collocation's strength, the LogDice does not. However, the LogDice considers zero as insignificant and has a maximum value of 14, meaning, the closer a collocation's score is to 14, the stronger it is. Given this study's objectives, which are to compare the production of two levels of language learners and three different types of collocations, a LogDice scale was devised. The scale of rating collocations that is used in this study, will verify collocations according to three levels of scores: low, medium and high. Previous researchers, such as Ellis et al. (2008) and Granger and Bestgen (2014) have established a band of threshold scores in order to identify fixed collocations at three levels low, medium and high. However, these thresholds were different between researchers. Ellis et al. (2008) had a band of 3.3/ 6.7/ 11 for low, medium and high respectively whereas Granger and Bestgen (2014) had non-collocation for 3<, low collocation for 3-4.99, medium collocation for 5-6.99, and high collocation for 7 and more.

However, due to the use of the LogDice in this study, which is different to the MI score used in previous studies, a proposed division of a LogDice scale is generated for the purpose of this study. The low level collocations will have a score from above zero, as zero and below show insignificant values i.e. non-collocations, and the scale will stop at 14, as it is the maximum value. Thus, dividing the scale into levels makes collocations with scores of between 0.1-3.5 as low, 3.6-7 as medium, 7.1-10.5 as high, and 10.6-14 as advanced. As the aim of this division is to ease the process of classification of Saudi learners' produced collocation, taking into consideration their relatively low-level of proficiency, the scale applied will use the scoring between zero and 10.5. However, any collocation produced by the learners above 10.5 will be mentioned and indicated as an exceptional instance produced by them. Rychly (2008: 9) writes that the maximum value in the LogDice i.e. 14 shows that collocates are almost co-occurring only with each other, meaning they are low frequency collocations and maybe rare collocations, which are not commonly produced by learners (Siyanova and Schmitt, 2008; Durrant and Schmitt, 2009; Fernández and Schmitt, 2015). Rychly (2008: 9) further asserts that the value of most collocations produced by learners is usually less than 10. This proposed division, additionally, will assist in comparing collocations produced by learners from the two proficiency levels to show whether the learners' levels are correlated to low, medium or high level collocations. For example, a pre-intermediate student and another intermediate wrote about the good weather they were enjoying on their trips by using two different expressions. The pre-intermediate level student produced *fantastic weather* whereas the intermediate level student produced *windy weather*. *Fantastic weather* scores 4.3 on the LogDice, which makes a medium level collocation on the scale while *windy weather* scores 7.4 on the LogDice making it a high level collocation on the scale. The use of this scale adds further value to the use of the association measurement LogDice by indicating and classifying differences or similarities between produced collocations and the levels of learners rather than just determining their production of collocations and non-collocations as less idiomatic combinations. It is an effective method to trace L2 development in learners' production of collocations to investigate their production across a scale of an association measurement as indicated by Durrant and Schmitt (2009:168) and Granger (2018:233).

Finally, after identifying each groups' production of acceptable collocations and their level of fixedness in the BNC using the number of citations in the corpus and LogDice score through the Sketch Engine tool, descriptive statistics will be used to describe and compare the two levels' results. It should be noted that repeated examples of collocations, when produced in one text, are counted as one example. However, when the same collocation occurs in another text, it is counted as another example such as *the* *worst vacation*. No matter how many times a single student wrote it in her text, it is counted as one example, whereas it is counted again as it occurs in other texts. This is due to the study interest in investigating which collocations are produced by learners rather than how frequent they use them.

4.5.3 Native-speaker informants' judgement

Some researchers such as Kuo (2009) and Laufer and Waldman (2011) employed native speakers to identify collocations produced by learners in addition to the use of, respectively, the BNC and collocation dictionaries, as references. However, they did not refer to association measurements as their main focus was erroneous collocations. The case, in this study, is different as it is investigating learners' production of collocation according to their fixedness by using statistical measurements i.e. the LogDice available in corpus tools. Thus, the use of native speakers comes as a third step in the analysis, and concerns only candidate VAN combinations extracted from the texts which were not identified in the BNC. They are those that did not reach the threshold score - five times in the BNC – and/ or scored a zero or less on LogDice. Even though corpora contain a large collection of native-speaker language and are beneficial in linguistic studies, they may not be completely representative of all native speakers' language. This is especially evident when investigating the production of L2 learners, whose language may include errors and mistakes and not be up to the level of native speakers' language. To overcome this issue, native-speaker informants' judgement for combinations not identified were sought.

This procedure is based on previous work by researchers such as Nesselhauf (2003) and Parkinson (2015). Despite the limitations in this approach, this sort of L2 language that is not found in corpora, can be judged by the intuitions of native speakers. These judgements however, will be subjective depending on the individuals' language knowledge and opinions. Additionally, this approach cannot provide any information about a specific collocation's strength and fixedness. The phraseological approach is merely showing if these combinations can be used, produced, and comprehended by native speakers or not. While the evidence taken from corpora is usually based on a random, various and a good number of texts, the evidence given by the native speakers is ought to be less in number, frequency and variations. The same native speakers, who are limited in number being participants, judge a given list of collocations. The application of this approach is mostly similar to Parkinson's. As described in Section 3.2, Nesselhauf's interpretations of native-speaker informants' responses were trying to classify the type of collocations according to restrictedness, which was not a very clear nor easy procedure. Parkinson interpreted the native-speaker informants' responses only according to acceptability. This study will follow Parkinson's use of this approach to judge acceptability only with no reference to restrictedness as did Nesselhauf. Those combinations, which native speaker informants agree on will be referred to as acceptable collocations while the remaining list of them, which native speaker informants disagree on or leave them undecided, will be referred at as the less idiomatic combinations.

To obtain these judgments, 20 British English native speakers were engaged as native-speaker informants, in this stage of the data analysis. They were volunteer undergraduate students at the University of Leeds, from the School of Education, majoring in English. The participants are aware of and knowledgeable about collocation as they have already covered this topic in a module that covers analysing texts according to lexical patterns such as collocation and idiom. Thus, they are found to make reliable candidates to judge the collocations. Before using this judgement with the undergraduate information, it was trialled with a native-speaker postgraduate student, whose judgement was also used. After the pilot, the collocations were divided into two lists to make it easier for the native-speaker informants, an example for each collocation was added, and a short instruction about the task was given. The informants preferred to have the information written even though they were given a brief about the study and task orally. Examples for all given collocations in the judgement list were taken from the participant learners' written texts. These examples were also proofed for grammatical and spelling mistakes to ensure the native-speaker informants focused on the task of judging collocations.

This judgement task began by giving the native-speaker informants a brief about the study and nature of the task (see Appendix Four for the judgement form). In the task, the informants were asked to judge each collocation decide if it was an acceptable or non-acceptable collocation. If they did not have a precise response, they would mark the collocation as 'not sure'. The task also provided an option for the informants to write down any alternatives or suggestions for a given collocation. The given example for each collocation was provided within a context to illustrate the representation of the five-word span between collocates, which was used in the study. The native-speaker informants spent about ten to 12 minutes to complete this task. Eventually, five native-speaker informants evaluated each collocation, which is considered to be a systematic procedure following the number of citations a collocation should appeared in the BNC.

Following this stage, the judgement responses from the native speakers were entered into an Excel spreadsheet for analysis. As described in Chapter Three, Nesselhauf (2003) referred to a scale of three informants' approvals, and Parkinson (2015) referred to a scale of 11 informants' approvals. Both of them used an odd number of informants to ease the process of reflecting informants' opinions. For example, even though the collocation *forgot passport* scores 2.9 on LogDice, it appears only three times in the BNC. Therefore, it was not considered a collocation as the frequent co-occurrence is also an important criterion besides exclusivity. By referring this candidate VAN combination *forgot passport* to native-speaker informants, four native-speaker informants agreed it was an acceptable collocation while one native-speaker informant did not. This shows that more than three responses considered this to be a collocation. In the same way, if the three responses were to find the collocation non-acceptable, then the candidate VAN combination was found to be non-acceptable and as a result was not considered to be a collocation for the purposes of this study. If three responses were marked as ‘unsure’, then the candidate VAN combination was considered as undecided as well as when it received an equal number of responses in two categories of the three options: acceptable, non-acceptable and unsure. For example, if two native-speaker informants find the collocation to be acceptable whereas two native-speaker informants mark it as ‘unsure’, and the last native-speaker informant may find it non-acceptable. As a judgement on a candidate VAN combination by native-speaker informants is not clear, it would be considered as undecided.

This analytical procedure, consisting of extracting and identifying VAN combinations from learners’ written texts, allows us to investigate to what extent those learners are able to produce acceptable and fixed collocations, and identify what types of collocations. As the same procedure is applied to both levels of learners, i.e. pre-intermediate and intermediate, it further allows the comparison between them as proposed in the third research question. However, similarities and differences which would occur between the two levels may possibly be related to a variety of issues, which were explored in the literature review chapters upon relevant studies outcomes. Whereas those issues are not investigated in depth or thoroughly in this study, there were places – in the examples of less idiomatic combinations, specifically - where it had to be referred to, as the discussion chapter will show. Therefore, and to support the discussion of such instances, the following tools were referred to: the Arabic-English dictionary for the Use of Students (Hava, 1915), the Concise Arabic-English Lexicon of Verbs in Context (Abdou and Hassanein, 2011), and the Oxford Arabic Dictionary (Arts, 2014), when discussing collocations that possibly were influenced by the learners' L1.

4.6 Summary

This chapter has explained why the study adopted the three approaches in a triangulation methodology. The CIA approach, the frequency-based approach, and the cross-sectional approach were used to answer the three research questions and ensure methodological credibility. This was important as this analysis of the Saudi learners' written texts was conducted manually and investigated qualitatively, specifically when comparing the two levels of learners. There is always an advantage for the researcher to combine the qualitative and quantitative approaches (McEnery and Wilson, 1996). This is applied in the study through using the statistical tests for association measurements of collocations.

The context description reflects on the need for this study to apply the analytical procedure followed, starting from the extraction procedure of candidate VAN combinations. As it was decided to do this extraction manually, a systematic procedure as detailed in the chapter had to be used. While this process includes some limits i.e. less accuracy than electronic software programme, it still has much to contribute to the validity of this study. Not many studies have investigated learners' written texts manually, and if so, they have focused on adjacent collocations or just one type of collocation. In the method applied in this study, the aim is to contribute to the literature by examining different types of collocations co-occurring in a wider span than adjacent collocates.

The frequency-based approach is applied following previous studies to identify collocations under investigation. There was a need to use this approach in order to identify collocations according to their exclusivity in a statistical way, which is viewed as being more reliable and replicable than qualitative approaches. The phraseological approach has been criticized for its intuitive and subjective judgement of collocations, which can be a poor guide (Hunston, 2002), especially in a similar case like this study investigating acceptable collocations considering their level of fixedness. However, this phraseological approach was applied as an alternative when candidate VAN combinations could not be determined using the corpus for reference. Referring to the BNC is also employed in a double-procedure to identify collocations produced by learners. The first step is to find whether or not candidate VAN combinations extracted from learners' written texts were conventional and occurred in native-speakers' language though an examination of their frequency. The second step is to identify those candidate VAN combinations' fixedness through the LogDice measure on the corpus tool. The application of five occurrences to decide collocations' existence in the BNC is found widely in previous research whereas the use of the LogDice measurement is not. However, the LogDice was found the most appropriate association measurement to be applied for this study and context as has been explained in Chapters Two, Three and Four, Sections 2.4.1, 3.2 and 4.5.2. It was further supported by the argument raised by Fernández and Schmitt (2015), and Gablasova et al. (2017a) for its suitability in language learning research (LLR).

These three steps of the analytical procedure allow the interpretation and comparison of the results of the two Saudi learner levels' production of collocation. This is to show the problematic areas and possible similarities and differences qualitatively as the data is from a relatively small number of texts and needs close exploration. Chapter Five considers details this methodology in some detail by describing it through two case studies and then shows the analysis of all data investigated regarding the Saudi learners in their two levels (pre-intermediate and intermediate).

Chapter 5 Data Analysis

5.1 Introduction

This chapter outlines the analytical procedure discussed in the previous methodology chapter. This will be done with reference to two case studies. The first is an example of a piece of lower-level writing (pre-intermediate), and the second is of a piece of higher-level writing (intermediate). The purpose of presenting these case studies is to illustrate the process of extraction and identification of the collocations. This process was followed for all of the data for the two levels of Saudi learners. After the two case studies, the remaining parts of the chapter discuss separately the analysis of each level of the Saudi learners (see all writing samples in Appendix Five and Six).

5.2 Case study 1

5.2.1 Extracting candidate VAN combinations from a lower-level writing text.

An essay written by a pre-intermediate-level learner was analysed. It consists of 200 words on the topic of “The Worst Vacation”. This essay and the rest of the essays are included in Appendix Five. First, to analyse the text and extract candidate VAN combinations, each sentence was broken up into clauses. The starting sentence of this text, *'Life has a lot of experience something good and something bad.'* is a clause as it is, so it is not divided:

1 Life has a lot of experience something good and something bad

In each clause, as nouns are the nodes of the candidate VAN combinations, all the nouns were underlined and then any verbs, adjectives, and nouns collocating with the noun node were identified (Section 4.5.1). All identified candidate collocates are within the five-word span appearing syntactically before and after the noun node. For example, the noun *life* forms a candidate VAN combination with the verb *has*, and the noun *experience* also forms another with the verb *has*. Both nouns make candidate VAN combinations of the noun-verb and verb-noun types. As quantifiers are not included, as indicated in Chapter Four, Section 4.5.1, *a lot of* is not considered in any of the extracted combinations. The noun *something* collocates with the adjectives *good* and *bad* respectively in adjective-noun type of VAN combinations.

2 Whe [when] the experines [experience] was good,
3 we would feel happy.
4 But If the experience was bad,
5 we would feel sad.

Of these four clauses, (2) and (4) include the same noun *experience*, which collocates with the adjective *good* and then with the adjective *bad*, forming adjective-noun from VAN combinations. Clauses (3) and (5) are not marked for any candidate VAN combination as they contain no nouns.

6 Before 10 years ago I visited my uncle with my mother
7 because he was very sick
8 and the doctors told him never leaved [left] his bed.
9 First, I played with his son.
10 We started run
11 and played football.
12 Then, I sleppted [slipped] on a slant,
13 when I was running.

Clause (6) includes the nouns *years*, *uncle,* and *mother*. While *years* and *mother* do not make any candidate VAN combinations within their span, *uncle* makes a candidate verb-noun combination with the verb *visited* being its collocate. Also, the nouns *doctors, bed* and *football* in clauses (8) and (11) make candidate combinations of noun-verb and verb-noun types as follows: *doctors told, leave bed* and *played football*. The two nouns *son* and *slant* in clauses (9) and (12) do not create any candidate VAN combinations with any surrounding collocates as they are parts of prepositional phrases, which are excluded in this study (Section 4.5.1, p. 91-92). Another example of a possible combination in the above clauses between nouns and verbs, according to the types included in this investigation, is *doctors leave*, which appears in each other's window, but not considered as a candidate VAN combination due to the syntactic criteria followed (Section 4.5.1).

14 My accident made big sound
15 because I fell on the flor [floor]
16 and my face crash to the hall [ball].

Clauses (14), (15) and (16) include the nouns *accident, sound, floor, face* and *ball*, of which three make candidate noun-verb and verb-noun combinations as in *accident made, face crash*, and *made sound*. The noun *sound* also collocates with the adjective *big*, forming a candidate adjective-noun combination. While the adjective *big* further appears in the span of the noun *accident* within a clause boundary, they do not form a candidate VAN combination because according to the syntactic level, the adjective *big* modifies the noun *sound* not the noun *accident* (Section 4.5.1, p. 94). The nouns *floor* and *ball* do not belong to any candidate VAN combination as they belong to prepositional phrases.

17 I was standing,
18 when my uncle leaved [left] his bed and came to know the problem!
19 he said:
20 What's the matter?
21 We answered: nothing.

In clause (18), four candidate VAN combinations can be extracted. Two are noun-verb type; *uncle left* and *uncle came*, and the other two are verb-noun types; *left bed* and *know problem*. The two verbs *came* and *know* do not make candidate VAN combinations with the noun *bed* because they are not directly syntactically related according to the criteria followed in this analysis, unlike the case of the nouns *uncle* and *problem*. Clause (20) contains the noun *matter* which collocates with the verb *is* but is not extracted as a VAN combination because the verb *be* is not included in this study as explained in Chapter Four, Section 4.5.1.

22 But he loocked [looked] to my face and said
23 You should play in the room.
24 Firstly, I couldn't understand
25 how could he knew.
26 Then, I saw his son was looking to my face and smiled.

Clauses (22), (23) and (26) include the nouns *face*, *room,* and *face*, which are not counted as candidate VAN combinations as they appear alone within prepositional phrases. The noun *son* in clause (26) makes two candidate VAN combinations as a verb-noun type in *saw son,* and noun-verb type in *son looking*.

27 At last, I saw my face in a meror [mirror].
28 I felt shy because of my face
29 I was wite [white]!! The pouder [powder] of ball on my face.

Two nouns are identified in clause (27), which are *face* and *mirror,* where one of them makes a candidate verb-noun combination as in *saw face*. The noun *face* in clause (28) does not form any candidate VAN combination due to being a part of a prepositional phrase. Clause (29) has three nouns *ball*, *powder* and *face*, but none of them form a candidate VAN combination.

30 After this memory I never run on a slant.

31 I learned the lesson

32 and I must be quite [quiet]

33 when I visit sick person.

Clauses (31) and (33) make two candidate verb-noun combinations: *learned lesson* and *visit person*. *Memory* and *slant* are nouns in clause (30) that do not meet the criteria of the analysis to make candidate VAN combinations. The last clause (33) includes another candidate combination: *sick person*, an adjective noun collocation.

The above section outlines how candidate VAN combinations were extracted from learners' texts, as the first step in the analytical process. The following section discusses the second step in the analysis: identifying these extracted VAN combinations in the BNC. All extracted candidate VAN combinations from this text are listed in Appendix Seven according to their types.

5.2.2 Identifying extracted candidate VAN combinations in the BNC from the lower-level writing texts

All candidate VAN combinations extracted from the text are searched for in the BNC using Sketch Engine (https://old.sketchengine.co.uk/auth/corpora/), according to the methodology given in Chapter Four, Section 4.5.2. This is carried out in two stages in order to confirm whether or not the collocations produced by Saudi learners are acceptable. To be considered an acceptable collocation, candidate VAN combinations should, firstly, reach the assigned threshold score (Chapter Four, Section 4.5.2) of five citations or more in the BNC to show frequent occurrences and conventional availability according to the meaning used by the learners. Secondly, they should score more than a zero on the LogDice. After that, all identified collocations are classified according to their level of fixedness as low, medium and high level collocations (Chapter Four, Section 4.5.2) in order to show to what extent learners produce acceptable and fixed collocations and to show possible similarities and differences between the two levels of learners.

This part of the analysis is organised in Table 5.1 to include information about candidate combination types, conventionality and fixedness for ease of reference and discussion.

Table 5.1: Collocations identified from the first case study text with their types and scores in the BNC and LogDice scores

Candidate Combinations	**Combination Type**	**Citation in the BNC**	**LogDice Score**	**Acceptable Collocation**	**Level of Fixedness**
Life has	Noun-verb	≥5	7.5	Yes	High
Has experience	Verb-noun	≥5	6.5	Yes	Medium
Visited uncle	Verb-noun	≥5	4.8	Yes	Medium
Doctors told	Noun-verb	≥5	6.8	Yes	Medium
Left bed	Verb-noun	≥5	5.8	Yes	Medium
Played football	Verb-noun	≥5	8.7	Yes	High
Accident made	Noun-verb	≥5	3.6	Yes	Medium
Made sound	Verb-noun	≥5	6.7	Yes	Medium
Face crash	Noun-verb	<5	NA	NA	NA

Candidate Combinations	**Combination Type**	**Citation in the BNC**	**LogDice Score**	**Acceptable Collocation**	**Level of Fixedness**
Uncle left	Noun-verb	≥5	2.7	Yes	Low
Uncle came	Noun-verb	≥5	3.0	Yes	Low
Know problem	Verb-noun	≥5	6.2	Yes	Medium
Saw son	Verb-noun	≥5	5.1	Yes	Medium
Son looking	Noun-verb	≥5	5.2	Yes	Medium
Saw face	Verb-noun	≥5	7.6	Yes	High
Learned lesson	Verb-noun	≥5	9.6	Yes	High
Visit person	Verb-noun	≥5	5.0	Yes	Medium
Something good	Adjective-noun	≥5	6.4	Yes	Medium
Something bad	Adjective-noun	≥5	5.7	Yes	Medium
Experience good	Adjective-noun	≥5	5.8	Yes	Medium
Experience bad	Adjective-noun	≥5	6.4	Yes	Medium
Big sound	Adjective-noun	≥5	4.9	Yes	Medium
Sick person	Adjective-noun	≥5	5.6	Yes	Medium

Table 5.1 shows all candidate VAN combinations extracted from this learner's text with their types, and whether or not they reached the threshold score in the BNC. One combination did not meet the threshold score from the noun-verb type: *face crash*. In this case this VAN combination was not frequent enough, i.e. less than five times in the BNC, and did not show conventionality with the learner's use. For example, *face crash* appeared in the learner's text as in "*my face crash into the ball"* as a subject to the verb crash. This combination appeared in the BNC nine times, but only two citations can be considered similar such as in "*a conflagration of raw heat crashed into her face*" and "*when I crashed down onto my face*". Also, *face* in these two citations, is in a prepositional phrase, which does not fall into the criteria used in the study. The rest of the citations either had the two words in two different sentences or clauses, or had a different meaning from the learner's as in *mountain face*, *a crashing axe* and *crashing face*. Therefore, it was considered as a non-acceptable collocation, and LogDice was not used (Section 4.5.3). It was, however, referred to the native speaker informants for their judgement.

The rest of the candidate combinations that reached or exceeded the threshold score were checked for their LogDice score using Sketch Engine online software, and then according to the proposed classification given in the methodology chapter, identified collocations were listed as in the table. Most of the identified collocations were medium level collocations, which means that they scored between 3.6-7.0 on LogDice, whereas only two collocations were low level collocations scoring between 0.1-3.5. Four collocations were found to be high level collocations scoring between 7.1-10.5. The table also shows that all adjective-noun combinations were identified as fixed and of medium level. All verb-noun collocations were also identified as fixed and found between the two classifications of medium and high level. Noun-verb combinations were different as one of them was not identified, and the identified collocations were low, medium and high level.

5.2.3 Unidentified VAN combinations from the lower-level writing text.

Combinations such as *face crash* were not identified in the BNC, were referred to the judgement of native speaker informants as a last procedure for identification. However, this procedure is limited to only verifying whether or not the combinations were accepted as collocations by English native speakers, and does not produce further statistical information that indicates whether or not the collocations are fixed, unlike when the corpus-based approach was used.

When *face crash* was referred to native speaker informants, it was judged as an non-acceptable collocation as all native speaker informants agreed it was an non-acceptable collocation and did not provide any further corrections or suggestions for an alternative acceptable collocation. However, the collocation list in the BNC which I obtained using the collocations tool in Sketch Engine shows some fixed collocations, which could convey the same meaning of the combination produced by the learner but with greater idiomaticity. For example, *face hit* and *face smash* are found with LogDice scores 6.4 and 5.2 respectively.

5.3 Case study 2

5.3.1 Extracting candidate VAN combinations from a higher-level writing text

The second essay analysed in this section was written by a higher-level learner (intermediate). It is a 200-word essay on the topic of “The Advantages and Disadvantages of Living in a Village”. The same analytical procedure given in the first case study to extract candidate VAN combinations from the text was followed. This process, as detailed above, starts by dividing sentences into clauses, then looking for noun nodes and their collocates from verbs, adjectives, and nouns within the five-word span, and extracting those candidate VAN combinations according to their syntactic relations. The analysis of this case study is presented in a table (Table 5.2), rather than using detailed explanations, as in the previous section in order to avoid repetition. The table includes all clauses found in this written text. All essays written by the two levels' learners, including those of the two case studies, are included in Appendix Six. As previously explained in Chapter Four, Section 4.5.2, that repeated examples of extracted candidate VAN combinations are counted as one instance when appearing in the same text such as *small village* in the table below.

Table 5.2: Clauses given in order from the second case study text with candidate VAN combinations extracted and their types

Clauses	**Extracted Candidate VAN Combinations**	**Type of the Extracted Combinations**
they have to think of their future and *improve* their *life*	Improve life	Verb-noun
any other thing can *make* the *air* polluted	Make air	Verb-noun
it *has* a lot of advanteges [*advantages*]	Has advantages	Verb-noun
they should not *forget* their home *town* and visit it every once while	Forget town	Verb-noun
Have a lot of *disadvantages* and a lot of advantages	Have disadvantages	Verb-noun
any other *thing* can *bother* them	Thing bother	Noun-verb
any other *thing* can *make* the air polluted	Thing make	Noun-verb
I *recommend people* to live in a big city	Recommend people	Verb-noun

Clauses	**Extracted Candidate VAN Combinations**	**Type of the Extracted Combinations**
I recommend people to live in a *big city*	Big city	Adjective-noun
they should not *forget* their *home* town and visit it every once while	Forget home	Verb-noun
they should not forget their *home town* and visit it every once while	Home town	Noun-noun
people can have a pure and fresh air people can have a very calme [calm] and quite [quiet] live [life] with city nois [noise]	People have	Noun-verb
people can have a very calme [*calm*] and quite [quiet] live [*life*] with city nois [noise]	Calm life	Adjective-noun
people can have a very calme [calm] and quite [*quiet*] live [*life*] with city nois [noise]	Quiet life	Adjective-noun
people can have a very calme [calm] and quite [quiet] live [life] with city nois [noise]	City noise	Noun-noun
So many people still live in *small villages* there are so many disadvantages of living in a *small village* There are a few advanteges [advantages] of living in *small villages* So many people still living in *small villages* for some reson [reason]	Small villages	Adjective-noun
people will not be *able* to be more cefictecaded [sophisticated] or open mind	People able	Adjective-noun
people will not be able to be more cefictecaded [sophisticated] or *open mind*	Open mind	Adjective-noun

Clauses	**Extracted Candidate VAN Combinations**	**Type of the Extracted Combinations**
there is no *important* fucllitys [*facilities*] such as hospitals with good doctors	Important facilities	Adjective-noun
there is no important fucllitys [facilities] such as hospitals with *good doctors*	Good doctors	Adjective-noun
people can *have* a pure and fresh *air*	Have air	Verb-noun
people can have a *pure* and fresh *air*	Pure air	Adjective-noun
people can have a pure and *fresh air*	Fresh air	Adjective-noun
any other thing can make the *air polluted*	Air polluted	Adjective-noun
they still stick with their tardtions [*traditions*] and *customs*	Traditions customs	Noun-noun
there is no *schools* or univircities [*universities*]	Schools universities	Noun-noun
Have a lot of *disadvantages* and a lot of *advantages*	Advantages disadvantages	Noun-noun

All candidate VAN combinations were extracted following the same procedure explained previously. However, this text includes an example of an extraction procedure, where the combination includes more than two words and, thus, can be extracted in two ways. Although this study is limited to investigating two-word collocations, some combinations occurred in the data including more than two words. Therefore, and for the reasons discussed in Chapter Four, Section 4.5.2, the following method was used: for example, *forget home town* is a three-word combination, which was extracted as three candidate combinations *forget home*, *forget town* and *home town*. According to the meaning of the combination, this one can be extracted as two combinations. Nonetheless, there are some other instances where the meaning of the combination cannot be split; therefore, only one combination is extracted. Such examples will be noted in the relevant sections.

5.3.2 Identifying extracted candidate VAN combinations in the BNC from the higher-level writing text

After extracting all candidate VAN combinations, they are searched for in the BNC using Sketch Engine online software to identify them using the same methods applied in the first case study. Candidate VAN combinations should reach the threshold score of five citations, or more, in the BNC for conventionality, and then should score more than a zero on the LogDice for fixedness. They are organized in Table 5.3 along with information about their types, citations and fixedness.

Table 5.3: Collocations identified from the second case study text with their types and scores in the BNC and LogDice scores

Candidate Combinations	**Combination Type**	**Citation in the BNC**	**LogDice Score**	**Acceptable Collocation**	**Level of Fixednes s**
Improve life	Verb-noun	≥5	6.8	Yes	Medium
People have	Noun-verb	≥5	8.5	Yes	High
Thing make	Noun-verb	≥5	7.8	Yes	High
Make air	Verb-noun	≥5	4.9	Yes	Medium
Have air	Verb-noun	≥5	4.9	Yes	Medium
Thing bother	Noun-verb	≥5	4.9	Yes	Medium
Recommend people	Verb-noun	≥5	3.5	Yes	Low

Candidate Combinations	**Combination Type**	**Citation in the BNC**	**LogDice Score**	**Acceptable Collocation**	**Level of Fixednes s**
Has advantages	Verb-noun	≥5	5.8	Yes	Medium
Forget town	Verb-noun	<5	NA	NA	NA
Forget home	Verb-noun	≥5	4.9	Yes	Medium
Have disadvantages	Verb-noun	≥5	3.0	Yes	Low
Small villages	Adjective-noun	≥5	7.7	Yes	High
People able	Adjective-noun	≥5	6.8	Yes	Medium
Open mind	Adjective-noun	≥5	7.2	Yes	High
Important facilities	Adjective-noun	≥5	4.4	Yes	Medium
Good doctors	Adjective-noun	≥5	5.4	Yes	Medium
Pure air	Adjective-noun	≥5	5.7	Yes	Medium
Fresh air	Adjective-noun	≥5	9.6	Yes	High
Air polluted	Adjective-noun	≥5	6.0	Yes	Medium
Calm life	Adjective-noun	≥5	2.1	Yes	Low
Quiet life	Adjective-noun	≥5	6.2	Yes	Medium

Candidate Combinations	**Combination Type**	**Citation in the BNC**	**LogDice Score**	**Acceptable Collocation**	**Level of Fixednes s**
Big city	Adjective-noun	≥5	7.6	Yes	High
Traditions customs	Noun-noun	≥5	7.3	Yes	High
Schools universities	Noun-noun	≥5	7.4	Yes	High
City noise	Noun-noun	≥5	3.6	Yes	Medium
Home town	Noun-noun	≥5	7.4	Yes	High
Disadvantages advantages	Noun-noun	≥5	9.4	Yes	High

Table 5.3 shows that all candidate VAN combinations extracted from this learner’s text were identified as acceptable collocations except for *forget town*, which did not reach the threshold score. This combination occurs ten times in the BNC, but only on two occasions is it similar to the learner’s use as in “*But Ivan, who died aged 84, did not forget Minehead, the Somerset town*” and “*Merchant princes did not forget their native towns*”. The rest of the citations were either in separate sentences or clauses, or where the noun *town* and the verb *forget* are not syntactically related.

All collocations identified from the adjective-noun and noun-noun types were found to be acceptable collocations from all three levels: low, medium and high. Noun-verb collocations were all also identified as acceptable collocations appearing in two levels, medium and high whereas the verb-noun collocations identified which were found to be acceptable collocations were only found to be low and medium level.

5.3.3 Unidentified VAN combinations from the higher-level writing text

The only candidate VAN combination that was not identified from this text is *forget town*, and thus was referred to the judgement of native speaker informants. Two native speaker informants marked it as an acceptable collocation while three native speaker informants thought it was unacceptable with no further corrections or suggestions for alternative acceptable collocations. Therefore, this combination was considered as a non-acceptable collocation.

5.4 Analysis by level

After describing the analytical procedures of extracting and identifying collocations in depth through the two case studies, the following sections will review the analysis of each level of learners separately. All candidate VAN combinations extracted from the two levels' written texts are organized in tables as given in the example of Case Study Two, and listed in the Appendix Seven and Eight. The remaining sections of the chapter will focus on the second and third step of the analysis: collocations identified using both the corpus and native speaker informants.

5.4.1 Pre-intermediate level analysis

This section analyses the lower-level Saudi learners' written texts (pre-intermediate), and is divided according to the type of collocations identified. Collocations that were not identified using the BNC and were referred to native speaker informants, will be discussed along with the less idiomatic combinations, in the last part of this section.

5.4.1.1 Verb-noun collocations identified in pre-intermediate-level writing

A total of 84 candidate verb-noun combinations were extracted of which 85.7% were found to be acceptable collocations. Table 5.4 gives a random selection covering the range of scores of these collocations identified for this analysis while the full list of 72 verb-noun collocations identified from this level's texts, are in Appendix Nine. Forty-nine of the collocations identified were of medium level scoring between 3.6–7.0 on LogDice such as: *heard knock, buy coffee, called doctor, buy things, take pictures* and *have space*. Only six of the identified collocations were low level, scoring between 0.1-3.5, *doing exams, find taxi, saw fountain, saw dress, saw snake* and *get flu*. Among the 17 high level collocations identified, scoring between 7.1-10.5, are *found people, visited friend, called police, ate lunch, take care, ask people, told story* and *spent day*. One of the high level collocations produced by this level of learners, *learned lesson*, was found to be the top verb collocate for the noun *lesson*, which scored 9.6 on the LogDice (Section 4.5.2).

Table 5.4 gives examples of those collocations identified in the contexts in which they appeared in from learners' texts with their citation and score numbers.

Table 5.4: Sample of acceptable verb-noun collocations identified from pre-intermediate-level texts with context and LogDice scores

Verb-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Spent vacation	I *spent* that *vacation* with my books	≥5	4.0
Tell parents	how to *tell* my *parents*	≥5	6.5
finished exam	I *finished* my *exams*	≥5	4.6
Had lunch	After we *had* our *lunch*	≥5	4.5
got information	we *got* some *information* about that	≥5	6.5

Verb-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
saw window	we *saw* our *windows* open!	≥5	6.6
asked service	he *asked* the *room* serves [*service*] to clean our rooms	≥5	5.4
Asked security	we *asked* a securty [*security*] gured [guard]	≥5	4.1
Miss plane	we will *miss* the *plane*	≥5	5.1
Get bags	we went to airport to *get* our *bags*	≥5	5.7
Found bags	we didn't *found* our clothes and *bags*!	≥5	4.8
Found clothes	we didn't *found* our *clothes* and bags!	≥5	5.0
Call taxi	My dad tryed [tried] to *call* a *taxi*	≥5	4.5
Lost day	We *lost* our first *day*	≥5	6.5

Some combinations, as discussed earlier, have more than two words such as *asked room service* and *asked security guard*. The table above gives two examples, where the first combination extracted was *asked service* only because *room* is a noun modifying service while the second combination could be extracted in two ways as *asked security* and *asked guard. Asked service* and *asked security* were identified in the BNC, but *asked guard* did not reach the minimum requirement of five mentions and was thus referred to native speaker informants' judgement.

This analysis also shows that learners used some nouns with different verbs, and were able to produce acceptable collocations with each of them. For example, *taxi* was found once collocating with the verb *find*, which is a low level collocation, and in the other instance collocating with *call*, which is a medium level collocation. Other examples are the nouns *lunch, exams, bags* and *people*. *Lunch* was produced in a medium level collocation as *had lunch*, and in a high level collocation as *ate lunch*. *Exams* had also two levels of collocations produced by learners as: *doing exams*, low level, and *finished exams*, medium level. *People* and *bags* were collocated with two verbs as well. *Bags* was produced as *get bags* and *found bags,* which are medium level collocations, and *people* was produced as *ask people* and *found people* which are high level.

5.4.1.2 Noun-verb collocations identified in pre-intermediate-level writing texts

A total of 38 candidate noun-verb combinations were extracted of which 84.2% of them were found to be acceptable collocations. The full list of the identified noun-verb collocations from this level's texts, which include all 32 acceptable collocations, are in Appendix Nine. Table 5.5 shows examples of these identified collocations in the contexts they appeared in from learners' texts with their citation and score numbers. Most of the collocations identified were of medium level, scoring between 3.6–7.0 such as: *father said, parent bought, bus came, brother get* and *restaurant called*. There are five high level collocations, scoring between 7.1-10.5, *life has, things happened, day came, house keep* and *father told*. Only two low level collocations, scoring between 0.1-3.5 were among the collocations identified; they are: *uncle came* and *dad tried*.

Table 5.5: Sample of acceptable noun-verb collocations identified from pre-intermediate-level texts with context and LogDice scores

Noun-Verb Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Mother asked	my *mother asked* me	≥5	6.7
Trip leave	our *trip* going to *leave* at 7.30 p.m.	≥5	3.9
Pain came	the *pain came* strong	≥5	4.5
Car work	the *car* didn't *work*	≥5	5.6

Noun-Verb Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Problem happened	there is a big *problem happened*	≥5	4.7
Policeman told	some *police man [policeman] told* as [us] who did that!	≥5	4.7
Garden have	big *garden* have a small space of fruit garden	≥5	4.8
Brother came	my big *brother came* to the Airport after 10 minutes with the ticket	≥5	5.2
Mother want	my *mother want* to back to home, Jeddah	≥5	6.0

Some of the nouns were commonly used by learners, and produced in different combinations with verbs, which were identified as acceptable collocations. For example, *father, mother, brother* and *sister* were found in collocations such as: *father said, father came, father called, father told, mother asked, mother want, mother took, brother came, brother get, sister started* and *sister know*. All of these combinations were identified as medium level collocations, except for *father told* that was identified as a high collocation.

5.4.1.3 Adjective-noun collocations identified in pre-intermediate-level writing texts

A total of 58 candidate adjective-noun combinations were extracted, of which 74.13% were found to be acceptable collocations. Most of the collocations identified, 34, were medium level, such as: *high buildings, beautiful area, big sound, good student, clean place* and *long vacation*. Eight high level collocations were identified; *bad/ worst things, deep sleep, long time, big problem, dark colours, windows open* and *whole story*. Only one low level collocation was identified, *horrible day*. The full list of 43 acceptable collocations from this adjective-noun type identified from texts from this level are in Appendix Nine. Table 5.6 gives examples with relevant information.

Table 5.6: Sample of acceptable adjective-noun collocations identified from pre-intermediate-level texts with context and LogDice scores

Adjective-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Final exams	I was doing my *final exams*	≥5	5.6
Food delicious	the *food* wasn’t *delicious*	≥5	6.4
Little sister	my *little sister* lost	≥5	5.9
Beautiful dress	I saw a *beautiful dress* in a shop	≥5	6.9
Big day	The *big day* cam [came]	≥5	6.6
Weather good	the *weather* wasn’t *good*	≥5	6.1
Good hotel	We spen [spent] our therd [third] day in a road to find some *good hotel*	≥5	5.6
Nice place	it was clean and *nice place*	≥5	6.0
People nice	*people* was very *nice*	≥5	6.1
Restaurant small	The *restaurant* was *small*	≥5	5.7

Some adjectives were common in the learners’ writing at this level such as *good, bad,* and *big*. Among the collocations identified using these adjectives are: *something good, good experience, good hotel, good student, something bad, bad experience, bad things, big sound, big day, big brother, big garden, big snake* and *big problem*. There were also some nouns that were modified with different adjectives, which still made acceptable collocations such as, *small restaurant* and *modern restaurant, nice place* and *clean place, beautiful hotel, famous hotel* and *good hotel,* and *sick brother* and *little brother*.

5.4.1.4 Noun-noun collocations identified in pre-intermediate-level writing texts

A total number of 17 candidate noun-noun combinations were extracted, of which 58.8% were found to be acceptable collocations. All of these are listed in Table 5.7 showing their context from the learners' writing, citation and LogDice scores.

Table 5.7: Sample of acceptable noun-noun collocations identified from pre-intermediate-level texts with context and LogDice scores

Noun-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Living room	I sat with my family in *living room*	≥5	7.6
Security guard	we asked a securty [*security*] gured [*guard*]	≥5	8.8
Fruit garden	the bus of our tour came and took us big garden have a small space of *fruit garden*	≥5	6.3
Room service	he asked the *room* serves [*service*] to clean our rooms	≥5	6.1
Sister brother	In 2014 my *sisters* and *brother* and i [I] we decaded [decided] to go to new york city	≥5	10.5
Gifts shops	We saw the high billdings [buildings] and the modern restaurants and a lot of candys [candies] and *gifts shops*	≥5	7.5
Breakfast area	We went to the *breakfast area*	≥5	2.9
Clothes bags	we didn't found our *clothes* and *bags*!	≥5	6.5
Police station	we went to *police* staistion [*station*] and told him the whole story	≥5	9.7

The table shows that this type of collocation is unlike previous types, in that most of these noun-noun type collocations were identified as being high level collocations: *living room, security guard, sister brother, gifts shops* and *police station*. One of these high level collocations, *brother sister,* is at the top score of the scale band used in this study with a score of 10.5 (See Section 4.5.2). All nouns making these noun-noun collocations were different and not repeated in this level of learners' data unlike collocations such as *the worst vacation* that occurred frequently in the data. Only one collocation identified was considered low level; *breakfast area* scoring 2.9. Three collocations identified were medium level: *fruit garden, clothes bags* and *room service*.

5.4.1.5 Unidentified VAN combinations produced by pre-intermediate-level learners

A total of 39 candidate VAN combinations were extracted, but unidentified in the BNC, representing 19.7% of the total production. Most of the unidentified combinations were verb-noun and adjective-noun combinations while the fewest were noun-verb and noun-noun. This means that they did not reach the threshold score, appearing less than five times, and therefore, they were not checked for their score on the Log Dice. As an example, *took shot* appeared in the BNC with reference to photography or, *shots of drink* and *gun shots* while the writer meant a medical injection. Another example is *tourist guide*, where the learner produced the combination in a subject-verb form whereas examples in the corpus were found as compound nouns as in '*tourist guiding'* or '*tourist guided walk'*. Other combinations were not identified because they did not occur at all in the corpus such as: *shingle hotel, candies shops, worst vacation* and *tired brother*. Only one combination reached the threshold score, but its LogDice score was less than zero, which is *had vacation* with -0.027. All of these less idiomatic combinations are listed in Appendix 11 with their types and citation numbers.

All unidentified combinations were referred to native speaker informants' judgement. Table 5.8 lists them with the contexts in which they appeared in learners' texts, and the judgement they received by the native speaker informants. As explained in Chapter Four, Section 4.5.3, there were three possible decisions to be made for each combination; acceptable, non-acceptable or unsure. The native-speaker informants decided that 25 combinations are acceptable collocations while 14 combinations were considered non-acceptable and none were considered unsure. Those non-acceptable combinations decided by the native speaker informants are considered in the study to be less idiomatic combinations.

Table 5.8: Candidate VAN combinations that were referred to native speaker informants from the pre-intermediate level writing texts with their context and judgement

Candidate VAN Combination	**Context**	**Native-Speaker Informants' Judgement**
Arranged bag	I woke up early and *arranged* my *bag*	Non-acceptable
Asked guard	we *asked* a security *guard*	Acceptable
forgot passport	He *forgot* the *passport*	Acceptable
quick order	we were very hungry So we *quick* our *order*	Non-acceptable
Forgot tickets	we *forgot* our *tickets* at home	Acceptable
Took shots	When we went to the hospital they told my father that was a food poisonous. After we *took* the *shots* we went to the Airport	Non-acceptable
Forget bags	we forget our *bags* in airport	Non-acceptable
Had vacation	it was the worst *vacation* I *had* ever in my life	Non-acceptable
Face crash	my *face crash* to the ball	Non-acceptable
Father surprised	my *father surprised* us with tickets	Acceptable
Father screaming	my *father* was *screaming* because of his stomach	Acceptable
Tourist guide	the *tourist guide* us	Acceptable
Waiter gave	the *waiter gave* us our food	Acceptable

Worst vacation	it was the *worst vacation*	Acceptable
Sister lost	my little *sister lost*	Acceptable
Beautiful photos	we took *beautiful photos* to remember	Acceptable
Sister sick	my *sister* started feeling *sick* then my brother	Acceptable
Bored time	When we arrived in Istanbul after a long *bored time*	Non-acceptable
Huge road	I lost my myself in a *huge road*	Non-acceptable
Dancing fountain	we saw the *dancing fountain*	Acceptable
Candies gifts	We saw a lot of *candies* and *gifts* shops	Non-acceptable
food poisonous	When we went to the hospital, they told my father that was a *food poisonous*	Non-acceptable
long planeting	We arrived to China after *long planeting*	Non-acceptable
poisoning snake	that isn’t a *poising snake*	Non-acceptable
weather fantastic	The *weather* was *fantastic*!	Acceptable
shingle hotel	We checked in *shingle hotel*	Non-acceptable
daddy meal	my *daddy meal* was very cold	Non-acceptable
Candies shops	We saw a lot of *candies* and gifts *shops*	Acceptable
Building restaurants	We saw the high buildings and the modern *restaurants*	Acceptable

The native speaker informants were given the option to provide further collocations if they wished. Some offered alternatives for the given combinations, which can be considered acceptable collocations. The suggested alternatives varied between structural and lexical. For example, the combination *daddy meal* was corrected to *daddy's meal* by adding the possessive ‘s’. Likewise, *long planeting* was corrected to *long planning* while the context does not indicate the meaning of planning.

Other combinations were corrected lexically, as required, forming acceptable collocations such as *food poisonous,* which was corrected to *food poisoning*, and *poisoning snake*, which was corrected to *poisonous snake*. As *poisonous snake* scores 8.8 on the LogDice showing a fixed high-level collocation. The list of collocations in the BNC retrieved through the Sketch Engine tool shows a further stronger collocation; *venomous snake* that scores 9.2 on the LogDice. While *poisonous food* could be an acceptable collocation by native speakers, and could only be a matter of writing the collocation in the correct order, the context in which *food poisonous* appeared in the student text, suggests the meaning of the collocation *food poisoning*, as corrected by the informants. Some verbs were corrected by some of the native speakers, such *forgot tickets* being corrected to *left tickets* even though most native speaker informants agreed on it as *forgot tickets*. Also, *quick order* was corrected to *order quickly* and *made the order quickly*, *arranged bags* was corrected as *organized bags*, and *huge road* was corrected as *long road*. Using Sketch Engine to search the BNC (described in Chapter Four, Section 4.5.2), for strong collocations of greater idiomaticity than those less idiomatic combinations produced by learners, conveying similar meanings, it suggests *pack bags* and *prepare bags*, which score 9.1 and 3.9 respectively on the LogDice, rather than *arrange bags*. Also, it suggests the fixed collocations *long road* and *wide road*, which score 6.6 and 6.1 respectively on the LogDice, with greater exclusivity and idiomaticity than *huge road*.

Some combinations were not given corrections at all, although the native-speaker informants judged them as non-acceptable, such as in the combinations *shingle hotel* in *we checked in shingle hotel* and *face crash* in *my face crash to the ball*. However, the lists of collocations in the BNC I obtained by using the collocations tool in the Sketch Engine could include some fixed collocations, which can be used to convey similar meanings to those required by the learners, but with greater idiomaticity. For *shingle hotel*, there are *beach hotel* and *seaside hotel* with scores 7.4 and 4.0 respectively on the LogDice.

5.4.2 Intermediate-level analysis

This section analyses the writing of higher-level (intermediate) Saudi learners following the same organization described previously in this chapter. Similarly, this section considers each type of collocation identified, and then discusses collocations that were not identified in the BNC, and were referred to native speaker informants, along with the less idiomatic combinations.

5.4.2.1 Verb-noun collocations identified in intermediate-level writing texts

A total of 63 candidate verb-noun combinations were extracted, of which 88.8% were found to be acceptable collocations. Table 5.9 gives some of the collocations identified for this analysis, while the full list of the 56 verb-noun collocations identified from this level texts is given in Appendix Ten. Following the procedure explained in Chapter Four, Section 4.5.2 to identify collocations in the BNC using the LogDice, most of the collocations identified were medium level collocations scoring between 3.6–7.0 such as: *forget home, see family, affect personality, choose place, visit cities* and *own car*. High level collocations scoring between 7.1-10.5, were the next most frequent, such as: *find things, have time, make life* and *live life*. The lowest number of collocations identified were low level collocations, scoring between 0.1-3.5 such as: *have entertainment, have disadvantages, find clinics* and *recommend people*.

Table 5.9: Sample of acceptable verb-noun collocations identified from intermediate-level texts with context and LogDice scores

Verb-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Enjoy weather	while you are enjoing [*enjoying*] the windy *weather*	≥5	5.9
Visit village	They had to *visit* a small *village*	≥5	6.7
Need job	you will *need* a *job* which is cheap to afford this cheap life there	≥5	6.6
Afford life	you will need a job which is cheap to *afford* this cheap *life* there	≥5	3.1
Smell smoke	you can not [can't] *smell* the *smoke* of the factories, big restaurants and cars	≥5	8.2
Finish education	I *finish* my educiton [*education*] in the city	≥5	3.8
Has amenities	village *hasn't* many entertainment and emenities [*amenities*]	≥5	0.7
Has centres	a village *hasn't* good health centers *[centres]*	≥5	5.2
Own bike	you should *own* your own car or *bike*	≥5	4.5
See advantages	Some people will *see* a lot of *advantages* if you live in a village	≥5	5.1
Get job	the people in a village can't *get* the best *job* with good and high salaries	≥5	8.2
Find parks	You are not able to *find parks* or playgrounds	≥5	3.5
Make call	You need to *make* a *call*.	≥5	6.8
Love place	you *love* quite [quiet] please [*place*] anyway	≥5	5.3

Some combinations included more than two words such as *forget home town* and *has health centres*, thus, they were extracted as *forget town*, *forget home* and *has centres*. The latter combinations were identified as acceptable collocations using the corpus, but *forget town* was not because it did not meet the threshold score conveying accpeatbaility, so it was referred to native speaker informants' judgement.

The analysis of this level of learners also showed that they produced a variety of acceptable collocations of this verb-noun type using the same nouns. For example, the noun *life* was found in collocations such as: *improve life* and *like life* as medium level collocations, *live life* and *make life* as high level collocations and *afford life* as a low level collocation. Also, *time* was produced in two high level collocations; *waste time* and *have time*, and *city* was produced in three medium level collocations as *love city*, *make city* and *visit city*. In addition, the noun *job* was produced in a medium level collocation *need job* and in a high level collocation *get job*.

5.4.2.2 Noun-verb collocations identified in intermediate-level writing texts

A total of 16 candidate noun-verb combinations were extracted, of which 15 were found to be acceptable collocations. As there were few of this type and only one was not identified, they are all listed in Table 5.10 in the contexts they appeared in from the learners' texts and their citation and score numbers.

Table 5.10: Sample of acceptable noun-verb collocations identified from intermediate-level texts with context and LogDice scores

Noun-Verb Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
People have	*people* can *have* a pure and fresh air	≥5	8.5
Thing make	any other *thing* can *make* the air polluted	≥5	7.8

Noun-Verb Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Thing bother	any other *thing* can *bother* them	≥5	4.9
People prefer	lots of *people* prefare [*prefer*] living in a big city and even see it as tortutre [torture]	≥5	5.8
People hate	*people hate* living in villages	≥5	5.0
Village have	some *villages* a lot *have*	≥5	5.0
City contains	this things make the *city* contun [*contains*] lots of pollitoned [pollution]	≥5	4.8
People think	Some *people think* that	≥5	8.5
People see	Some *people* will *see* a lot of advantages if you live in a village	≥5	8.1
People get	the *people* in a village can't *get* the best job with good and high salaries	≥5	8.5
Villages got	*villages got* alot [a lot] of trees	≥5	3.9

As the above table shows, learners at this level produced only medium and high level collocations, almost equally, of this noun-verb type. In addition, most of the collocations used the noun *people*, and the other popular nouns were *village* and *thing*. Some collocations were repeated in the data such as *village have/has* and *people have*. The only combination that was extracted, but found unidentified by referring to the BNC is *village force*, which will be discussed in a later part of this section.

5.4.2.3 Adjective-noun collocations identified in intermediate-level writing texts

A total of 92 candidate adjective-noun combinations were extracted, of which 84.7% were found to be acceptable collocations. Most of the collocations identified were medium level collocations such as: *important facilities, new name, actual place, great thing* and *big markets*. High level collocations were the next most frequent, such as: *long way, good* *idea, good people, big city* and *different countries*. Low level collocations were the least frequent type at this level such as: *cheap life, kind people, crowded place* and *good salaries*. The full list of the 78 adjective-noun collocations identified from this level texts are in Appendix Ten, Table 5.11 lists some examples.

Table 5.11: Sample of acceptable adjective-noun collocations identified from intermediate-level texts with context and LogDice scores

Adjective-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Small villages	So many people still live in *small villages*	≥5	7.7
Good doctors	there is no important fucllitys [facilities] such as hospitals with *good doctors*	≥5	5.4
Quiet life	people can have a very calme [calm] and quite [*quiet*] live [*life*] with city nois [noise]	≥5	6.2
Long hours	you can forget about *long* houres [*hours*] driving	≥5	7.9
Powerful thing	you can find home is the most *powerful thing*	≥5	2.6
Old thing	Some might consedr [consider] it an *old thing* to it	≥5	6.6
Green areas	You will be able to clear yourself and cure your soul with the *green areas* all around you	≥5	4.8
Same level	the advantages and disadvantages to living in a village might be on the *same level*	≥5	7.2
Cheap job	you will need a *job* which is *cheap*	≥5	4.4
Fresh food	the *food* there *fresh*	≥5	7.5
Healthy life	Living in village it’s a good idea for healty [*healthy*] *life*	≥5	5.4

Some adjectives were common in these learners' texts, such as *good*, *small*, *big* and *great*. Among the collocations identified which included these adjectives are: *good idea, good people, good place, good thing, good doctors, good side, good mood,* and *good centres*, of which the first four are high level collocations and the latter four medium level collocations. Further examples with *small* and *big* are: *small villages, big cities, big hospitals, big markets, big schools* and *big restaurant* that varied among the three levels of fixedness. There were also some nouns that were produced with different adjectives and still made acceptable collocations such as: *place* in *great place, crowded place, quiet place, peaceful place, natural place* and *actual place*.

5.4.2.4 Noun-noun collocations identified in intermediate-level writing texts

A total of 74 candidate noun-noun combinations were extracted, of which 78.3% were found to be acceptable collocations. Table 5.12 includes some examples of this type of collocation showing their context from the learners' writing, citation and LogDice scores. The full list, of 58 acceptable collocations, is given in Appendix Ten.

Table 5.12: Sample of acceptable noun-noun collocations identified from intermediate-level texts with context and LogDice scores

Noun-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Traditions customs	they still stick with their tardtions [*traditions*] and *customs*	≥5	7.3
City noise	people can have a very calme [calm] and quite [quiet] live [life] with *city* nois [*noise*]	≥5	3.6
Advantages disadvantages	living in village have *advantages* and *disadvantages*	≥5	9.4

Noun-Noun Collocations	**Citation from the Text**	**Citation in the BNC**	**LogDice Score**
Road traffic	there are a few stores with a few *roads* and no ~~a~~ trafic [*traffic*]	≥5	9.1
Birds animals	there is no nois [noise] at all just the *birds* and *animals*	≥5	8.5
Human being	The home town location of the *human being* effects on his body, brian [brain] and heart	≥5	6.5
Shops cinemas	when there are not many intertaining [entertaining] places to go like coffee *shops*, malls, cenimas [*cinemas*] and barks [parks]	≥5	5.3
Car station	there is many *cars* gas staican [*station*]	≥5	5.9
Health centres	a village hasn't good *health* centers *[centres]*	≥5	7.3
Villages cities	I would like to say that both *villages* and also *cities* are good places to live in	≥5	5.9

Only two collocations from this type were identified as low level collocations; *rush life* as in "*you like the rush and noisy life",* where *rush* can be understood as modifying *life*, and *schools institutes* as in "*people can't have many options for education like, universities, global schools and educational institutes"*, where *schools* and *institutes* are in a list. Medium level collocations were the most frequent, followed by high level collocations as given in the above table.

There were some combinations that included more than two words, as a result, they were extracted as two separate combinations. For example, the combination *home town location*, was extracted as *home town*, *home location* and *town location* all of which were identified in the corpus. *Home town* scored 7.4 as a high level collocation while *home location* and *town location* scored 4.4 and 5.3 respectively as medium level collocations. Another example is *car gas station* that was extracted as *car station* and *gas station*, which were identified as a medium level collocation for *car station* scoring 5.9, and a high level collocation for *gas station* scoring 7.1.

5.4.2.5 Unidentified VAN combinations produced by intermediate-level learners

A total of 38 candidate VAN combinations were extracted, but unidentified in the BNC which represents 15.5% of the total production. The most frequent types of those unidentified combinations were noun-noun and adjective-noun combinations with 16 and 14 combinations respectively, while verb-noun and noun-verb types were the least with seven and one combinations respectively. Those combinations which were not identified in the BNC due to not conveying conventionality by not reaching the threshold score of five appearances, were not checked on the LogDice. As an example, *love farms* was cited three times in the BNC, but none of them was similar to learner's "*I love the green farms*". Two of the BNC citations had the verb love and farm in a separate sentence and clause, and the third as "*farm workers love the job*". Another example is *good clinic*, which was found in the corpus 12 times, but appeared only four times as *good* modifying the noun *clinic*. Other combinations were not identified because they did not occur at all in the corpus such as: *malls schools, markets bookstores, mall cinemas, shining stars, well-equipped hospitals* and *well-developed transportation*. All of these combinations are listed in Appendix 11 with their types and citation. The following section discusses these in relation to native speaker informants' judgement.

As all unidentified combinations were referred to native speaker informants' judgement, Table 5.13 shows them with the contexts in which they appeared in the learners' texts, and the judgement they received from native speaker informants. As discussed (Section 4.5.3), the informants had to decide if a candidate collocation was acceptable, non-acceptable or unsure. The native-speaker informants decided that 22 combinations were acceptable while 15 were non-acceptable, and remained undecided on one.

Table 5.13: Candidate VAN combinations that were referred to native speaker informants from the intermediate level writing texts with their context and judgement

Candidate VAN Combination	**Context**	**Native-Speaker Informants' Judgement**
Forget town	they should not *forget* their home *town*	Non-acceptable
Find playgrounds	You are not able to *find* parks or *playgrounds*	Non-acceptable
Cure soul	You will be able to clear yourself and *cure* your *soul*	Undecided
Like rush	wither you *like* the *rush* and noisy life	Non-acceptable
Have malls	they will *have* a lot~~s~~ of *malls*	Acceptable
Contains pollution	this things make the city *contains* lots of *pollution*	Non-acceptable
Love farms	I *love* the green *farms*	Acceptable
Villages force	Some *villages* might *force* you to drive a long way	Non-acceptable
Small clinics	You can only find *small clinics*	Acceptable
People scared	*people* oftenly *scared* of what things they have never tried	Acceptable
Noisy voices	you will get rid of all the *noisy voices* of the crowd city and the cars	Non-acceptable
Noisy life	wither you like the rush and *noisy life*	Acceptable
Clear life	wither you like the rush and noisy *life* or the peaceful and *clear* one.	Non-acceptable
School complete	The hospitals and *school* its not *complete* like the city	Non-acceptable

Candidate VAN Combination	**Context**	**Native-Speaker Informants' Judgement**
Hospitals complete	The hospitals and *school* its not *complete* like the city	Non-acceptable
Big mall	because of not having *big* markets, *mall*, bookstores	Acceptable
Big bookstores	because of not having *big* markets, mall, *bookstores*	Acceptable
Green farms	I love the *green farms*	Acceptable
Global schools	the people can't have many options for education like, universities, *global schools* and educational institutes	Acceptable
Well-equipped hospital	the *hospitals* aren't *well-equipped*	Acceptable
Good clinic	there aren't *good clinics*	Acceptable
Well-developed transportation	the public *transportation* isn't *well-developed*	Non-acceptable
Parks playgrounds	You are not able to find *parks* or *playgrounds*	Acceptable
Phone station	to drive a long way just to get to the *phone station*	Acceptable
Shining stars	watching the sky full with the *shining stars*	Acceptable
Forest farms	the fresh air Because of *forest* and *farms*	Non-acceptable
Factories restaurants	you can't smell the smoke of the *factories*, big *restaurants* and cars	Non-acceptable
Shops malls	places to go like coffee *shops*, *malls*, cinemas and parks	Acceptable
Mall cinemas	places to go like coffee shops, *malls*, *cinemas* and parks	Acceptable
Mall parks	places to go like coffee shops, *malls*, cinemas and *parks*	Acceptable

Candidate VAN Combination	**Context**	**Native-Speaker Informants' Judgement**
Cinemas parks	places to go like coffee shops, malls, *cinemas* and *parks*	Acceptable
Markets mall	because of not having big *markets*, *malls,* bookstores	Acceptable
Markets bookstores	because of not having big *markets*, malls, book stores [*bookstores*],	Acceptable
Mall bookstores	because of not having big markets, *malls, bookstores*	Acceptable
Entertaining places	there are not many *entertaining places* to go to	Acceptable
Malls hospitals	they will have ~~a~~ lots of *malls*, *hospitals* and big school, university	Non-acceptable
Malls schools	they will have ~~a~~ lots of *malls*, hospitals and big *school*, university	Non-acceptable
Crowd city	You will get rid of all the noisy voices of the *crowd city*	Non-acceptable

Some of the native speaker informants offered alternative collocations. For example, adjectives were corrected in combinations such as *noisy voices*, to *loud voice, clear life* was corrected to *calm life,* and *complete hospitals* was corrected to *equipped hospitals*, and the verb in *cure soul* was corrected to *heal soul*. A list of suggestions made by native speaker informants are in Appendix 13. The list of collocations in the BNC, I obtained using the collocations tool in Sketch Engine has some fixed collocations with greater idiomaticity, which could be conveying similar meanings to those required by the students. For example, *loud voice* and *high voice* rather than *noisy voices,* which score 8.5 and 6.3 respectively on the LogDice, and *comfortable life* and *peaceful life* rather than *clear life,* which score 4.9 and 4.5 on the LogDice. Also, the collocations *nourish soul,* *purify soul* and *free soul* are found with scores 5.4, 5.4 and 4.9 respectively on the LogDice, which could be used rather than *cure soul*. Conversely, *heal soul*, which was suggested by the informants appears only once in the BNC.

Some collocations were not given corrections at all, although the native-speaker informants judged them as non-acceptable, such as *like rush, contains pollution, crowd city* and *well-developed transportation*. The list of collocations in the BNC retrieved using the collocations tool in the Sketch Engine includes some alternatives for those less idiomatic combinations such as *experience rush* with a 4.3 LogDice score, and *suffers pollution, produces pollution* and *includes pollution* with 6.1, 4.7 and 4.6 scores respectively on the LogDice. *Busy city* and *crowded city* are found for *crowd city,* and *school ready* for *school complete* in the BNC using the collocations tool in Sketch Engine with these LogDice scores, 5.7, 5.1 and 4.6 respectively. Although *efficient transportation* is found as a fixed collocation with 4.4 score on the LogDice, it only has three citations in the BNC.

5.5 Summary

This chapter has described how collocations produced were extracted, identified and analysed in the two sets of Saudi students' writing. The decision to manually extract candidate combinations in this study, and the relatively low-level of the learners' writing, meant that identifying collocations was complex. Therefore, it was important to illustrate the analysis of the written texts through case studies and also to clarify the process of identifying collocations through using the BNC, and LogDice association measure, and then by referring to native speaker informants.

Pre-intermediate level learners wrote a total of 2124 words in eight texts with between 200-300 words per text. 197 candidate VAN combinations were extracted, of which 157 met the frequency threshold score in the BNC and scored above zero on the LogDice, and so were classified as acceptable collocations with a level of fixedness. The descriptive statistical analysis shows that pre-intermediate-level tend to produce more verb-noun and adjective-noun combinations at 42.6% and 29.5% respectively of the total number while the other two types, noun-verb and noun-noun represented 19.2% and 8.6% respectively.

Intermediate-level learners wrote a total of 2034 words in eight texts, with about 200-300 words per text, where 243 candidate VAN combinations were extracted. After checking their occurrences in the BNC and scores in LogDice, 205 were found to be acceptable collocations with a level of fixedness. Intermediate-level produced more combinations of adjective-noun and noun-noun types with 37.4% and 30.4% respectively of the total production of VAN combinations. The other two, verb-noun and noun-verb types represented 25.5% and 6.5% respectively.

The analysis of the two levels of Saudi learners' written texts, i.e. pre-intermediate and intermediate, shows that both levels are able to produce acceptable collocations with a reasonable degree of success. The percentage of identified combinations which were identified as acceptable collocations from both levels' writing is 82.3%, of which the majority were medium level collocations. It is notable that of the rest of the acceptable collocations produced, more were high level collocations than low level ones. Higher level learners produced 84.3% of their total collocation production as fixed, while the lower level was somewhat less successful with 79.6%. While the higher level produced more combinations of adjective-noun and noun-noun combinations, they succeeded in producing most of their verb-noun and noun-verb types as acceptable collocations with 88.7% and 93.7% respectively. In contrast, the lower level produced more verb-noun and adjective-noun combinations, and had most of their verb-noun and noun-verb types as acceptable collocations with 85.7% and 84.2% respectively.

The analysis of the native-speaker informants’ judgement identified the collocations that can still be considered acceptable even though they failed to be identified using the BNC corpus. Less idiomatic combinations, which are VAN candidate combinations extracted from the learners’ texts but were not identified through either method i.e. corpus and native-speaker informants, represented only 7% of the total combinations produced. This means that although Saudi learners show an ability to produce fixed and acceptable collocations, the collocational production was sometimes erroneous for them. Of the VAN candidate combinations produced by the pre-intermediate level 7% were considered to be less idiomatic and non-acceptable, where verb-noun, adjective-noun and noun-noun types were the most frequent while noun-verb type was the least i.e. one combination. The same pattern was repeated at the intermediate level as of the 5% of candidate combinations deemed to be less idiomatic, the three types verb-noun, adjective-noun and noun-noun were found to be the most frequent types while the noun-verb type was the least i.e. one combination. The difference in number between those less idiomatic combinations was not major between the two levels of learners. Thus, learners’ proficiency level was not evident in their production of the types of collocations investigated.

Finally, it should be acknowledged that some of the acceptable collocations that are identified by native speaker informants could be found in the list of collocations in the BNC using the collocations tool in Sketch Engine with notable scores on the LogDice. However, they were not identified in the early stages of the process when referring to the BNC citations and then the LogDice score in the Sketch Engine tool for two reasons. As explained in Chapter Four, Section 4.5.2, firstly, the collocation should co-occur five times or more in the BNC to show conventionality to what the learner produced, then secondly, a LogDice score is obtained for the collocation. Those acceptable collocations which may have a significant score on the LogDice, could either co-occur in the BNC less than five times i.e. the frequency threshold, such as *fantastic weather*, or co-occur within a different span boundary than the one used in the study i.e. five-word window limited with a clause and syntactic relation, such as *global schools*, or a different meaning than that of the learner's such as *took shot*.

The following chapter will discuss how the analysis detailed here is used to address the research questions and relate that to the relevant literature and the possible impact on the language learning process.

Chapter 6 Findings and Discussion

6.1 Introduction

This chapter starts by outlining the main findings of the research in response to the three research questions which provoked this study. The first research question; "Do Saudi-foundation year students at university produce acceptable collocations in writing? If so, what are the types?" is addressed by the findings of the overall analysis of writing from both levels. The second research question: "Which less idiomatic combinations do Saudi foundation-year students produce in writing?**"** is addressed by extracting from learners' written texts less idiomatic combinations which were not identified as either fixed or acceptable collocations. The third research question: "What are the similarities and differences between acceptable collocations produced by two levels of Saudi foundation-year university students, studied in their written texts?" is addressed by looking at the similarities and differences in the production of the acceptable collocations extracted from the written texts of each level.

The research findings are then discussed in relation to three main themes in the Literature: firstly, the Saudi foundation-year students' ability to produce acceptable collocations and whether or not this result is related to their language learning context, secondly, the learners' production of less idiomatic combinations, and possible problems for their production, thirdly, the relationship between Saudi foundation-year students' language proficiency levels and their written production of acceptable collocations.

6.2 Findings

Research Question One: **Do Saudi foundation-year students at university produce acceptable collocations in writing? If so, what are the types?**

The analysis of the written texts produced by pre-intermediate and intermediate level students indicates that learners from both levels produce acceptable collocations in their writing. Most of the extracted candidate verb, noun, and adjective (VAN) combinations reached the threshold score in the BNC meaning conventionality as well as being identified as acceptable collocations in three levels of fixedness according to the LogDice measurement in Sketch Engine. Thus, most of the extracted combinations are found to be acceptable collocations of the types investigated: verb-noun, noun-verb, adjective-noun and noun-noun.

A total of 442 VAN combinations were extracted from the learners' written texts, of which 364 were found to be acceptable collocations. This shows that Saudi foundation-year university students are able to produce a significant percentage, about 82%, of acceptable collocations in their writing, even though they are not explicitly taught collocations. The majority of acceptable collocations produced were found to have LogDice scores between 3.6 and 7.0, that is, were medium level collocations; the next most frequently produced were high level collocations, those with scores between 7.1 and 10.5. Least frequent were low level collocations that scored less than 3.5. Some of the high level collocations produced by these learners were top level collocations such as: *learned lesson* with 9.6, and *advantages disadvantages* with 9.4.

This finding shows that these Saudi foundation-level university students produce such acceptable collocations in a variety of different forms. It also shows their ability to use vocabulary they have learned in classrooms, or have acquired outside the classroom, in combinations that can be identified as acceptable collocations. However, many of them should not be problematic as they carry a transparent meaning. It is still good to find the four different types of collocations: verb-noun, noun-verb, adjective-noun and noun-noun, across both levels, seven of the eight groups follow the same pattern in that the most frequent collocations found were medium level, followed by high level and low level collocations being the least frequent. Only one group, noun-noun collocations, produced by the lower level group, did not follow this pattern, with high level collocations being the most frequent, followed by medium level and, as with the other groups, the lower level collocations were the least frequent. This result suggests that learners may produce some types of collocations differently than other types. This is addressed in the second research question.

Research Question Two: **Which less idiomatic combinations do Saudi foundation-year students produce in writing?**

The less idiomatic combinations are those that were not found in native speaker language use through corpus or phraseological analyses. As previously discussed, extracted VAN combinations that were not found to be acceptable collocations were referred to the judgement of native speaker informants. Only a few of these combinations were not accepted by native speaker informants as collocations. They are considered in the study as less idiomatic combinations. They may have been produced by learners for certain reasons. However, this study focuses on their types in relation to learners' proficiency levels.

Pre-intermediate level learners produced 197 candidate VAN combinations, of which 39 do not occur in the BNC. In the second stage of identification, which was through the phraseological approach by referring to native speaker informants, 25 were identified as acceptable collocations and the remaining 14 combinations were either non-acceptable or undecided. Intermediate-level learners produced more candidate VAN combinations than the lower-level learners: 243 combinations, with a smaller number of them being unidentified in the BNC; 38 combinations. Native speaker informants identified 22 as acceptable collocations and the remaining were found to be less idiomatic combinations; 15 as non-acceptable and one as undecided.

In total, 77 combinations across both levels were given to native speaker informants to judge. Forty-three of them were judged to be acceptable and the rest, 34, were not. Of these non-acceptable combinations, the noun-verb type was the least frequent while the other three types, verb-noun, adjective-noun and noun-noun appeared almost equally. The learner's proficiency level does not seem to be related to these results; 16 less idiomatic combinations were produced by higher level learners while 18 less idiomatic combinations were produced by the lower level. While a low level of language proficiency can result in problematic written production of collocations, a higher level does not seem to guarantee a perfect production of collocations. However, the issue of producing non-acceptable collocations in learners' written texts does not suggest a challenging area for both levels as found from this study findings. Based on the corrections and suggestions for acceptable collocations provided by native speaker informants, this finding will be discussed below.

Research Question Three: What are the similarities and differences between the acceptable collocations produced by two levels of Saudi foundation-year university students, studied in their written texts?

The study findings show that similarities as well as differences occurred in the two levels' written production of collocations. Despite the difference in the topic written about, both levels produced a significant number of acceptable collocations of all types investigated. There were also similarities between the levels in using certain nouns and adjectives more frequently to produce collocations. Both levels produced more medium and high level collocations, and less low level ones. According to this study, students with a higher level of language proficiency are not necessarily able to produce more or better collocations than lower level students. Across the levels of fixedness of acceptable collocations, similarities between the levels of learners' production were more obvious than the differences.

However, there are some slight differences between the two levels' written production of collocations, such as in the number of acceptable collocations compared to the less idiomatic combinations. Of the collocations produced by higher levels about 85% of them were fixed, while for the lower level the figure is nearly at 79%. Similarly, only 5% of the higher level's collocation production was deemed to be less idiomatic while for the lower level it was slightly more at nearly 7%. However, these percentages are only drawn upon the descriptive statistics used in the study, and to show whether or not such differences are significant between the two levels' production of collocations, in depth statistical tests should be carried on.

Differences also occurred in relation to the production of the different types of collocations. For example, the higher-level learners showed a tendency to produce a more complex form of adjective-noun collocations than the lower level. They produced two-word hyphenated adjectives in adjective-noun collocations e.g. well-equipped hospitals and well-developed transportation, which did not occur in the lower level written texts. They also tended to combine nouns successfully to produce more noun-noun collocations than the lower level learners. However, the lower level learners tended to produce the majority of the few noun-noun collocations they produced, as high-level collocations.

Finally, both levels produced a different number of combinations among the different types, and both of them tended to succeed in producing the majority of those two types verb-noun and noun-verb combinations as acceptable collocations but were less successful with adjective-noun and noun-noun acceptable collocations. These results show that although both levels differed in their production of different types of combinations, they were similar in producing acceptable collocations with the forms, verb-noun and noun-verb. Thus, these students with different levels of language proficiency produced collocations similarly, albeit only with a slight different.

6.3 Discussion

6.3.1 Saudi learners' production of different types of acceptable collocations

My study has suggested that, contrary to the author's expectations, Saudi foundation-year university students are able to produce acceptable collocations in three levels of fixedness in their written texts despite the fact they are not explicitly taught collocations in their English lessons. This result is consistent across both levels, which suggests that these students are acquiring the ability to produce written collocations independently by being able to combine what they have learnt and acquired, both inside and outside the classroom, into collocations. These collocations were made up of different types: verb-noun, noun-verb, adjective-noun and noun.

Huang's (2001) study, discussed earlier in Chapter Two, claims that learners, specifically low-level learners, face difficulty in producing restricted collocations i.e. fixed collocations and idioms, compared to free combinations. However, he suggests that the restricted collocations are particularly more challenging when compared to idioms because such low-level learners do not usually tend to use idioms anyway. In my study, Saudi learners who are relatively low-level learners are able to successfully produce acceptable collocations with a majority of the combinations produced identified in the BNC and found to be fixed using the LogDice association measurement. Differences between the results of this study and those of Huang could be due to the nature of the investigation. While Huang conducted a controlled task-based study investigating quantitatively the difference in learners' use of free combinations, collocations and idioms, this study investigates the use of collocations in a less controlled environment by analysing learners' free writing and a reference to an association measurement test.

Looking at the types of collocations produced by Saudi learners, verb-noun and noun-verb collocations were found to be acceptable more often than adjective-noun and noun-noun collocations. This was consistent across both levels of learners' writing. Martyńska (2004), Shehata (2008) and Kuo (2009) compared the production of verb-noun and adjective-noun collocations by their learners, and found different results. In contrast, in task-based studies, Martyńska found that his Polish learners were equally good at both those types, Shehata found that her Arab learners were better in producing verb-noun than adjective-noun collocations, which coincides with the results of this study.

While this study is like Kuo's in investigating learners' production of collocations in their written texts and finding that both learners produced a high number of collocations, this production was found different in the types of collocations. Kuo's Taiwanese learners tended to be better at producing adjective-noun type than verb-noun types, whereas the Saudi learners were opposite. However, Kuo's definition of collocations was based on collocational acceptability using frequency only whereas this study identified collocations acceptability and fixedness using frequency and exclusivity in the corpus and phraseology by native speaker informants, meaning a higher criterion to examine EFL learners' production of collocations. Saudi learners showed their ability to produce acceptable collocations with identified levels of fixedness rather than only frequent collocations, which may be more challenging to learners. Additionally, Kuo's learners were university students majoring in English, with a clearly higher proficiency level than the Saudi learners on a foundation-year university course who participated in this study, the results from both studies are similar in that all students produced a significant number of collocations.

Nesselhauf (2004) principally investigated restricted, i.e. fixed, verb-noun collocations in German learners' written texts, but without making a statistical reference to association measurements. She found that this type of collocation, i.e. verb-noun, is problematic for her learners and they tend to mix up verbs, like *pay* and *take*, in a collocation such as *take care*, and write *pay care*. However, Saudi learners were able to produce the verb *take* in an appropriate collocation with no problems such as *take care*. *Take* is a de-lexical verb and could be investigated according to its particular kind of collocation in relation to learners' L1 as many languages include this type of verb and learners' production of L2 collocations could be influenced by their L1. As this study does not examine the effect of L1 in depth in Saudi learners' production of collocations, a possible explanation for this difference in findings could be a result of learners' different L1s. Such a finding in this study is further significant when compared to Nesselhauf's because the collocation *take care* was not only identified in the BNC, but also it was found to have a high score. Applying statistical measurements to test collocational association will always add value and consistency to an investigation (Gablasova et al., 2017b), especially if the association measurements are not based on intuitions such as the phraseological approach, which is used by Nesselhauf, and also this study, to judge and identify collocations produced by the learners. Thus, association measurements are not subjective when interpreting results and reliable when comparing between studies.

Saudi learners showed a higher level of competence when using particular adjectives in collocations when compared to Chinese learners' production in their written texts. Fan (2009), who investigated the production of adjective-noun collocations, found them challenging for Chinese learners, including collocations with very common adjectives such as *big* and *deep*. However, in this study Saudi learners produced collocations such as *deep sleep*, which is a fixed collocation with a high score on LogDice. Also, these Saudi students produced many acceptable collocations using the adjective *big* such as *big garden, big day, big problem, big companies* and *big brother*, with two of them being high level collocations and the rest medium. Such a difference in the results of the two studies could be because of the way each study identified and investigated collocations. Fan's learners were given the task of a picture to describe, which may limit the variety of collocations they could produce, and were compared to the use of native speakers who also have done the same task.

While differences between this study's findings and other similar studies could be related to the nature of the task and the analytical methods followed, studies can produce different results because of how collocations are identified. Among the studies that used the frequency and exclusivity of a collocation, as in this study, to investigate the production of collocations by learners, is one by Siyanova and Schmitt (2008). However, unlike this study they referred to the MI score, which tests rare exclusivity, and not the LogDice index. They analysed the production of adjective-noun collocations by EFL and ESL Russian learners by referring to the collocation's frequency in the BNC and exclusivity using the MI score. They found that those Russian learners were able to produce accurate collocations, but with limited success as only 45% of the total number of collocations produced reach this MI score threshold of fixedness. It would have been interesting to know if the same result would be obtained if the LogDice had been used, as in this study, testing a collocation's exclusivity rather than rare exclusivity. Under these circumstances it is possible that the results for Russian students may have been similar to those of this study. This would have been especially interesting given the significant differences between Arabic and Russian.

Durrant and Schmitt (2009), who also focused on learners' use of adjective-noun collocations, refer to a collocation's frequency by using the *t-score* test and exclusivity by using the MI score test to draw on the gap between those two measurements according to learners' use of collocations. Their research found that Turkish and Bulgarian learners used high-frequency collocations more than those of a low frequency, which indicates that, as could be expected, learners tend to rely more on frequent collocations, such as free combinations, which they may have acquired both inside and outside the classroom, while underusing exclusive combinations such as the restricted collocations, which they would have had less exposure to. Their findings also highlighted the grey area between free combinations and restricted collocations, as indicated previously by Huang (2001), but through using statistical tests. This study investigated this gap research into learners' production as well as addressing it by applying the LogDice measurement, which combines the testing of frequent and exclusive collocations. This application is valuable in that it analyses learners' written production of acceptable collocations without separating relevant results by relying only on frequency or rare exclusivity. In this study, the adjective-noun collocations produced most frequently by Saudi learners were medium level ones.

By showing a practical example of applying the LogDice measurement to test the production of collocations by EFL learners in their written texts, and its suitability to relatively low-level learners such as in this context, this study contributes to the knowledge of collocational production. Fernández and Schmitt (2015), who studied frequent and infrequent lexical 2-gram collocations, suggest that a test like LogDice is an appropriate way to analyse learners' collocational knowledge. However, their study tested Spanish learners' production of collocation only by using the raw frequency, *t-score* and

MI score, which led them to the conclusion that a measurement such as the LogDice could be more efficient in examining learners' production, than the methods they used. This is due to the features the LogDice has through combining frequency, infrequency and fixedness of collocations, which are communicative collocations that are produced by learners. Fernández and Schmitt's learners tended to produce the majority of their collocations with a correlation to raw frequency and t-score (high frequency collocations) rather than less correlation with the MI score (infrequent and rare collocations).

By the use of the LogDice, which has never been done before in this type of research, this study is contributing to the knowledge of collocation production. The use of the LogDice has been suggested by researchers as a suitable method of examining learners' production of collocations, which lie between high frequency and rare exclusivity. The application of the LogDice in this study and results which show Saudi learners' ability to produce a high number of acceptable collocations with a level of fixedness is in line with studies which suggest the feasibility of the LogDice as a measurement such as those by Durrant and Schmitt (2009) and Fernández and Schmitt (2015). The LogDice score can indicate the learners' real knowledge of collocation by showing their production of frequent as well as exclusive collocations. Therefore, I conclude that the use of the LogDice gives a reliable and accurate result to the real production of collocations by Saudi learners in their written texts.

6.3.1.1 Saudi learners' production of collocation in relation to their language learning context

It was found that Saudi university students produced a high number of acceptable collocations of different types, even though they are not being given explicit instructions about them in their foundation year. This finding supports the hypothesis that EFL learners are capable of assimilating collocations and learning to produce them independently. Wray (2000:264) says that learning collocations is not a single phenomenon, as it takes input from a range of sources to acquire collocational knowledge. This assertion could be applied to Saudi students, who seem to be picking up collocations naturally and tending to use them successfully through other channels than their classroom experiences alone.

Rather than limiting the examination of learners' production of collocations in a list of given or suggested collocations by the researcher, such as Huang (2001), analysing learners' written texts gives a wider scope to the investigation. For example, this study shows that learners were able to give a variety of adjectives for some nouns when writing about given topics. They were able to produce the noun *hotel* in acceptable collocations such as *beautiful hotel*, *famous hotel* and *good hotel*, all of which are medium level collocations. Those collocations were produced by pre-intermediate learners whose textbook includes a section in a unit called "In a hotel", but only provides some relevant vocabulary such as: *reception, dining room, gym* and *swimming pool*. Many of the collocations produced used vocabulary other than that which the students were given in the text book.

Students in this study tended to combine adjectives with nouns in acceptable collocations according to their individual language knowledge and personal experience, especially given that the topic of the writing was to describe a holiday, which is clearly a very personal experience. Nonetheless, the influence of the taught materials cannot be totally ignored. Pre-intermediate level learners produced some successful verb-noun collocations such *took photos* and *took pictures.* While only *took photos* was mentioned in the textbook, a learner produced *took pictures* as a collocation by using the same verb *take* but with another noun than the one she was taught. This also supports the idea of the natural acquisition and production of collocations by these learners. This is especially noticeable for frequent collocations such as those with the verb *take*.

Other examples from the intermediate level learners' written texts are those collocations produced with noun *city*. In the vocabulary section of one of their textbook units, they were given some adjectives that are common with this noun. However, none of those given adjectives were used by the learners in their writing, instead they created their own combinations. The textbook gives *ancient, young, modern, capital, historic* and *busy* as adjectives combining with *city* while learners produced combinations such as *big city, city noise* and *crowd city,* of which the first two were identified as acceptable collocations whereas only the last one was not. The fact that the subject of the writing composition was slightly different to the topic in the course book could be an explanation for the different choice of adjectives by the students. However, the learners still had the ability to produce acceptable collocations with their choice of word combinations. Thus, their production is not solely dependent on what they are given in classroom.

Shehata (2008), who tested Arab learners in a task-based investigation and found similar results to this study's findings, attributed the results to the fact that language teachers usually focus on verbs more than adjectives. While this study does not investigate teachers' influence directly, there was a very obvious example of this suggesting a relationship between teachers and their learners' production of collocations. The topic this lower level wrote about was "The Worst Vacation". Even though the class textbook used *holiday* not *vacation*, which is expected as the textbooks use British English, the teacher used the word *vacation*. Subsequently, all relevant collocations produced by the learners used the noun *vacation*, not *holiday,* such as *spent vacation, had vacation* and *long vacation*. This would suggest that the teacher has more influence than the textbook on learners' language production.

Hunston (2002:193) writes that language learners tend to use the language in their own way according to what they have learnt, and they should not be forced to use “lexical chunks” exclusively. This supports the idea of learners’ reliance on what is familiar to them, meaning high frequency collocations, rather than what is rare and mostly exclusive to native speakers, the likelihood that are generally low frequency collocations. Most textbooks give examples of what the authors believe to be typical native speaker language, but this may not include high frequency collocations (ibid.:43-44). Indeed, while some textbooks use corpora for their lexical and grammatical choices (Burton, 2012:104), they do not seem to use collocational research to support the choice of collocations included. Burton claims that awareness of collocations should spread among language teachers as well as textbook publishers in order to encourage a proper use of corpora in learning and teaching materials. As some text book examples may sound challenging to learners, Hunston suggests that students seek the security of producing language which is already known to them such as the examples discussed above.

It is one of this study’s limitations that it does not investigate other influences learners may experience outside classrooms. A follow-up task such as an interview with the learners would have been useful in discovering some of the possible external influences on learners’ which has affected their production of collocations. The study findings are indicative for future research of a similar context taking these limitations into consideration.

6.3.2 Saudi learners’ production of less idiomatic combinations

There were very few candidate VAN combinations, which were extracted from the Saudi learners’ writing, but were not identified in the corpus nor by native speaker informants. Still, they are considered interesting to this study as they could indicate which combinations could be more challenging for the students to produce accurately. The Saudi learners in this study at both levels produced more less idiomatic collocations of some types than of others. Indeed it was noticeable that of the four types of collocations studied, the noun-verb type was produced most successfully while the other three types had an almost equal failure rate. Thus, it could be said that language proficiency does not automatically lead to a better production of certain types of acceptable collocations.

Whether or not using corpus approaches to analyse learners' written production of collocations, studies have shown different results in relation to the different types of collocations produced by learners. Shehata (2008), who investigated Arab learners' use of collocations in a task-based study, found similar results to this corpus-based study in that learners have been more successful in producing verb-noun combinations than adjective-noun ones.

This is in contrast to Kuo's (2009) Taiwanese learners who produced more accurate adjective-noun than verb-noun collocations. Furthermore, Martyńska (2004) and Chiu and Hsu (2008) conducted task-based studies and reached different results to each other, and to this study, when investigating the noun-verb collocations. While Martyńska found that Polish learners faced difficulty in producing this type compared to the verb-noun type, Chiu and Hsu found their Taiwanese learners did equally well with verb-noun and noun-verb types. This study contrasts with both of them by showing that Saudi learners are most successful with noun-verb types with the other types, which show an almost equal failure rate. It could be speculated that the reason for this lies with the learners' L1 and this would benefit from more, comparative, studies.

To support this aforementioned speculation, it is worth looking at Parkinson (2015) who examined the production of noun-noun collocations. She studied leaners with three different L1s: Tswana ESL learners, Spanish EFL and Mandarin EFL learners. Her results showed that Mandarin EFL learners were most successful in producing noun-noun collocations, which could be linked to the fact that Mandarin is a language which allows the noun-noun phrase system. However, Spanish does not allow the noun-noun phrase system and it was the Spanish students who were the least successful in producing these types of collocations. This structure does exist in Arabic and could support Parkinson's findings, especially as it seems that Saudi higher level students with a greater knowledge of the English language are better able to produce this type of collocations than lower level learners. However, this needs a further investigation in the future.

It has been agreed by researchers such as Siyanova and Schmitt, (2009) and Fernández and Schmitt, (2015) that learners tend to use collocations which are frequent as well as exclusive. It was unsurprising that most of the Saudi learners' collocational production in this study was identified as fixed while only a few were found to be less idiomatic combinations. Also, the topics learners wrote about are likely to have an impact on the results of the study as Gablasova et al. (2017:175) state that: "the effect of topic must be taken into account". They wrote about general topics, which could be one of the reasons for the production of a high number of acceptable collocations. Saudi learners may have relied on what is known and familiar to them thus suggesting that their production of collocation is not as problematic as that of other learners' results from different studies. Such differences as in the methodology used as well as the topics used may contribute to different findings.

6.3.2.1 Possible reasons for learners' production of less idiomatic combinations

The results of this study have shown a very high level of success in the production of acceptable collocations. The very low failure rate suggests that the Saudi students in this study were able to overcome many of the problems identified in other studies as being the cause of such failure. Wray (2000:270) states that, although some implicit approaches of learning the language by natural exposure can be effective, they still have to be administered because they might confuse learners struggling to distinguish between what is acceptable and what is not acceptable to native speakers. To Wray, learners can learn certain vocabulary when exposed to TV, radio, the Internet, or direct contact with native or non-native language speakers, but they do not always acquire them in a native-like way. Kuo (2009:143) also thinks that this L2 learning process overlaps between exposure and instruction.

Fan (2009:111), investigated Chinese ESL learners, and claims that learners' main problem in producing collocations is their lack of L2 exposure. The reason for this could be due to the nature of the learners' L1s as Fan suggests that Chinese learners' L1 had more influence than their exposure to L2. However, the findings of this study are not consistent with Fan's hypothesis, as the students in this study are not in an L2 environment yet are successfully producing a high percentage of acceptable collocations. Given that both Chinese and Arabic have different alphabets and grammatical structures from English this result suggests that for each of these languages there are areas which are closer to English than others. As Granger (2018) claimed that the big distance between learners' L1 and L2 may result in a better use of collocations as this enables them to distinguish between their mother and target language easily.

Siyanova and Schmitt (2008:431) stated that what makes language learners struggle with collocations is that they produce "unconventional combinations" that are not used in the same way by native speakers. This was also noted by Brashi (2009), who studied Saudi learners' collocational competence, stating that learners found it difficult to combine the vocabulary they had learnt; therefore, they end up with a language that does not sound native-like. The noun-noun combination *shingle hotel* is a good example of this, where a learner apparently knew the two nouns previously and tried to create a collocation out of them, especially as the textbook does not include the word *shingle*. This is an example of language extension which has failed. Other similar failed verb-noun combinations produced by learners was *cure soul,* and *arranged bags.* As they were judged as non-acceptable collocations by the native informants, they were corrected by them as *heal soul* and *organized bags.* The collocation list in the BNC obtained using the collocations tool in Sketch Engine shows some fixed collocations such as *nourish soul* and *purify soul* for *cure soul* and *pack bags* and *prepare bags* for *arranged bags* (Chapter Five, Section 5.4.1 and 5.4.2). While in my study the learners have produced a higher percentage of acceptable collocations than in the studies of Siyanova and Schmitt (2008) and Brashi (2009), the conclusions they reached in respect of the possible reasons behind the production of non-acceptable collocations find some support in this study.

There are some cases that show that while learners were exposed to and taught some combinations from their textbooks, they would still fail to produce accurate collocations. For example, the two combinations *food poisonous* and *poisoning snake* were not identified in the BNC nor by the native-speaker informants. However, they are very close to the acceptable collocations; *food poisoning* and *poisonous snake*. It is clear that the student who produced them was confused between the two collocates for the two nouns *food* and *snake*. Siyanova and Schmitt (2008:430) stated that sometimes when a learner knows two synonyms, they may assume that they are equal and could be used in different collocations and contexts, which is not always the case. This type of error was previously identified by Howarth (1998) and Nesselhauf (2003). Nesselhauf attributed this type of error to blending and mismatching, such as the example *pay care*, in her study, which mistakenly blends the two collocations: *pay attention* and *take care*. In this study, a lower-level learner wrote *took shots*, (in this context meaning received an injection). However, it was corrected as *took medicine* and *receive shots* showing that *took shots* was possibly produced mistakenly by blending and mismatching.

The issue of producing less idiomatic combinations by EFL learners as well as Saudi learners should not necessarily be considered to be the result of one particular factor such as their language proficiency levels as originally proposed in this study. Other important factors could include their exposure to L2, the relationship between their L1 and L2 and their language learning environment, which may play an equal role as a learner's proficiency levels in their written production of acceptable collocations. Indeed, it could be hypothesised that these other factors are actually more important than the learner's proficiency level.

Granger (2018) suggests that the differences in the number of learners' collocational errors recorded are related to way the research is conducted; for example, she refers to Nesselhauf's (2004) study where learners produced a large number of errors when investigating their grammatical and lexical accuracy. Whereas in Laufer and Waldman's study (2011) learners produced few errors when lexical accuracy only was investigated. This current study is similar to Laufer and Waldman (2011) in that it identifies only lexical collocations and similarly a low percentage of errors were recorded. As a result, it could be said that the nature of the investigation plays a role in the results and different findings across studies.

Despite the different contexts and nature of the methodology, research has shown that there is a link between the types of collocations and learners' L1, however it is not clear if it is a positive or negative link. Granger (2012:11) stated that L1 influence is the most widely discussed in learner corpus research, and that most corpus-based studies are consistent with it being the reason behind most EFL learners' inappropriate collocations. Laufer and Waldman (2011:651) show that even learners who are aware of L2 collocations through classroom instruction, still face difficulty in using collocations which can be influenced by their L1. They found, in three different levels, that 89% of their Hebrew learners' errors were caused by L1 interference. Saudi learners' outcomes in this study can correspond with Laufer and Waldman's results as some examples occurred in the two levels, and their given corrections by the native speaker informants suggest the possibility of the learners' L1 influence.

For example, a pre-intermediate level learner produced *forgot bags* as in "*we forgot our bags*", which was not identified in the BNC nor by native speaker informants' judgement. Other two instances also occurred in the same level where the learners used the verb *forgot* with other subjects such as *tickets* and *passport* as in "*we forgot our tickets*" and "*he forgot the passport*". While these two combinations *forgot tickets* and *forgot passport* were not identified in the BNC, the majority of native speaker informants agreed on them as acceptable collocations. Informants who considered them as non-acceptable collocations provided the verb *left* as a correction. In Arabic, the Arabic verb meaning *forget* is used normally in a context such as the sentences produced by the learners. The informants provided a correction for the collocation using the verb *leave*. In the Arabic context, people '*forget things'*, but in English the usual structure is that something is left or accidentally left, not forgotten. The Arabic-English Dictionary for the Use of Students (Hava, 1915:768) gives an example of the use of the Arabic verb *naseya* نسيmeaning *forget* in English, as in *small things left by a traveller*, which explains the learner's use of this verb in the collocation. Also, the Concise Arabic-English Lexicon of Verbs in Context (Abdou and Hassanein, 2011) gives two forms of the Arabic verb as *nasa* نسى and *naseya* نسي (meaning *to forget* in English) through the example *naseeto meftahee fel maktab* نسيت’ مفتاحي في المكتب , meaning in a literal translation *I forgot my key in the office*. This suggests that the learner may have been influenced by her L1 using a literal translation to its L2 equivalent. This study is consisting with Granger's work highlighting the importance of L1 influence.

Learners may guess a word using their L1 (Kasahara, 2011:2) such as in congruent and incongruent collocations (Kurosaki, 2012). Conversely, Granger (2018:232) suggests that the greater distance between learners' L1 and L2 affects their production of collocations positively as learners would be able to differentiate between the two languages more clearly. Learners in this study could relate, in only a very few instances, their L1 to their production of collocations. However, from this study it is not clear if the greater distance between Arabic and English affects learners' production of collocations positively, as suggested by Granger. Using a much larger dataset or an Arabic corpus could help to answer this assumption. Equally, it is not clear if any similarities between L1 and L2 have a positive effect as posited by Kasahara. This is an area where further research could be useful to develop clearer theories for the collocation production by Arabic L1 learners of English.

6.3.3 Similarities and differences between pre-intermediate and intermediate-level learners' production of acceptable collocations

Studies such as Farooqui's (2016) suggest that the learners' language proficiency level will have a positive impact on their collocational use; however, most studies looking at this specifically reported only a slight difference, such as Abdul-Fattah and Zughoul (2001) and Laufer and Waldman (2011), while others such as Chiu and Hsu (2008) found there was no significant differences in learners' use of collocations across different language proficiency levels.

This study makes a contribution in examining the production of acceptable collocations by two levels of EFL learners from the same L1. Most studies that have tested the impact of learners' language proficiency levels have examined only one type of collocation, usually the verb-noun type such as Laufer and Waldman (2011), Huat (2012) and Ebrahimi-Bazzaz et al. (2015). Studies that have examined the use of a variety of types of collocations among different levels of learners did not use similar methodology to this study, for example, task-based studies such as Chiu and Hsu (2008). The importance of analysing learners' written texts for their production of collocation is that it provides results that are much closer to the learners' real knowledge of the language than those of controlled tasks, which could be completed by guesswork. Even though Saudi learners' written production of collocation might be influenced by the classroom textbooks or the teachers, as suggested in Section 6.3.1, this effect is likely to be equal in both levels of learners as they are studying in the same context. This use of a free writing task can, indeed, add to the validity of this study when comparing the production of both levels whether for a better use in the higher level learner production or less efficient use in the lower level learner production.

The most obvious similarity is that, the learners at both levels produced a very high percentage of acceptable collocations as well as very few less idiomatic combinations as discussed in the previous sections. Due to the features given of the association measurement (see Chapter Two, Section 2.4.1) such as having a maximum value, it was feasible to identify the acceptable collocations in this study according to levels of fixedness i.e. low, medium and high level collocations. Most noticeably is that learners of both levels tend to produce more medium level collocations than high and low level collocations, with an exception for noun-noun collocations. In addition, both levels produce more verb-noun and noun-verb collocations in their writing than other types.

As an example, both levels produced a variety of different collocations using the verb *find*, which is a common and frequent verb for learners at these levels. The lower level produced collocations such as *find place* and *find family, find hotel* and *find clothes* whereas the higher level produced collocations such as *find home, find things, find parks* and *find peace*. Both levels of learners produced these acceptable collocations with all three levels of fixedness i.e. low, medium and high. Such collocations could be seen as simple and not really challenging nor needing the learners to be creative in producing them. However, they still succeeded when combining the verbs with nouns in noun-verb collocations in high-level ones such as: *life has, things happened, day came* and *father told*, which were written by the pre-intermediate level and *people have, people get, people think* and *things make*, which were written by the intermediate level. These examples also show that both levels' learners rely on frequent words. Even though it could be argued that the used bands of the LogDice in this study maybe low, learners showed their ability to use familiar verbs to produce them correctly in different collocations, especially when considering the different topics each level wrote about. Eventually, the majority of the two levels' production fall between medium and high level collocations. Paquot (2017:20), who investigated three levels of EFL learners, suggests that by using the BNC as well as COCA can provide useful information about the frequent words used by the students in their writing and about the learners' general knowledge about the language. This use of frequent words may explain the similarities observed across both levels of Saudi learners' written production of collocations, who tended to use high frequency words to form collocations.

This study is consistent with Paquot's (ibid.) in finding that the differences in learners' production of collocations is related to the types of collocations more than the learners' proficiency levels. Paquot found a small difference in her learners' production of adjective-noun collocations and the difference became more obvious in the verb-noun collocations. Similarly, the two levels' in this study showed differences in the production of adjective-noun and verb-noun collocations across the levels. However, the results from this study indicate that verb-noun collocations are not as challenging as adjective-noun collocations across the two levels investigated. Even though the differences are very small, Saudi higher level learners' production across the different types of collocations is still better than the lower level learners'. Such differences could be a result of the different association measures used, whereas Paquot used the MI score, testing strongly related to low frequency collocations, this study used the LogDice, testing exclusive frequent collocations. Similar to their production of verb-noun and noun-verb types, both levels of Saudi learners used common and frequent adjectives such as *good, bad, big* and *small* to produce collocations. The lower level produced these adjectives in mostly medium level collocations such as: *something bad, bad experience, good student, good hotel, small space, small restaurant, big day* and *big snake*. Among with some high level collocations using the same adjectives such as in *worst thing* and *big problem*. The higher level also produced the common adjectives given above in medium level collocations in examples such as *good doctors, good mood, bad side, big market* and *big school.* However, high level collocations using the same adjectives are found more frequently in both number and variety than in the lower level writing. *Small village*, *small house*, *big city*, *big companies*, *good idea*, *good job* and *bad thing* are examples of high level collocations found in the intermediate level students' written texts.

In contrast, noun-noun collocations produced by the Saudi learners were distinguishable between the intermediate and the pre-intermediate level writing. The pre-intermediate learners were more likely to produce quite well-known collocations such as *living room, security guard, room service* and *police station* whereas the intermediate learners seem to be more confident and creative when combining two nouns together as modifiers or collocations appear in the same list (or clause). Parkinson (2015) suggests that L2 exposure in ESL learners when compared to EFL counterparts is behind their creative uses of noun-noun collocation. However, in the context of this study, with L2 exposure being limited and unknown, this result could be due to the learners' higher level of English, which enables them to be more ambitious in their production of collocations. Intermediate level learners produced noun-noun collocations such as *rush life, traditions and customs, schools and universities, home town, roads and traffic, birds and animals, coffee shops, buses and trains, health and education* and *gas station*. The obvious variant in this case is the learners' proficiency levels and the different classroom textbooks used. Yet, due to the nature of the topics learners wrote about, being related to personal interests, the effect of textbooks might have less of an impact, especially, as the collocations produced were rarely used by the learners in both levels. Despite the fact that they are taught certain verbs, adjectives and nouns in their lessons, the Saudi learners showed capability in combining them into various different collocations. Particularly, in the use of the higher level writing which could indicate a result of their learning development being more creative and thus enabling them to combine nouns into acceptable collocations.

The English proficiency level of Saudi learners in this study did not seem to show a clear-cut or major impact on their written production of acceptable collocations; sometimes the groups performed equally, and at other times they showed a slight difference. However, Kim's (2003) study, whose learners were also not taught collocations, had different results. Kim found a strong link between Korean learners' proficiency and their production of different types of collocations, specifically in relation to their L1 influence. Kim tested high-school Korean EFL learners for their collocational knowledge in relation to their language proficiency with three out of four types of collocations and by using translation tasks and bilingual dictionaries. His results also suggest that students have better knowledge of vocabulary and how to use it as individual words than as collocations. Nevertheless, in this study learners, while not being taught collocations, and only being exposed to vocabulary such as Kim's learners, Saudi learners showed good collocational competence. This study only partially supports the conclusions reached by Kim in relation to language proficiency levels and collocational knowledge, that this correlation can correspond to the types of collocations being produced by the learners.

Similarly, higher language proficiency may not always guarantee a better performance by language learners as this study and others suggest. Lesniewska (2006) stated that language extension occurs when learners use their knowledge about the language to create additional language. Nesselhauf (2004), who investigated German EFL learners' production of verb-noun collocations, was not very confident that a learner with a higher language proficiency is necessarily a better producer of collocations. Chiu and Hsu (2008) support Nesselhauf in finding no relationship between learners' collocational production and language proficiency. According to Kuo (2009), a more language-proficient learner might extend their knowledge about the language to produce collocations, although they may not always succeed in producing acceptable ones. Saudi intermediate learners tend to be more confident in their language production; as a result, they extend their knowledge by using the L2 creatively in collocations. However, this creative language extension to produce collocations can result in either success or failure. In some occasions and most probably by chance, they create acceptable collocations such as *noisy life*, or they fail to do so such as in *noisy voices*. This study supports the findings of Nesselhauf, Chiu and Hsu and Kuo in respect of the successful production of collocations across different language levels.

This study also suggests that creative language extension, as described in the previous paragraph, is not exclusive to intermediate language learners. It can occur in pre-intermediate writing as well. According to Kuo (2009), language extension is one of the main causes of collocational errors and learners need to be taught collocation explicitly.

These types of errors appeared in the pre-intermediate Saudi learners' writing in the following two examples. The first is an adjective-noun collocation, *long planeting,* as in "*We arrived in Chine [China] after long planeting"*, which was not identified in the BNC neither by native-speaker informants' judgement. One of the informants provided a correction as *long planning.* However, the context as produced by the learner shows that the meaning she wanted to communicate is not *long planning*; it is most likely *flight.* Due to the learner's low proficiency level, it seems that she generated the word *planeting* from *plane* or *airplane*. The second is *quick order* in "*quick our order*" as a verb-noun combination, where the learner meant "make a quick order" or "make the order quickly" by using the adjective *quick* as a verb. However, the interpretation of this combination overlaps between the lexical choices made by the learner based on the collocation *quick order,* where *quick* is an adjective for the noun *order*, and the learner's extension of her grammatical knowledge of examples that have the same adjective-verb forms such as: *fast* and *slow*. Native speaker informants gave corrections such as: *placed an order in the quick order line, quickly ordered* and *made our order quickly.* This study agrees with Kuo, in that both levels studied show evidence of creative language extension causing collocational errors.

The two levels of Saudi learners; pre-intermediate and intermediate, show some similarities and differences in their written production of collocations, which does not indicate a strong impact of the learner's proficiency level on their written collocations. Granger (2018:234) writes that research investigating learners' production of collocation in relation to their proficiency levels revealed two things. The first, Granger and Bestgen (2014), shows that the more advanced the learners are i.e. higher intermediate and above, the more acceptable collocations they produce generally or in a specific type of collocation. This is especially obvious when the gap between learners' levels is significant such as in Laufer and Waldman's (2011) investigation, which indicated a significant difference only between advanced and basic level learners, but not with the level in between them. The second Granger describes is with reference to Paquot's (2017) study, which suggests a link between different proficiency levels of learners' production of collocations to the type of collocation investigated. This study does not support the first view as the results do not show a big difference between levels. Also, the two levels are adjacent not as in Granger's study. However, this study appears to be in line with the second view which is illustrated through Paquot's findings which suggest that learners' production of different types of collocations can correlate with their level of proficiency. It is an issue that occurred in various studies and contexts, whether indicating a positive, negative or no relation, such as (Kim, 2003; Chiu and Hsu, 2008; Kuo, 2009).

6.4 Summary

The results of this study, when compared with other related studies in the literature, showed that the methodological approaches used and the definitions of collocations used in the studies can affect the results.

One of the strengths of this study is that it considered free-written texts produced by learners. This allowed considerable freedom of expression by the students. There is an advantage in analysing this type of free writing using corpora, being less controlled than task-based approaches that may encourage learners to only produce collocations that are familiar to them (Granger, 2018). Thus, the results of the study showed that students were able to create their own collocations from previously known or recently learned (from the course book) vocabulary, or from outside influences. In this respect this study is consistent with Wray (2000) when observing that it takes input from a number of sources to produce acceptable collocations, which leads to one of the main limitations of this study is that the possible source of this language knowledge was not investigated.

While collocation across its different types has been proven to be a challenging grey area for EFL learners (Nesselhauf, 2004, Fan, 2009, Durrant and Schmitt, 2009), the results of this study showed Saudi learners' production of collocations was more or less equal with the verb-noun and noun-verb types were the most frequent and successful. This supports other research done with Arabic L1 learners such as Shehata (2008). Studies looking at the production of other collocation types with different L1s produced different results, which may suggest that learners' L1 does have an impact on collocational production and in different types.

Some researchers such as Fan (2009) believe that unsuccessful production of collocations can be the result of lack of L2 exposure. However, this study is not consistent with this, as the learners are based in the country of their L1, with limited L2 exposure, yet achieved a vast majority of successful collocations.

In respect of approaches used by researchers, studies referred to in this study such as Sivanova and Schmitt (2008) and Kuo (2009), applied corpus approaches, but differently. It is not clear if this difference in results is due to the ability of the students, or the different approaches and measurements used to identify collocations.

EFL learners' research used to rely on either investigating high frequency collocations through the *t-score* test or low frequency collocations through the MI score, which showed a gap in studying learners' real use of collocations that could be addressed through using the LogDice (Fernández and Schmitt, 2015; Gablasova et al., 2017a). This study serves to reinforce Granger's (2018) belief that the results obtained when researching collocations depends very much on the way the research is conducted and with so many variables: the task type (controlled or free), the methodology and approach, (corpus-based or phraseology) or the association measurement of collocations, (LogDice, MI score, t-score) and the researcher's decision about what constitutes a successful collocation.

By using the LogDice, it was further possible to produce a scale to enable the comparison between the two level's productions of collocations. Thus, it did not only show the learners' ability to produce collocations, but also showed a level of fixedness of these collocations produced. The examples mentioned in the discussion show that both pre-intermediate and intermediate learners used sometimes the same adjectives, verbs and nouns, but only intermediate learners were able to produce a wider variety of higher level collocations.

Finally, looking at the issue of language proficiency level and successful production of collocations, this study came to similar conclusions as Chiu and Hsu (2008) who suggested that there is no great difference between adjacent language proficiency levels. There is a tendency across levels to use familiar and frequently used nouns, verbs and adjectives, and naturally, the more frequent the word, the more likely it is to make successful collocations with other words. The two levels of Saudi learners showed the ability to combine words they already know and produce them as collocations. This similarity is also shown in the fewer number of the less idiomatic combinations that higher level learners produced than lower level learners. However, the most noticeable difference between the levels was in the production of noun-noun collocations.

The study concludes by suggesting that Arab EFL learners are capable of producing a very high percentage of acceptable collocations, which are frequent as well as exclusive and in different types even when they are not receiving explicit teaching on collocations. Thus, each study's understanding of collocation and how to approach it contributes to its findings and to what extent learners produce it.

Chapter 7 Conclusion

7.1 Introduction

While the production of collocation by EFL learners has been shown to be challenging (Huang, 2001), and even though it has been widely addressed in the literature, there are still some areas which needed further investigation. This study focuses on three of them: to explore if Saudi university students produce acceptable collocations and in what types, to investigate which less idiomatic combinations those learners produce, and what possible similarities and differences in their production of acceptable collocations can be observed at different proficiency levels.

By following the understanding of collocation developed by Sinclair (1991), Hunston (2002), Stubbs (2002) and Brezina et al. (2015), this study investigated collocations produced by Saudi learners in their written texts, according to frequency, span and exclusivity. The findings show that Saudi university students are able to produce a very high percentage of acceptable collocations. This was found for both levels of learners investigated: pre-intermediate and intermediate, and across the different types of collocations. These findings contribute to the literature by showing learners' capability to produce and develop their written production of acceptable collocation with a level of fixedness without being taught collocations explicitly. This could be happening due to some external influences outside the classroom as well as formal teaching. This study produced different findings to most studies, which suggest that students do not produce fixed collocations satisfactorily; these include work by Nesselhauf (2004), Siyanova and Schmitt (2008) and Fan (2009). Only a very few studies had similar results to this one; for instance, the work by Kuo (2009).

This study's most important finding supports a claim that external influences could impact positively on learners' collocation production, given that in this study they produced such high percentages of acceptable collocations in three levels of fixedness without them being given explicit instructions about collocations in the classroom.

In order to answer the research questions, it was important to combine two approaches of studying collocations: corpus-based and phraseological approaches. Although the two approaches have been used together in a similar way in previous studies such as Parkinson (2015), this study identified some limitations in applying them. In addition, the manual analytical procedure used to extract candidate combinations and the application of the association measurement the LogDice to identify each collocation's fixedness contributes methodologically to the literature. The study concludes with implications related to language teaching and learning in the context of EFL learners.

7.2 Implications of the study

Scholars have proposed that language teachers often focus their lessons on grammar rather than lexical patterns. Hill (2000) argues that teachers tend to correct grammar issues and overlook collocational mistakes and that, as a result, learners continue to make these mistakes. While the learners in this study have been given more attention grammatically than lexically, as consistent with Hill's claims, this study does not suggest that they produce a lot of lexical mistakes. Indeed, they were able to produce lexical combinations to a very high extent. However, it could be considered useful, as suggested by other researchers to teach collocation specifically in the classroom. For example, this could be done through focusing on some certain types that have been overlooked in the literature, i.e. noun-noun collocations. Lewis and Conzett (2000) suggest that language teachers should make learners aware of their production of collocations. This can possibly be done through discussing learners' production, whether correct or incorrect, in their writing or by being given tasks in the classroom. Correct collocational production can be further encouraged by showing the learners some original native speaker output such as that available in corpora, especially as the teachers in this study have shown some possible influence on learners' collocational production somehow i.e. in the discussed examples of *holiday* and *vacation* (Chapter Six, Section 6.3.1). Teachers can refer to corpora, such as has been done in this study, to evaluate collocations produced by learners in classrooms. They can examine the production of their learners and compare it to the language of native speakers, and thus, show them examples from naturally-occurring L2 language. Kasahara (2011:2) assumes that teaching learners through L2 contexts helps learners not only to understand the collocations, but also enables their production within grammatical frameworks and creating of successful meanings.

The learners in this study are only exposed in the classroom to traditional methods of language learning such as textbooks, audio CDs, and rarely any additional materials. While textbook exercises tend to explain vocabulary in some sentences and reading paragraphs, learners may benefit from collocational knowledge that is beyond this. Hussein (1998) states that teaching individual vocabulary items with a greater collocational focus and encouraging greater exposure to L2 media and books could be more effective than using classroom textbooks alone. Such techniques could help learners to acquire as much knowledge as possible from more naturally produced language compared to especially created examples in textbooks. Amending teaching materials requires also modifying teachers' classroom practices. According to Hunston (2009), teaching students collocation can be as simple as teaching pronunciation or word classes as just another part of vocabulary lessons. If two words are taught explicitly as a collocation and in a given context, the chance of producing less idiomatic combinations would be reduced. Thus, collocations could be learnt by both teaching and exposure. The two methods complement each other, as the first constructs the use of the language, and the second gives an understanding of the native-like use of the language.

Although this study has not directly investigated pedagogical practice, it is still able to link some implications drawn from the results of this study to the field of L2 learners and collocation research. Many studies in the literature, as discussed in Chapter Six, suggest that the issue of successful collocation production could be related to the learner’s L1; as a result, teaching materials should consider learners’ context and L1. Granger (1998) believes that not all language learners require the same material. Brashi (2009:27) suggests that even the differences in learners’ native language such as the different dialects and patterns in Arabic specifically, can be used to encourage successful production of collocations.

Bahns (1993:56) writes that there are still some collocations which are unique to the L2 and should be taught specifically to EFL learners. Huang (2001:15) also suggests that learning a new word implies learning its cultural and semantic aspects, one of which is the collocations it can make. Learners may tend to carry across their L1 language knowledge to the L2, which may not necessarily produce new collocations successfully. Thus, it could be beneficial for teachers to study their students’ actual language use and production, providing an opportunity to spot problematic areas and then track changes or developments in a specific group of learners, which can assist the teaching-learning process. Granger (2009) believes in the importance of comparing the production of L2 learners in relation to their L1, as it helps teachers to adjust teaching materials to learners' needs and interests. Even when learners are at an advanced level (Nesselhauf, 2004), or using bilingual dictionaries (Kim, 2003; Bahumaid, 2006), their L1 can still influence their production of collocations, though this study did not investigate that issue specifically.

The relative unimportance of learners’ language proficiency has also been observed in this study. As the difference between the two levels was minimal in the accurate production of collocations as well as in the less idiomatic combinations, the aforementioned implications apply to both lower and higher levels of learners. This also could suggest that Saudi learners are picking up L2 unconsciously, and producing acceptable, according to this study, L2 collocations. Even though there were many examples of easy or non-challenging collocations, the fact that the non-acceptable collocations were very few is still important. While not being given explicit lessons on collocations, and most probably limited in L2 exposure, this implicit language learning is interesting in Second Language Acquisition (SLA).

7.3 Contributions of the study

Because of the need of the study to examine the written texts closely, it was decided to analyse them manually. This was done to extract candidate VAN combinations before checking them in the BNC for identification. In order to do this, a manual extraction procedure has been generated to follow through all the samples systematically. It was a challenging step in the methodology because of the learners’ low level, the weak grammatical structures they produced, and the issues caused by poor handwriting. Learners did not follow regular or even writing patterns in their texts. Several pilots were conducted before the methodology was finalized. Following Evert (2008), three criteria were considered to extract those combinations: according to surface, textual and syntactic co-occurrence. These criteria have led to how candidate combinations were identified and extracted from the texts before identifying them in the corpus. A two-word combination appearing within a five-word span as the surface co-occurrence and limited within a clause as the textual co-occurrence while having a syntactic relation. This set of criteria is also in line with Granger's (2018:230) view on how collocations are traditionally defined as "pairs of words that are in a syntactic relation, and display restricted commutability and some degree of semantic opacity or specialization". This method of analysing learners' writing provided the opportunity to explore the data in great depth and detail, and thus assisting with certain decisions throughout the analysis, such as the learners' production of combinations within a specific window of five-word being limited to a clause. Most studies in the literature usually investigate adjacent collocations or one type of collocation, or rely on extraction software programmes. As a result, a procedure had to be found that allows this investigation of collocations produced by this level of learners, yet taking into account the originality of the data, and the systematic application of extracting candidate combinations. In addition to the methodological contribution to the literature that this study provides by way of the manual analysis to extract candidate combinations, this study shows a practical application for the use of the LogDice score. Using the statistical association measurement, LogDice, particularly through the proposed scale, facilitates the study of collocations produced by learners, which as suggested by Granger (2018:239) does not only give information about collocations and non-collocations, but provides an in-depth insight and interpretation of the results. LogDice scores facilitated the comparison between the two levels of learners, as well as between the different types of collocations produced.

Collocation can vary according to the scope of the study, its researcher's understanding and approaches as well as participants under investigation. This study has demonstrated this by showing that EFL learners from different contexts are still not similar to each other in their production of collocations in their written texts. This study, along with others that have investigated collocations according to frequency and exclusivity (Siyanova and Schmitt, 2008; Parkinson, 2015; Fernández and Schmitt, 2015), further demonstrates that the way collocation is defined and approached depends on the research context and interest of the researcher, and as a result, can lead to different findings. Studies referred to have examined learners' production of collocation separately according to high frequency by using the t-score, and/ or exclusivity by using the MI score, which has always suggested a gap between learners' knowledge and production of collocation. Therefore, this study aimed to investigate learners' production of collocations according to the LogDice; testing high frequency as well as exclusivity. The findings suggest that these results resemble those results from studies investigating collocations according to high frequency as in Kuo's (2009) study. It has been suggested that learners tend to rely on high frequency collocations (Hunston, 2002; Durrant and Schmitt, 2009; Fernández and Schmitt, 2015), this study also supports this. This study further shows that EFL learners are capable of producing acceptable collocations of different types. Thus, this study contributes to the understanding that how collocations are defined and approached, according to exclusivity or rare exclusivity, can affect the research findings. However, this should not necessarily mean that EFL learners can always be credited with producing good results when using the LogDice.

Furthermore, the study contributes with its findings to understanding how the EFL learners' context can affect production of collocations. This study has shown similarities with other studies which support the theory that learners' language proficiency has only a slight influence on their production of collocation. The findings from this study suggest that there was not a major difference in the number of acceptable collocations produced by the two levels of learners as well as to the level of fixedness of the collocations. However, the only notable difference is that the higher level produced more as noun-noun collocations. Thus, the higher level showed some development in their production of collocations of this particular type compared to the lower level, which suggests that learners' proficiency levels could have an effect on particular types of collocation only when they share the same L1.

This research contributes to the literature by combining the study of these different types of collocations using corpus-based and phraseological approaches to explore the written production of two levels of EFL learners in an Arab context. This contribution supports the findings of previous studies that the types of collocations play a major role in EFL learners' use of collocations. Nonetheless, this study was not able to indicate additional influences such as suggested by the previous studies i.e. L1 and L2 environment. Thus, the study needs further experimental approaches to show precisely the effect of some influences such as L1 and external influences outside the classroom. It is not only the combination of corpus and phraseological approaches, but also combining learners' writing analysis with experimental methods can add more to the study rigorous.

In the Arabic learners' context, this study is important because very little research has been done: specifically with this L1 and there has been no previous study which applies corpus-based approaches. Siyanova and Schmitt (2008: 455) state that the importance of applying collocation research using corpora relies on the usefulness of showing language patterns used by certain language speakers, and can be used as a reference in teaching and testing. This study, and others, have found that the variations in learners' production is firstly linked to the types of collocations under investigations (Martyńska, 2004). Task-based studies on Arab learners such as Shehata's (2008) study have shown similar findings to this study, indicating that Arab learners' production of verb-noun collocations is better than adjective-noun collocations. Other studies suggest the influence of learners' L1 on collocation production such as Yamashita and Jiang (2011) and Fernández and Schmitt (2015). This would be a very useful area for further research in the Arab context. Thus, the study has bridged a gap between Arab L1s and other EFL learner's L1 studies on the written production of collocation as well as reflecting on a comparison between two levels of learners across different types of collocations.

7.4 Limitations of the study

There are a number of issues that were identified as limitations of the study. The first issue is methodological. In the first stage of the analysis, which was to extract candidate VAN combinations, the manual analysis was used rather than an electronic software programme. Given the study's focus and aims, it was decided that this is a better approach, especially given that the written texts were handwritten. The manual analysis, as well as the software programmes, may have some drawbacks. While the software may be better in any systematic application, they lack the facility to deal with low level writing with weak grammatical structures and lexical ambiguity. Some interesting examples can be found through human intervention rather than electronic software even though it is not guaranteed to be free from human error. Thus, using manual analysis should have valid and systematic steps otherwise it cannot be replicated.

Another limitation in the methodology is that the study was only analysing learners' written texts with no further experimentation. In addition, the writing samples were relatively small. The study covered and analysed the findings of only a small number of learners. Nevertheless, as has been discussed earlier (Chapter Four), a low number was chosen particularly because the study wanted to investigate learners' written production in depth qualitatively, as no existing studies of Arab learners have done so previously. This study is further limited to investigating pre-university students, who are learning English as a requirement of their university foundation-year. Different results may have occurred if the study was conducted with students majoring in English, or producing academic writing, unlike the study participants' who wrote about general topics i.e. personal experiences.

The last and most serious methodological limitation is in the use of corpora. First, the study relied on using the BNC only while there is a chance that learners could be influenced to produce another variety than the British English that they are taught. Wray (2002) stated that L2 learners are sensitive to the collocations they come across and tend to remember them. Although this should not affect their production of collocations because the study is testing their ability to combine words to form acceptable collocations with a level of fixedness, some examples will not appear in the BNC. Possibly, due to topic limitations of the corpus, learners' unconscious production of a different variety of English could be a reason for this non-appearance in the BNC. For example, in Saudi Arabia, most of the shopping centres are known as 'malls', an American expression, rather than 'shopping centres'. Collocations containing the word *mall*, such as *have mall, mall shop,* and *big mall* appeared in intermediate-level texts and were not identified in the BNC. The possibility of EFL learners' mixed production of the English varieties while being taught British English was observed by McKenzie (2004) with Japanese learners, Siyanova and Schmitt (2008) with Russian learners, Zhang and Hu (2008) with Chinese learners, Bikeliene (2015) with French, Spanish, and Russian learners and Hameed and Fatima (2016) with Saudi learners. However, due to the combined approach this study applied i.e. corpus-based and phraseological, native speaker informants judged them as acceptable collocations while being British English. This could indicate a further limitation for the BNC, which is its size and date. Thus, a combination of two or more corpora of different English types and possibly sizes and dates should be applied to minimize such limitations.

The other limitation related to corpora is in the use of the association measurement; the LogDice. Even though the reasons for using this measurement are justified in the literature review and methodology chapters, the study suggests a limitation to the application of this score. It has been suggested that the LogDice score gives a better indication of the learners' production of collocations rather than only applying the t-score which indicates only high frequency collocations, or the MI score which indicates only very exclusive, low frequency collocations (Fernández and Schmitt, 2015). This study's findings have shown the suitability of the LogDice score for examining learners' production of collocations by combining the two criteria of frequency and exclusivity. However, Granger (2018) suggests that the corpus-based study of learners' collocation errors alone does not necessarily indicate their good use or few errors. She (ibid., 2019:233) claims that "learners generally tend to play it safe, using collocations of which they are sure". In a similar manner, as the LogDice tests, learners' production of exclusive as well as high frequency collocations, may only show collocations that learners feel safe producing. They may use the same combinations repeatedly, or may be with a little creativity to produce new combinations. Thus, the use of the LogDice score alone is not enough to show learners' real knowledge of collocation in their written production. It would be more valid to combine this application with a further procedure. Nonetheless, this method could be used as an indication of learners' production more than testing their perception of collocation (ibid., 2018:228-229).

7.5 Suggestions for future research

Based on the limitations outlined in the previous section, further areas are suggested for consideration for future research. Even though LogDice could be the most suitable association measurement to identify learners' production of collocations, it would still be interesting to use the MI score testing exclusivity of low frequency collocations. There are no similar investigations which have been conducted on Arab learners to compare directly with this study's results. Also, most of the studies using association measurements combined the application of t-score and MI score. None have combined the LogDice with another test. Also, an experimental study could be useful in examining possible influences on learners' production of collocations such as L1 interference, explicit teaching of collocations or external influences and exposure to L2. This study could be further followed by some elicitation approaches such as given tasks and questionnaires or interviews with students. There is a need to combine the corpus and phraseological approaches with additional tasks.

Additionally, as this analysis has been finalized in early-mid 2018, it is worth noting that a newer version of the BNC is being launched at the end of 2018, and a revised analysis could be conducted with this new version of the BNC. This could lead to an update of the study findings and contribute further to existing research. The limitations of the current BNC's size and outdatedness will most probably be refined. Therefore, it will be advantageous to use a newer version. However, a combination of two native corpora i.e. the BNC and COCA could be beneficial, which may reduce the limitations of the current BNC, especially as the learners' production is not specific to any variety of English, or topics and vocabularies they may use.

Finally, a similar study could be conducted to test language proficiency effects on learners' production of collocations on a longitudinal approach, or between non-adjacent levels. The hypothesis of this study can be tested on a small number of learners over a longer period of time in a longitudinal study. As it was found that the Saudi EFL learners were able to produce a large number of acceptable collocations across both levels, research could be conducted comparing beginner level with advanced learners, and possibly with a larger number of learners. Furthermore, a comparison could be conducted between students of general English such as the learners in this study and specialized students such as those majoring in English.

References

ABDUL-FATTAH, H. S. & ZUGHOUL, M. R. 2001. Collocational competence of Arabic speaking learners of English: A study in lexical semantics. *ERIC document reproduction,* pp.1-19.

ABDUL-FATTAH, H. S. & ZUGHOUL, M. R. 2003. Translational collocational strategies of Arab learners of English: A study in lexical semantics. *Babel*, vol. 49, no. 1**,** pp.59-81.

ARTS, T. 2014. *Oxford Arabic dictionary: Arabic-English, English-Arabic*, Oxford University Press.

ARAZY, O. and WOO, C., 2007. Enhancing information retrieval through statistical natural language processing: a study of collocation indexing. *Mis Quarterly*, pp.525-546.

BAHNS, J. 1993. Lexical collocations: a contrastive view. *ELT journal,* vol. 47**,** no. 1**,** pp.56-63.

BAHNS, J. & ELDAW, M. 1993. Should we teach EFL students collocations? *System,* vol. 21**,** no. 1, pp.101-114.

BAHUMAID, S. 2006. Collocation in English-Arabic translation. *Babel,* vol. 52, no. 2**,** pp.133-152.

BARNBROOK, G., MASON, O. & KRISHNAMURTHY, R. 2013. *Collocation: applications and implications*, Springer.

BENSON, M., BENSON, E. & ILSON, R. F. 1986. *Lexicographic description of English*, vol, 14. John Benjamins publishing.

BIBER, D., CONRAD, S. & CORTES, V. 2004. If you look at…: lexical bundles in university teaching and textbooks. *Applied linguistics,* vol. 25, no. 3**,** pp.371-405.

BIKELIENĖ, L. 2015. Lithuanian learners’ English: British or American? *Verbum,* vol. 6**,** pp.29-40.

BLOOR, T. & BLOOR, M. 1995. The functional analysis of English: a Hallidayan approach. London: Arnold.

BRASHI, A. 2009. Collocability as a problem in L2 production. *Reflections on English language teaching,* vol. 8**,** no. 1, pp.21-34.

BREZINA, V., MCENERY, T. & WATTAM, S. 2015. Collocations in context: a new perspective on collocation networks. *International journal of corpus linguistics,* vol. 20**,** no. 2, pp.139-173.

BROWN, D., 2011. What aspects of vocabulary knowledge do textbooks give attention to? *Language teaching research*, vol. 15, no. 1, pp.83-97.

BURTON, G., 2012. Corpora and coursebooks: destined to be strangers forever? *Corpora*, vol. *7*, no. 1, pp.91-108.

CHIU, C. Y. & HSU, J.-Y. 2008. Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL journal,* vol. 10**,** no. 1, pp.181-204.

CHENG, W., GREAVES, C. and WARREN, M., 2006. From n-gram to skipgram to concgram. *International journal of corpus linguistics*, vol. 11, no. 4, pp.411-433.

COWIE, A. P. & HOWARTH, P. 1996. Phraseological competence and written proficiency. *British studies in applied linguistics,* vol. 11**,** pp.80-93.

COWIE, A. P. & HOWARTH, P. 1996. Phraseology-a select bibliography. *International journal of lexicography,* vol. 9**,** no. 1, pp.38-51.

DORNYEI, Z. 2007. Research methods in applied linguistics. Oxford University press.

DURRANT, P. & SCHMITT, N. 2009. To what extent do native and non-native writers make use of collocations? *IRAL-international review of applied linguistics in language teaching,* vol. 47**,** no. 2, pp.157-177.

EBRAHIMI-BAZZAZ, F., SAMAD, A. A., BIN ISMAIL, I. A. & NOORDIN, N. 2015. Verb-Noun collocations in written discourse of Iranian EFL learners. *International journal of applied linguistics and English literature,* vol. 4**,** no. 4, pp.186-191.

ELLIS, R. 1994. *The study of second language acquisition,* Oxford University press.

ELLIS, R. & BARKHUIZEN, G. P. 2005. *Analysing learner language*, Oxford University press, USA.

EVERT, S. 2008. Corpora and collocations. *Corpus linguistics. An international handbook,* 2**,** pp.1212-1248.

FAN, M. 2009. An exploratory study of collocational use by ESL students–A task based approach. *System,* 37**,** 110-123.

FAROOQUI, A. S. 2016. A corpus-based study of academic-collocation use and patterns in postgraduate Computer Science students' writing. *Doctoral dissertation,* University of Essex.

FERNANDEZ, B.G. and SCHMITT, N., 2015. How much collocation knowledge do L2 learners have? *ITL-international journal of applied linguistics*, vol. 166, no. 1, pp.94-126.

FIRTH, J. R. 1957. Modes of meaning. Papers in linguistics 1934-51, pp.190–215. Oxford University press.

FIRTH, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistics analysis.*

FIRTH, J. R. & PALMER, F. R. 1968. *Selected papers of JR Firth, 1952-59 ed. FR Palmer,* London & Harlow.

FORSBERG, F. 2010. Using conventional sequences in L2 French. *IRAL-international review of applied linguistics in language teaching*, vol. 48**,** no. 1, pp.25-51.

FRAENKEL, J. R., WALLEN, N. E. & HYUN, H. H. 1993. *How to design and evaluate research in education*, McGraw-Hill New York.

GABLASOVA, D., BREZINA, V., MCENERY, T., 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning: A Journal of Research in Language Studies*, vol. 67, no. S1, pp.155-179.

GABLASOVA, D., BREZINA, V., MCENERY, T., 2017. Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information. *Language Learning: A Journal of Research in Language Studies*, vol. 67, no. S1, pp.130-154

GRANGER, S. 1998. Prefabricated patterns in advanced EFL writing: collocations and lexical phrases. *Phraseology: theory, analysis and applications***,** pp.145-160.

GRANGER, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching,* vol. 33**,** pp.13-32.

GRANGER, S. 2012. Learner corpora. *The encyclopaedia of applied linguistics*. Blackwell publishing ltd.

GRANGER, S. 2018. Formulaic sequences in learner corpora: collocations and lexical bundles. In Siyanova-Chanturia, A. and Pellicer-Sanchez, A. (Eds.). *Understanding formulaic language: a second language acquisition perspective*, chap. 12. Routledge.

GRANGER, S. & MEUNIER, F. 2008. *Phraseology: an interdisciplinary perspective,* John Benjamins publishing.

HALLIDAY, M. A. K. 2004. *Lexicology and corpus linguistics,* Bloomsbury publishing.

HAMEED, M. & FATIMA, P. 2016. Varieties of English and Saudi EFL learners' acquisition of spelling and pronunciation: Inculcating 'Good' language habits. *International J. Soc. Sci. & Education* vol. 6, no. 1, pp.50-54.

HANDL, S. 2008. Essential collocations for learners of English: the role of collocational direction and weight. *Phraseology in foreign language learning and teaching,* vol. 43**,** p.65.

ḤASANAYN, A. Ṭ., EL SEOUD, D. Y. A. & ABDOU, K. M. 2011. *The concise Arabic-English lexicon of verbs in context*, American University in Cairo press.

HAVA, J. G. 1915. *Arabic-English dictionary*. Catholic press.

HEID, U. 1994. On ways words work together-topics in lexical combinatorics. *Eurolex*, pp.226-257.

HENRIKSEN, B. 2013. Research on L2 learners' collocational competence and development–a progress report. *L2 vocabulary acquisition, knowledge and use. Eurosla monographs series* 2**,** pp.29-56.

HILL, J. 1999. Collocational competence. *Readings in methodology: a collection of articles on the teaching of English as a foreign language, 2006,* pp.162-167.

HILL, J. 2000. Revising priorities: from grammatical failure to collocational success. *Teaching collocation***,** pp.47-69.

HOWARTH, P. 1998. Phraseology and second language proficiency. *Applied linguistics,* vol. 19**,** no. 1, pp.24-44.

HUANG, L.-S. 2001. Knowledge of English collocations: an analysis of Taiwanese EFL learners. *Texas papers in foreign language education,* pp.113-132.

HUAT, C. M. 2012. Learner corpora and second language acquisition. *Corpus applications in applied linguistics***,** pp.191-207.

HUDDLESTON, R. & PULLUM, G.K. 2002. Preliminaries. *The Cambridge grammar of the English language,* chap. 1. Cambridge University press

HUNSTON, S. 2002. *Corpora in applied linguistics*, Cambridge University press.

HUNSTON, S. 2006. Corpus linguistics. *Linguistics,* vol. 7**,** pp.215-244.

HUNSTON, S. 2009. The usefulness of corpus-based descriptions of English for learners. *Corpora and language teaching,* vol. 33**,** p.141.

HUSSEIN, R. F. 1998. Collocations revisited. *Language and translation, journal of King Saud University* vol. 10, pp.39-47.

HYLAND, K., HUAT, C. M. & HANDFORD, M. 2012. *Corpus applications in applied linguistics*, A&C Black.

IBRAHIM, N., MUSTAFA, J., NAMVAR, F. & NOR, N. F. M. 2012. Analysis of collocations in the Iranian postgraduate students' writings. *3L: language, linguistics and literature, the Southeast Asian journal of English language studies,* vol. 18**,** no. 1, pp.11-22.

JICK, T. D. 1979. Mixing qualitative and quantitative methods: triangulation in action. *Administrative science quarterly,* vol. 24**,** no. 4, pp.602-611.

JOHANSSON, S. 1991. Computer corpora in English language research. *English computer corpora: selected papers and research guide***,** pp.3-6.

KASAHARA, K., 2011. The effect of known-and-unknown word combinations on intentional vocabulary learning. *System*, vol. *39*, no. 4, pp.491-499.

KILGARRIFF, A. and KOSEM, I., 2012. Corpus tools for lexicographers. In Granger, S. and Paquot, M. (Eds). *Electronic Lexicography*. Oxford University press.

KIM, N.-B. 2003. An investigation into the collocational competence of Korean high school EFL learners. *English teaching,* vol. 58**,** pp.225-248.

KUO, C. 2009. An analysis of the use of collocation by intermediate EFL college students in Taiwan. *Arecls,* vol. 6**,** pp.141-155.

KUROSAKI, S. 2013. *An analysis of the knowledge and use of English collocations by French and Japanese learners,* Universal-Publishers.

LARSSON, T. 2012. On spelling behavio (u) r: a corpus-based study of advanced EFL learners' preferred variety of English. *Nordic journal of English studies,* vol. 11, no. 3, pp.127-154.

LAUFER, B. & WALDMAN, T. 2011. Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language learning,* vol. 61**,** no. 2, pp.647-672.

LENNON, P. 1996. Getting 'easy' verbs wrong at the advanced level. *IRAL-international review of applied linguistics in language teaching,* vol. 34**,** no. 1, pp.23-36.

LESNIEWSKA, J. 2006. Is cross-linguistic influence a factor in advanced EFL learners' use of collocations. *Cross-linguistic influences in the second language lexicon. Clevedon: multilingual matters***,** pp.65-77.

LEWIS, M. & CONZETT, J. 2000. *Teaching collocation: Further developments in the lexical approach,* Cengage Learning.

LI, C. 2005. A study of collocational error types in ESL/EFL college learners' writing. *Unpublished MA thesis,* University of Ming Chuan, Taiwan.

LI, J. & SCHMITT, N. 2010. The development of collocation use in academic texts by advanced L2 learners: a multiple case study approach. *Perspectives on formulaic language: acquisition and communication***,** pp.22-46.

MARTON, W. 1977. Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level. *Interlanguage studies bulletin***,** pp.33-57.

MARTYŃSKA, M. 2004. Do English language learners know collocations. *Investigationes linguisticae,* vol. 11**,** pp.1-12.

MCENERY, T., WILSON. A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University press.

MCENERY, T., XIAO, R. & TONO, Y. 2006. *Corpus-based language studies: an advanced resource book.* Taylor & Francis.

MCKENZIE, R. M. 2008. Social factors and non-native attitudes towards varieties of spoken English: a Japanese case study. *International journal of applied linguistics,* vol. 18**,** no. 1, pp.63-88.

MEECHAI, D. & CHUMWORATHAYEE, T. 2015. Verb+ Noun collocational competence of Thai university EFL students: a comparative study of a regular program and an English program. *LEARN journal: language education and acquisition research network,* vol. 8**,** no. 2, pp.145-160.

MEL'ČUK, I. 1998. Collocations and lexical functions. *Phraseology: theory, analysis and applications,* pp.23-53.

NATION, I. S. 2001. *Learning vocabulary in another language,* Cambridge University press.

NATION, P. & SHIN, D. 2008. Beyond single words: the most frequent collocations in spoken English. *ELT journal,* vol. 62**,** no. 4, pp.339-348.

NELSON, M. 2000. Corpus-based study of the lexis of business English and business English teaching materials. *Unpublished PhD thesis*, University of Manchester, Manchester.

NESSELHAUF, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics,* vol. 24**,** no. 2, pp.223-242.

NESSELHAUF, N. 2004. Learner corpora and their potential for language teaching. *How to use corpora in language teaching,* vol. 12**,** pp.125-156.

NESSELHAUF, N. 2005. *Collocations in a learner corpus,* John Benjamins publishing.

PARKINSON, J. 2015. Noun–noun collocations in learner writing. *Journal of English for academic purposes,* vol. 20**,** pp.103-113.

PAWLEY, A. & SYDER, F. H. 2014. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. *Language and communication,* pp.191**,** 225. Routledge.

PAQUOT, M., 2017. The phraseological dimension in interlanguage complexity research. *Second language research*, pp.1-25.

RYCHLÝ, P. 2008. A lexicographer-friendly association score. *Proceedings of recent advances in Slavonic natural language processing, RASLAN***,** p.6.

SAMADJA, F. 1993. Retrieving collocations from text: Xtract. *Computational linguistics,* vol. 19**,** no. 1, pp.143-177.

SCHMITT, N. 2006. Formulaic language: Fixed and varied. *ELIA: estudios de Linguistica Inglesa Aplicada,,* vol. 6, pp.13-39.

SCHMITT, N. 2010. *Researching vocabulary: a vocabulary research manual,* Springer.

SHEHATA, A. K. 2008. L1 influence on the reception and production of collocations by advanced ESL/EFL Arabic learners of English. *Doctoral dissertation,* Ohio University.

SINCLAIR, J. 1991. *Corpus, concordance, collocation,* Oxford University press.

SINCLAIR, J. 2003. *Reading concordances: an introduction,* Pearson Longman.

SINCLAIR, J. 2004. *Trust the text: language, corpus and discourse,* Routledge.

SIYANOVA, A. & SCHMITT, N. 2008. L2 learner production and processing of collocation: a multi-study perspective. *Canadian modern language review,* vol. 64**,** no. 3, pp.429-458.

SIYANOVA-CHANTURIA, A. & MARTINEZ, R. 2015. The idiom principle revisited. *Applied linguistics,* vol. 36**,** no. 5, pp.549-569.

SOARS, L. & SOARS, J. 2011. New headway plus special edition: intermediate. Oxford University press.

SOARS, L. & SOARS, J. 2011. New headway plus special edition: pre-intermediate. Oxford University press.

STUBBS, M. 1995. Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of language,* vol. 2**,** no. 1, pp.23-55.

STUBBS, M. 1996. *Text and corpus analysis: computer-assisted studies of language and culture,* Blackwell Oxford.

STUBBS, M. 2001. *Words and phrases: corpus studies of lexical semantics,* Blackwell publishers Oxford.

STUBBS, M. 2002. Two quantitative methods of studying phraseology in English. *International journal of corpus linguistics,* vol. 7**,** no. 2, pp.215-244.

STUBBS, M. 2007. On texts, corpora and models of language. *Text, discourse and corpora: theory and analysis***,** pp.127-161.

STUBBS, M. 2009. Memorial article: John Sinclair (1933–2007) the search for units of meaning: Sinclair on empirical semantics. *Applied linguistics,* vol. 30**,** no. 1, pp.115-137.

TEUBERT, W. 2004. Language and corpus linguistics. *Lexicology and corpus linguistics***,** pp.73-112.

TSUI, A. B. 2004. What teachers have always wanted to know–and how corpora can help. *How to use corpora in language teaching,* vol. 12**,** pp.39-61.

WEBB, S., NEWTON, J. & CHANG, A. 2013. Incidental learning of collocation. *Language learning,* vol. 63**,** no. 1, pp.91-120.

WRAY, A. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied linguistics,* vol. 21**,** no. 4, pp.463-489.

WRAY, A. 2009. Future directions in formulaic language research. *Journal of foreign languages,* vol. 32**,** no. 6, pp.2-17.

XIAO, R. & MCENERY, T. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied linguistics,* vol. 27**,** no. 1, pp.103-129.

YAMASHITA, J. & JIANG, N. 2010. L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly,* vol. 44**,** no. 4, pp. 647-668.

ZHANG, W. & HU, G. 2008. Second language learners' attitudes towards English varieties. *Language awareness,* vol. 17**,** no. 4, pp.342-347.

Appendices

Appendix 1: Request Form

UNIVERSITY OF LEEDS

To the Vice Dean of the English Language Institute (ELI),

I am undertaking my PhD research at the University of Leeds, School of Education. This research is a study of the writing of foundation-year students at King Abdulaziz University, to find out how students use English collocation and what areas may cause problems. I will be investigating two different levels of learners and exploring the differences between them. The purpose of this research project is to give the researcher an insight into Saudi learners' production of English collocation, and to understand the connections between the level of the learners and the kind of language they use. This cross-sectional study will be using writing samples from Levels 3 and 4 (pre-intermediate and intermediate). There will be no interviews with students or teachers, or any further task. All information will remain confidential, anonymous, and be used solely for the purpose of the study. Data will only be saved on the University of Leeds hard drive (my own account), and my personal laptop which is protected by a password.

I am writing this, with the attached information sheet, to ask for your permission to give the consent form to students in order to find out if any students are willing to participate in the research. An explanation of the research project and how the data will be processed and used will be given to them at the same time (and included in the consent form). Students will be informed about the confidentiality and anonymity of their information. Any student who wishes to withdraw can do so up to two months after the data collection, by contacting me by email.

I will go to campus to collect the writing samples from teachers having applied a random sampling for students' sections and samples for each level from students' written texts. After I choose the samples from each level, the remaining copies of the samples will be destroyed. Only the samples that are chosen will be typed up and stored as electronic files, and then analyzed according to the research methodology.

Thank you,

Huda Y. Khoja

Lecturer at ELI, KAU

PhD student at the University of Leeds

E-mail: hykhoja@leeds.ac.uk

Supervisors:

Professor: Alice Deignan: a.h.deignan@education.leeds.ac.uk

Dr. Richard Badger: r.g.badger@education.leeds.ac.uk

Appendix 2: Information Sheet

UNIVERSITY OF LEEDS

Introduction:

I would like to conduct my research in exploring the use of English collocation in the writing of foundation-year students of two levels at the English Language Institute (ELI) at King Abdulaziz University (KAU).

Goals:

I aim to conduct a corpus-based research for the writing of two levels, Levels: 3 and 4. I will then be able to identify and investigate the English collocation that students are using, and if more advanced levels of English learners are better users than the lower levels. I would also like to explore any errors made by students, and whether they are relevant to certain types of collocations.

Why Levels 3 and 4?

I have chosen Levels 3 and 4 specifically for this cross-sectional study because these students are asked to write longer paragraphs than the lower levels, and there are overlapping topics between both levels.

Plan:

First, I will choose the sections from each level by taking every third section from Level 3 and the same from Level 4. The total will include 180 samples from each level. Then, I will photocopy the texts as I am only allowed to have copies from students' writing while originals should be kept with the teachers. I will follow the same random sampling by taking every third written text, making 30 samples from each level. After that, I will contact the students to find out if they wish to participate with their writing in the research or not, and if so, give them the consent form to sign. The written texts that I have chosen will be transferred electronically to my personal laptop to use and analyze while the rest will be discarded and destroyed. An analysis will be written for each text individually, and then an overall review will be made for each level. A comparative study between the two levels will take place as the final stage before writing the findings. There will not be any interviews with students or teachers, or further tasks. I will not be making individual judgments about students or the teaching.

Thank you for your interest in this research,

Huda Y. Khoja

Lecturer at ELI, KAU

PhD student at the University of Leeds

E-mail: hykhoja@leeds.ac.uk

Appendix 3: Student Consent Forms

UNIVERSITY OF LEEDS

Student Consent Form

I am carrying out a project which will investigate student writing to find out how students use English vocabulary ‘collocation’, and what areas of English causes them problems. I will be investigating two different levels of classes and exploring the differences between students at these two levels. The purpose of this research project is to give the researcher an insight into Saudi learners' production of English, and to understand the connections between the level of the learners and the kind of language they use.

All the student written texts I collect will be anonymized. The texts will only be used for research purposes, and will only be seen by the researcher and her supervisors. The written texts and their analysis will be stored on the University of Leeds hard drive (the researcher’s account), and the researcher's personal laptop which is protected by a password.

I would be very grateful if you would agree to let me have a copy of your writing, but I will understand if you do not wish to do so, and there will be no penalty if you do not want to be a part of the study. Those who give their consent, but wish to withdraw later, can contact me by e-mail within two months from this date, after which time it will be difficult to process the required data.

	Tick next to the statements you agree
1. I confirm that I have read and understand the description of the project, and I have had the opportunity to ask questions about it.	

	Tick next to the statements you agree
2. I understand that my name will not be used in the research or any publications and that the research will play no part in my evaluation as a student.	
3. I understand that I can withdraw my consent and have my data withdrawn at any point up to two months after collection.	
4. By completing the form, I agree to take part in the above evaluation	

Name of student	
Student's signature	
Date	

Thank you for your interest,

Huda Khoja

Lecturer at KAU & a PhD student at the University of Leeds: hykhoja@leeds.ac.uk

Supervisors:

Professor: Alice Deignan: a.h.deignan@education.leeds.ac.uk

Dr. Richard Badger: r.g.badger@education.leeds.ac.uk

ARABIC CONSENT FORM

Researcher's name:

Title of research project:

I have been fully informed about the general aims of this project. I understand that:

- It was my choice to participate in this research project and I may at any stage withdraw my participation
- Any data I provide will be used solely for research purposes.
- My data may be used anonymously in research publications, academic conferences or seminar presentations.
- All information I give will be treated as confidential
- The researcher will to preserve my anonymity

<u>موافقة على الاشتراك فى جمع بيانات لأغراض البحث</u>

اسم الباحث:

عنوان البحث:

لقد حصلت على معلومات عامة عن أهداف هذا البحث، و أفهم أن:

اشتراكي في البحث اخياري و يمكنني الانسحاب من الإشتراك في أي وقت (و لكن لا يمكنني الإنسحاب من محاضراتي في الجامعة إذا كان هناك جمع بيانات لبحث تمت الموافقة عليه من قبل إدارة المعهد)

المعلومات اللتي سأدلي بها و النماذج اللتي سأقدمها ستستخدم لأغراض البحث فقط

أفهم أن ما أقدمه من بيانات قد ينشر (بدون ذكر أسمي) في مجلات علمية أو مؤتمرات و ندوات أو حلقات بحث

ستحافظ الباحثة على سرية هويتي و بياناتي التي أدليت بها وستيخزن البيانات أو يتلفها لاحقا بطريقة مناسبة

Participants name/ الاسم	ID/الرقم الجامعي	Date/التاريخ	Signature/التوقيع

Appendix 4: Native-Speaker Informants' Judgement Form

What is your native language?

If it is English, please specify British or American:

The following table includes collocations identified from Arab university students' writing, but are not found in corpora. In this task, I would like you to state whether you agree or disagree to consider these combinations as collocations or not. In case, you cannot decide, please choose 'Not Sure', and provide your preferred collocates for any of the two words in a collocation. P.S. The two words in any collocation below may appear within a span of five words, not necessarily as adjacent. Each collocation below is given with an example.

Collocation	**Example**	**Agree**	**Disagree**	**Not Sure, This collocation is better**
arranged bag	I woke up early and *arranged* my *bag*			
asked guard	we *asked* a security *guard*			
forgot tickets	we *forgot* our *tickets* at home			
take shots	We had food poisoning and went to the hospital to *take* the *shots*			
villages force	Some *villages* might *force* you to drive a long way to get to work			
had vacation	it was the worst *vacation* I *had* ever in my life			
father screaming	my *father* was *screaming* because of his stomach			
tourist guide	the *tourist guided* us			
waiter gave	The *waiter gave* us our food			
sister lost	my little *sister* was *lost*			
beautiful photos	We took *beautiful photos*			
green farms	I love the *green farms*			
bored time	we arrived after a long *bored time*			
huge road	I lost myself in a *huge road*			

Collocation	**Example**	**Agree**	**Disagree**	**Not Sure, This collocation is better**
big bookstores	It does not have *big* markets, malls and *bookstores*			
candies shops	There are a lot of *candies* and gifts *shops*			
buildings restaurants	We saw the high *buildings* and the modern *restaurants*			
small clinics	I can only find *small clinics*			
mall parks	This area has many *malls*, *parks* and cinema			
transportation well-developed	the public *transportation* isn't *well-developed*			
parks playgrounds	You can't find *parks* or *playgrounds*			
phone station	I drove a long way just to get to the *phone station*			
malls schools	they will have lots of *malls*, hospitals and big *school*			
shining stars	lay down and watch the sky full of *shining stars*			
clear life	Do you like the rush and noisy *life* or the peaceful and *clear* one?			
school complete	The hospitals and *schools* are not *complete* like in the city			
markets mall	There are some big *markets* and *malls*			
malls hospitals	they will have lots of *malls*, *hospitals* and big school			
forget town	they should not *forget* their home *town*			
find playgrounds	You can't *find* parks or *playgrounds*			
cure soul	You will be able to *cure* your *soul* with green areas around you			
like rush	Do you *like* the *rush* and noisy life or the peaceful and clear one?			
forget bags	we *forget* our *bags* in airport			
have malls	They *have* a lot of *malls*			
contains pollution	Many things make the city *contains* lots of *pollution*			
love farms	I *love* the green *farms*			
face crash	my *face crashed* to the ball			

Collocation	**Example**	**Agree**	**Disagree**	**Not Sure, This collocation is better**
Crowd city	You will get rid of all the noisy voices of the *crowd city*			
father surprised	my *father surprised* us with tickets			
people scared	*people* are often *scared* of things they have never tried			
hospitals complete	The *hospitals* and schools are not *complete* like in the city			
big mall	There are some *big* markets and *malls*			
good clinics	there aren't *good clinics*			
worst vacation	it was the *worst vacation*			
global schools	people have many options for education like, universities, *global schools* and institutes			
hospitals well-equipped	the *hospitals* aren't *well-equipped*			
sister sick	my *sister* started feeling *sick*			
forest farms	There is fresh air where there are *forest* and *farms*			
factories restaurant	you can’t smell the smoke of the *factories* and *restaurants*			
mall cinemas	They have been to places such as *malls*, *cinemas* and theatres			
cinemas parks	I like to go to *cinemas* and *parks*			
mall bookstores	It does not have big markets, *malls* and *bookstores*			
entertaining place	there were not many *entertaining places*			
candies gifts	There are a lot of *candies* and *gifts* shops			
markets bookstores	There are some big *markets* and *bookstores*			
shops malls	there are many places to go to like coffee *shops* and *malls*			
Food poisonous	When we went to the hospital, they told my father that was a *food poisonous*			

Collocation	**Example**	**Agree**	**Disagree**	**Not Sure, This collocation is better**
Noisy voices	you will get rid of all the *noisy voices* of the crowd city and the cars			
Forgot passport	He *forgot* the *passport*			
Quick order	we were very hungry So we *quick* our *order*			
Noisy life	wither you like the rush and *noisy life*			
Dancing fountain	we saw the *dancing fountain*			
Long planeting	We arrived to China after *long planeting*			
Poisoning snake	that isn't a *poising snake*			
Weather fantastic	The *weather* was *fantastic*!			
Shingle hotel	We checked in *shingle hotel*			
Daddy meal	my *daddy meal* was very cold			

Appendix 5: Pre-intermediate Level Written Texts

Text One

Life has a lot of experience something good and something bad. Whe [when] the experines [experience] was good, we would feel happy. But If the experiences [experience] was bad, we would feel sad. Before 10 years ago I visited my uncle with my mother because he was very sick and the doctors told him never leaved his bed. First, I played with his son. We started run and played football. Then, I sleppted [slipped] on a slant, when I was running. My accident made big sound because I fell on the flor [floor] and my face crash to the hall [ball]. I was standing, when my uncle leaved his bed and came to know the problem! he said: What's the matter? We answered: nothing. But he loocked [looked] to my face and said: You should play in the room. Firstly, I couldn't understand how could he knew. Then, I saw his son was looking to my face and smiled. At last, I saw my face in a meror [mirror]. I felt shy because of my face I was wite [white]!! The pouder [powder] of ball on my face. After this memory I never run on a slant. I learned the lesson and I must be quite [quiet] when I visit sick person.

Text Two

One day my family were sitting in the living room talking about our next trip in summer, they desided [decided] to go to another country, I was doing my final exams, so I feel a bit nervous because I have to study hard now, I couldn't talk with them, Then I left to my room and slept early. In the first day of vacation I discovered that I faid [failed] in the exam, I felt very dispointed [disappointed] and I was worried how to tell my parents, I went out to buy some coffee and I visited my friend, It was a horrible day, I kept thinking how to tell my family about my result but, I didn't know any way. Suddenly, when I was sitting in my room I heard knock on the door, I said who? but no bod [nobody] answerd [answered] to me, I desided [decided] to get down to see what are happen there, unfourtonately [unfortunately] I found lots of people sitting with my family, they were waiting for me to celeprate [celebrate] by my graduation, they didn't know I failed in the exam, so I felt embarrased [embarrassed], my mother asked me why you look very sad you must be happy today we are going to travel tomorrow. Then I told her about my exam she surprised about that but she was very kind with me. She told my dad about me then they desided [decided] to stay at home on that vacation. I spent that vacation with my books, then I done well, I learnt from my experiance [experience] to study hard befor [before] exams and I never be shy of my parents, it was the worst vacation I had ever in my life , but since that summer I become very good student.

Text Three

The worst vacation I have ever had was in 2012. First of that my father surprised us with tickets to Malaysia then we were very excited and started thinking and planning. So after two weeks its the big day we went to the airport but it was very crawded [crowded] and our trip going to leave at 7.30 p.m. We were so late then my father said he forgot the passport and he couldn't go back to home because there were trafic [traffic]. So we travelled to Malaysia without my father and I was crying this was the first time we travelled without him. When we arrived at Malaysia we didn't know what to do. My father wasn't with us we felt lost but their [there] was my mother's friend and she help us so we stayed on a beautiful hotel and get a rest. On the second day my father came to Malaysia and we were so happy to saw him. After that, we went to a restaurant to had our lunch. The restaurant was small and the food wasn't delicious. After we had our lunch we went to the petronas twin tower. When we arrived there my little sister lost and we tried to found her. We called the police to help us. Then they found her and she was crying. After that we went to hotel because we were very exhausted. At the night befor [before] we got to bed my father was screaming because of his stomache [stomach] and we tried to called the doctor but no one ansered [answered] then we want to took my father to the hospital but it was raining and we couldn’t find taxi so we stayed in the hotel and I was worried about him. On the last day my mother want to back to home, Jeddah, also me and my sisters. So all of us arrived to airport but they said we couldn’t travel because of the weather. Then we stayed in Malaysia one day. After that we arrived at Jeddah. This was my worst vacation I have ever had and I don’t want it to happen again we didn’t enjoy on this vacation and a lot of bad things happened to us.

Text Four

One month ago, I sat with my family in living room. My family decided to travel to Dubai. I felt so happy because I finished my exams So I needed to relax in somewhere. In the morning of that day. I woke up early and arranged my bag. At 8.00 a.m we went to Airport. After two hours we arrived in Dubai. Then we sat in famous hotel. It was oppisit [opposite] of Kalifah [Khalifa] twore [tower] and It was clean and nice place. After that we ate lunch in a resturant [restaurant] and we saw the danceing fountin [dancing fountain]. Then, we went to Dubai mall because we wanted to buy somethings [some things]. Then, my mother took me my sister and she told me “take care of your sister”. After that I saw a beautiful dress in a shop then I bought it suddnely [suddenly], I lost my sister and I felt scared because my sister didn’t know anything. I searched her but I couldn’t see. After that I told my mother she angry for me. Then we asked a securty [security] gured [guard] but they didn’t see her. After that we saw her in candy seller and she was scared. Then I felt so happy. At last I said to my mother “I’m sorry mama bout [about] that”, and I learned of that day I have to more responsible and take care of give me somethings [some things].

Text Five

The worst vacation I have ever had was in 2012 in Chine [China]. My parent bought tickets to Chine [China] as a surprise to us. We were very exciting because that was our first vacation out of Saudi Arabia. Then, we started packing our bags with full of happieness [happiness] befor [before] tow [two] weeks from the traveling [travelling] day. The big day cam [came] and we were full of eneargy [energy]. We went to the Airport and everything was good until we remebered [remembered] that we forgot our tickets at home. We were fried [afraid] that we will miss the plane. Fortuentely [Fortunately] my big brother came to the Airport after 10 minutes with the ticket. We arrived to Chine [China] after long planeting [?] We lost our first day because we were tired and we just falled in deep sleep. In the morning of the second day the bus of our tour came and took us big garden have a small space of fruit garden into a house that keep them warm. While we tasting the fruit we heared [heard] a loudly screaming and that was my litele [little] brother. He saw a big snake coming close of him and my mom [mum] started crying in her child. Fortuentely [Fortunately] the tourist guide us that isn't a poiesing [poisoning] snake. Then, we took beautiful photor [photos] to remember. At the miednight [midnight] of our last day I just waked up when I feeled sick and I couldn't stop feeling sick until the morning. My dad tryed [tried] to call a taxi to take us to the houspital [hospital] while the pain came strong and I just started crying. Evrything [everything] came wort [worse] when my sister starte [started] feeling sick then my brother. When we went to the houspital [hospital] they told my father that was a food poiesnios [poisonous]. After we took the shots we went to the Airport because we were late. When we arrived to Jeddah we just had pessful [peaceful] feels. It was the worst vacation I have ever had and I would never get back to that place.

Text Six

Last summer vacation it was the worst vacation we have ever had. I went to Africa with my family. We was exciting to go but when we arrived to Africa everything was bad. When we arrived at hotel we suddly [suddenly] realized that we forget our bags in airport so we went to airport to get our bags. Scound [second] day we woke up early to go to jungle but when we arrived we found a lot of people and it was so traffice [traffic,] but when we wanted to go back the car didn't work so we sat with another people in jungle and we went to hotel with Saudian family by their car. Third day we wanted to go to beatch [beach,] but my brother get flu so we sat at hotel because we can't go without him. My father called a doctor to come but he didn't come I don't know why maybe because the weather wasn't good. After five days we went to Saudi Arabia and my father told us sorry about the worst vacation you have ever had that was my idea to go to Africa but we said don't worry we learnt from this vacation we must search about the place that we want to go to before we go and ask another people about it.

Text Seven

My worst vacation that I never had, was on dec-20-2013. My family decided to go in this long vacation to Istanbul- Turkey. The week before, me and my brother's gathered in my room to think about what should we have in this trip. We checked googel [google] maps and serched [searched] about guide book [guidebook] and we got some information about that. We were so happy and very exited [excited] for this holiday. When we arraived [arrived] in Istanbul after, along [a long] , bored time that I spent in the plane. I was very tierd [tired] and my brother also, So we spent our first days in a hotel. then, at the therd [third] day my father decided with him silf [himself] that we have to move this hotel because there is a big problem happened when he asked the room serves [service] to clean our rooms. We spen [spent] our therd [third] day in a road to find some good hotel, Suddenly the sky came with a dark colors [colours] and started raining. When my family started walking to find some place I lost my silf [myself] in a huge road and I didn't no [know] what can I do and my phone was with my mom [mum] so I started cry until some one [someone] help me. After that , when I found my family, we want to have a dinner in resturant [restaurant] that called 'Semit' we were very hungry So we quick our order, then when the witer [waiter] gave us our food, my dady [daddy] meal was very cold and he was very angry.

Text Eight

In 2014 my sisters and brother and i [I] we decaded [decided] to go to new york city. We was so exited [excited]. We checked in shingle hotel and after that we went to Times square. The beautiful area in new york city. We saw the high billdings [buildings] and the modern restaurants and a lot of candys [candies] and gifts shops. Also we saw a lot of Asian people and African people. It was great! Then we went to Statue of liberty in center [centre] of New York City. We took pictures and also have fun together. The weather was fantastic! Not very cold or very hot! We are lucky about that. people was very nice. In the evening we went to indian restaurant we order some chicken teka [tikka]! and garlic bread! It was dilicuse [delicious]! The next Morning it's the Worst thing in The Vacation! We went to the breakfast area and when we finished we back to our rooms we saw our windows open! then we didn't found our clothes and bags! It's so tirroble [terrible]. After that we went to police staistion [station] and told him the whole story. We didn't sleep that night! but in the next morning at 10 am some police man [policeman] told as [us] who did that! And they captured him. Then they back all the bags for as [us]. We toguht [thought] we couldn't found it but the police men did it. We had a worst vacation in the end.

Appendix 6: Intermediate Level Written Texts

Text One

So many people still live in small villages because they still stick with their tardtions [traditions] and customs. However, they have to think of their future and improve their life. On the other hand, there are so many disadvantages of living in a small village. Firstly, people will not be able to be more cefictecaded [sophisticated] or open mind because there is no schools or univircities [universities]. Secondly, there is no important fucllitys [facilities] such as hospitals with good doctors and that will cause losing a lot body. There are a few advanteges [advantages] of living in small villages. Firstly, people can have a pure and fresh air because there is no factories or any other thing [that] can make the air polluted. Secondly, people can have a very calme [calm] and quite [quiet] live [life] with city nois [noise] or any other thing can bother them. To sum up, I recommend people to live in a big city because it has a lot of advanteges [advantages]. But, they should not forget their home town and visit it every once while. So many people still living in small villages for some reson [reason]. [?] Have a lot of disadvantages and a lot of advantages.

Text Two

There are thousends [thousands] of villages around the globe, and every day you hear about a new village's name. Often people who lives in villages don't mind moving out. However, most people are not a fans of villages and would never conseder [consider] living in one. Familys [Families], who live in a village mostley [mostly] have been there for generations. They just can't seem to fit any where eles [else]. For example, living in a village gets you closer to your family and neighbours, not like if you were living in a big city you wouldn't be able to see your family maybe for weeks. Also, there are a few people living in a village and not a lot of cars, therefore, its really quite [quiet] and calming. Finally, in a village there are a few stores with a few roads and no a trafic [traffic], so you can forget about long houres [hours] driving. On the other hand, lots of people prefare [prefer] living in a big city and even see it as tortutre [torture] if they had to visit a small village. First of all, people hate living in villages because they can’t find things to do or have fun with. In a village you are not able to find parks or playgrounds. Secondly, there are no options when it comes to schools for your kids you only take what you get and that might be in a diffrint [different] village. Lastly, there are no big hospitals with excellent care. You can only find small cleniqes [clinics] with a few doctors. To sum up, people oftenly scared of what things they have never tried doing. So unless you give living in a village a chance. You can’t really decied [decide] wheather [whether] it fits you or not.

Text Three

Home is a place where you can do whatever you want and feel to do it. The actuall [actual] place where you can find home is the most powerful thing to affect you and your personality such as village. Some might consedr [consider] it an old thing to it but some others are just love it. At the end i think living in a village can be a great thing to choose. There are some disadvantages to living in a village. First of all there is no internet and you can't be in touch with anyone whenever you want. Some villages might force you to drive a long way just to get to the phone station if you need to make a call. And because of that you'll get used to live alone without knowing anything about anyone. Also it doesn't cost much to live in a village but you will need a job which is cheap to afford this cheap life there and to bring you some action so you won't have to kill yourself of being bored. On the other hand, there are some advantages to living in a village. Firstly, you will get rid of all the noisy voices of the crowd city and the cars. Secondly, you will find peace, you will find yourself because you'll have a lot of time with yourself. You will think about every beautiful thing in your life and you will have time to lay down watching the sky fell [full] with the shining stars while you are enjoing [enjoying] the windy weather. You will be able to clear yourself and cure your soul with the green areas all around you which will make you feel that everything is going to be okay. To sum up, the advantages and disadvantages to living in a village might be on the same level, it's all depends wether [whether] you like the rush and noisy life or the peacful [peaceful] and clear one. But remember to choose a place where you can find yourself and feel good about it.

Text Four

Living in village it’s a good idea for healty [healthy] life you know fresh air a natural that’s great anyway living in village have their advantages and disadvantages. Thare [there] are many ………………. dis [disadvantages] first like that there they havn’t [haven’t] schools or hospital they have but not like the citey [city] and the house is sumall [small] and made of wood and it dosn’t [doesn’t] save second in the other hand, when it came dark the road it’s gonna [is going to] be dengros [dangerous] Because the woolfs [wolves] ands [and] bears living in village [?] a good idea for a healthey [healthy] life, as you know a fresh air and a natural place with a good and genrase [generous] people but living in village have advantages and disadvantages. There are many disadvantges [disadvantages] for living in a village. First, the houses made of wood. The hospitals and school its not complete like the citey [city] its need a lot that if they have some villages dosn’t [doesn’t] have. Second, when it came dark they sleep early Because they have alot [a lot] of working in the morning They Work hard and there another reason that itsn’t save [safe] Because of animals like the woolfs [wolves] and bears its dengruse [dangerous]. There are many advantges [advantages] of living in village. First, the fesh [fresh] air Because [because] of forest and farmes [farms] there is no nois [noise] at all just the birds and animals it’s a peace [peaceful] place. Second, the food there fresh and more weather than the citey [city] Because, it’s came from the farm to me and the people are genruse [generous] and kind they help each other and alwyse [always] together. They know every thing [everything] about like the family. Living in viliage [village] a good idea if you can live with the disadvantages and you love quite [quiet] please [places] anyway, it have a lot of advantages I would live there.

Text Five

The home town location of the human being effects on his body, brian [brain] and heart. Living in a big city is not similer [similar] as living in a small village. Each has its good and bad sides. Living in a village gives you the ability to live in peace. By living there, you will not be able to listen to the noice [noise] of the people, cars, buses and trains It also makes you able to live a hithier [healthier] life. For example, you can not [can’t] smell the smoke of the factories, big restaurants and cars. In addition, living in a village will make your life easier, more simple [simpler] and more comfortable such as you don’t have to suffer in the trafic [traffic] every morning to go to work. But, there are always two sides to everything. Living in a village can be boring sometimes when there are not many intertaining [entertaining] places to go like coffee shops, malls, cenimas [cinemas] and barks [parks]. Moreover, living there can be a hard thing sometimes for people who eat out always and people who need to go to hospitals a lot. Most importantly, living in a village might be exhausting sometimes because of not having big markets, mall, book stores [bookstore], ect [etc.]. And that will lead the people who live in villages to need to visit the big cities from time to another. Living in a village is not a bad thing at all. All you need is to know where you can live happy, relaxed and satisfied.

Text Six

Living in a village to some people is a good thing and for others it's not good becuae [because] some like to be in quite [quiet] plaes [places] and some didn't like it. There are a lots of advantages and disadvantages about living in a village. It's diffrent [different] from one to another becuae [because] evrey [every] person think of this essay from his oun [own] opinion. Living in a village has many advantages. When you live there you will be at good mood all the time. The people in the village so kind and lovely all of them know ech athor [each other]. In the village you won't need a car you just can walk between the houses and have great time. For some people living in a village mean the end of there [their] life becuae [because] they love the city they will have a lots of malls, hospitals and big school university. The city so croded [crowded] pleas [place] there is many cars gas staican [station], this things make the city contun [contains] lots of pollitoned [pollution] and that not good pleas [place] you can live there. There are millons [millions] of people from diffrent [different] countres [countries] so they don't know each athor [other]. At the end I think that living in a village is a great thing. I wisch [wish] that I can live there after I finish my educiton [education] in the city becuae [because] I love the green farms and the old houses, I like teka [take] car [care] of the animals and farms.

Text Seven

Villages haven't entertainment and there aren't a lot of things to do. Some people think that it is cheap and it isn't hard. However, I think that there are several disadvantages to living in a village related to culture, health and education. Some people will see a lot of advantages if you live in a village. Firstly, the life style [lifestyle] won't be very expensive like, renting a home and shopping. Secondly, in the villages there aren't a lot of people and cars on the streets and that can't waste your time. In addition, villages are quiet because there aren't factories and they are not crowded. And the people in a village often are friendly and kind. On the other hand, I believe that there are some disadvantages to living in a village. For example, in a villages, the people can't have many options for education like, universities, global schools and educational institutes. Moreover, a village hasn't good health centers [centres]. For instance, the hospitals aren't well-equipped there aren't good clinics. Also, the people in a village can't get the best job with good and high salaries. And village hasn't many entertainment and emenities [amenities]. Top sum up, I think that living in big cities is better than living in a village. I don't recommend living in a village for many reasons.

Text Eight

Choosing the right place to live in is very important. Alot [A lot] of people might not prefer to live in the village due to their needs. However there are still some people who want to live a simple village life. There are alot [a lot] of advantages for living in a village. Firstly, its a great place for the elderly; because it's quite and isn't crowded. Secondly, living in a village is very safe; everybody know each other so you won't be worried if your kids are playing in the street, Moreover, it's such a clean environment to live in; villages got alot [a lot] of trees. On the other hand, there are still some disadvantages. One is, living in the city can be little boring; there are not alot [a lot] of ways for entertainment and not that many places to visit. Furthermore, there are not many job offers; because big companies are usually in the city. Finally, the public transportation isn't well-developed so you should own your own car or bike. To sum up, I would like to say that both villages and also cities are good places to live in. If you want to know whats [what’s] best for you you should take in considaration [consideration] all of your needs.

Appendix 7: Candidate VAN combinations extracted from pre-intermediate level written texts organized according to their types, and included in contexts they appeared in

Table One

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
Text One	
life has	*Life has* a lot of experience
has experience	Life *has* a lot of *experience*
visited uncle	I *visited* my *uncle* with my mother
doctors told	the *doctors told* him
left bed	never leaved [*left*] his *bed* when my uncle leaved [*left*] his *bed*
played football	and played *football*
accident made	My *accident made* big sound
made sound	My accident *made* big *sound*
face crash	my *face crash* to the hall [ball]
uncle left	when my *uncle* leaved [*left*] his bed
uncle came	when my *uncle* leaved [*left*] his bed and came
know problem	to *know* the *problem*
saw son	I *saw* his *son*
son looking	his *son* was *looking*
saw face	I *saw* my *face* in a meror [mirror]
learned lesson	I *learned* the *lesson*
visit person	when I *visit* sick *person*
Text Two	
doing exams	I was *doing* my final *exams*
found people	I *found* lots of *people*
heard knock	I *heard knock* on the door
tell family	how to *tell* my *family* about my result
tell parents	how to *tell* my *parents*
visited friend	and I *visited* my *friend*
buy coffee	to *buy* some *coffee*
mother asked	my *mother asked* me
told dad	She *told* my *dad* about me
spent vacation	I *spent* that *vacation* with my books
vacation had	the worst *vacation* I *had* ever in my life
become student	since that summer I become very good *student*
Text Three	
vacation had	The worst *vacation* I have ever *had*

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
	my worst *vacation* I have ever *had*
father surprised	my *father surprised* us with tickets to Malaysia
trip leave	our *trip* going to *leave* at 7.30 p.m.
father said	my *father said*
forgot passport	He *forgot* the *passport*
had lunch	to had our *lunch* After we *had* our *lunch*
father came	my *father came* to Malaysia
get rest	and get a *rest*
called police	We *called* the *police*
things happened	a lot of bad *things happened* to us
stayed day	we *stayed* in Malaysia one *day*
mother want	my *mother want* to back to home, Jeddah
father screaming	my *father* was *screaming* because of his stomache [stomach]
called doctor	to *called* the *doctor*
took father	to *took* my *father* to the hospital
find taxi	we couldn't *find taxi*
Text Four	
finished exams	I *finished* my *exams*
arranged bag	and *arranged* my *bag*
ate lunch	we *ate lunch* in a resturant [*restaurant*]
saw fountain	we *saw* the danceing fountin [dancing *fountain*]
buy things	to *buy* somethings [some *things*]
mother took	my *mother took* me my sister
take care	"*take care* of your sister" and *take care*
Give things	give me somethings [some things].
took sister	my mother *took* me my *sister*
saw dress	I *saw* a beautiful *dress* in a shop
lost sister	I *lost* my *sister*
sister know	my *sister* didn't *know* anything
told mother	I *told* my *mother*
asked guard	we *asked* a securty [security] gured [*guard*]
asked security	we *asked* a securty [*security*] gured [guard]
Text Five	
vacation had	The worst *vacation* I have ever *had* the worst *vacation* I have ever *had*
parent bought	My *parent bought* tickets to Chine [China] as a surprise to us
bought tickets	My parent *bought tickets* to Chine [China] as a surprise to us
day came	The big *day* cam [*came*]
forgot tickets	we *forgot* our *tickets* at home
miss plane	we will *miss* the *plane*

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
brother came	my big *brother came* to the Airport after 10 minutes with the ticket
lost day	We *lost* our first *day*
bus came	the *bus* of our tour *came*
garden have	big *garden* have a small space of fruit garden
took garden	and *took* us big *garden*
house keep	a *house* that *keep* them warm
have space	big garden have a small *space* of fruit garden
tasting fruit	we *tasting* the *fruit*
saw snake	He *saw* a big *snake*
mum started	my *mom [mum] started* crying in her child
tourist guide	the *tourist guide* us
took photos	we *took* beautiful photor [*photos*]
dad tried	My *dad* tryed [*tried*]
call taxi	to *call* a *taxi*
pain came	the *pain came* strong
sister started	my *sister* starte [*started*] feeling sick then my brother
told father	they *told* my *father*
took shots	we *took* the *shots*
Text Six	
vacation had	the worst *vacation* you have ever *had*
forget bags	we forget our *bags* in airport
get bags	to *get* our *bags*
found people	we *found* a lot of *people*
car work	the *car* didn’t *work*
brother get	my *brother* get flu
get flu	my brother get *flu*
father called	My *father called* a doctor
called doctor	My father *called* a *doctor*
father told	my *father told* us sorry about the worst vacation
ask people	*ask* another *people* about it
Text Seven	
vacation had	My worst *vacation* that I never *had*
checked maps	We *checked* googel [google] *maps*
got information	we *got* some *information* about that
spent day	we *spent* our first *days* in a hotel We spen [*spent*] our therd [third] *day* in a road
move hotel	to *move* this *hotel*
problem happened	a big *problem happened*
asked service	he *asked* the room serves [*service*]
clean rooms	to *clean* our *rooms*
find hotel	to *find* some good *hotel*
family started	my *family started* walking
find place	to *find* some *place*
found family	I *found* my *family*
have dinner	to *have* a *dinner* in resturant [restaurant]

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
restaurant called	resturant [*restaurant*] that called ‘Semit’
quick order	we *quick* our *order*
waiter gave	the witer [*waiter*] *gave* us our food
gave food	the witer [waiter] *gave* us our *food*
Text Eight	
saw buildings	We *saw* the high billdings [*buildings*] and the modern restaurants and a lot of candys [candies] and gifts shops
took pictures	We *took pictures*
have fun	and also have *fun* together
saw windows	we *saw* our *windows* open!
found clothes	we didn’t found our *clothes* and bags!
found bags	we didn’t found our clothes and *bags*!
told story	and *told* him the whole *story*
sleep night	We didn’t *sleep* that *night*!
policeman told	some *police man [policeman] told* as [us]
had vacation	We *had* a worst *vacation* in the end

Table Two

Candidate Adjective-Noun Combinations	**Citation from the Text**
Text One	
something good	Life has a lot of experience *something good* and something bad
something bad	Life has a lot of experience something good and *something bad*
experience good	Whe [when] the experines [*experience*] was *good*
experience bad	If the *experience* was *bad*
big sound	My accident made big sound
sick person	when I visit *sick person*
Text Two	
horrible day	It was a *horrible day*
final exams	I was doing my *final exams*
next trip	talking about our *next trip* in summer
worst vacation	it was the *worst vacation*
good student	I become very *good student*
Text Three	
food delicious	the *food* wasn’t *delicious*
restaurant small	The *restaurant* was *small*
beautiful hotel	we stayed on a *beautiful hotel*
worst vacation	The *worst vacation* I have ever had This was my *worst vacation*
big day	Its the *big day*
little sister	my *little sister* lost
sister lost	my little *sister lost*

Candidate Adjective-Noun Combinations	**Citation from the Text**
bad things	a lot of *bad things* happened to us
Text Four	
famous hotel	we sat in *famous hotel*
nice place	it was clean and *nice place*
clean place	It was *clean* and nice *place*
beautiful dress	I saw a *beautiful dress* in a shop
Text Five	
worst vacation	The *worst vacation* I have ever had It was the *worst vacation*
big day	The *big day* cam [came]
big brother	my *big brother* came to the Airport
long planeting	We arrived to Chine [China] after *long planeting*
deep sleep	we just falled in *deep sleep*.
big garden	and took us *big garden*
small space	big garden have a *small space* of fruit garden
little brother	that was my *litele* [*little*] *brother*
big snake	He saw a *big snake*
beautiful photos	we took *beautiful photor* [*photos*]
pain strong	the *pain* came *strong*
sister sick	my *sister* starte [started] feeling *sick* then my brother
sick brother	my sister starte [started] feeling *sick* then my *brother*
food poisonous	that was a *food poiesnios* [*poisonous*]
Text Six	
worst vacation	it was the *worst vacation* the *worst vacation* you have ever had
weather good	the *weather* wasn't *good*
Text Seven	
worst vacation	My *worst vacation* that I never had
long vacation	to go in this *long vacation* to Istanbul- Turkey
bored time	When we arraived [arrived] in Istanbul after, along [a long], *bored time*
long time	When we arraived [arrived] in Istanbul after, along [a *long*], bored *time*
tired brother	I was very tierd [*tired*] and my *brother* also
big problem	a *big problem* happened
good hotel	to find some *good hotel*
dark colours	the sky came with ~~a~~ *dark colors [colours]*
huge road	I lost my silf [myself] in a *huge road*
meal cold	my dady [daddy] *meal* was very *cold*
Text Eight	
beautiful area	we went to Times square. The *beautiful area* in new york city
high buildings	We saw the *high billdings* [*buildings*] and the modern restaurants and a lot of candys [candies] and gifts shops
modern restaurants	We saw the high billdings [buildings] and the *modern restaurants* and a lot of candys [candies] and gifts shops
weather fantastic	The *weather* was *fantastic*!

Candidate Adjective-Noun Combinations	**Citation from the Text**
people nice	*people* was very *nice*
worst vacation	We had a *worst vacation* in the end
worst thing	it's the Worst *thing* in The Vacation !
windows open	we saw our *windows open*!
whole story	and told him the *whole story*

Table Three

Candidate Noun-Noun Combinations	**Citation from the Text**
Text One	
None	
Text Two	
living room	My family were sitting in the *living room*
Text Three	
None	
Text Four	
living room	I sat with my family in *living room*
security guard	we asked a securty [*security*] gured [*guard*]
dancing fountain	we saw the danceing fountin [*dancing fountain*]
Text Five	
fruit garden	big garden have a small space of *fruit garden*
poisoning snake	that isn't a *poiesing* [*poising*] *snake*
Text Six	
None	
Text Seven	
daddy meal	my *dady [daddy] meal* was very cold
room service	he asked the *room* serves [*service*]
Text Eight	
sisters brother	In 2014 my *sisters* and *brother* and i [I] we decaded [decided]
candies gifts	We saw the high billdings [buildings] and the modern restaurants and a lot of *candys* [*candies*] and *gifts* shops
gifts shops	We saw the high billdings [buildings] and the modern restaurants and a lot of candys [candies] and *gifts shops*
candies shops	We saw the high billdings [buildings] and the modern restaurants and a lot of candys [*candies*] and gifts *shops*
buildings restaurants	We saw the high billdings [*buildings*] and the modern *restaurants* and a lot of candys [candies] and gifts shops
breakfast area	We went to the *breakfast area*
clothes bags	we didn't found our *clothes* and *bags*!
police station	we went to *police* staistion [*station*]
shingle hotel	We checked in *shingle hotel*

Appendix 8: Candidate VAN combinations extracted from intermediate level written texts organized according to their types, and included in contexts they appeared in

Table One

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
Text One	
improve life	and *improve* their *life*
people have	people can have a pure and fresh air people can have a very calme [calm] and quite [quiet] live [life] with city nois [noise]
Have air	people can *have* a pure and fresh *air*
thing make	any other *thing* can *make* the air polluted
make air	any other thing can *make* the *air* polluted
thing bother	any other *thing* can *bother* them
recommend people	I *recommend people*
has advantages	it *has* a lot of advanteges [*advantages*]
forget town	they should not *forget* their home *town*
forget home	they should not *forget* their *home* town
have disadvantages	*Have* a lot of *disadvantages* and a lot of advantages
Text Two	
see family	to *see* your *family* maybe for weeks
people prefer	lots of *people* prefare [*prefer*] living in a big city
visit village	to *visit* a small *village*
people hate	*people hate* living in villages
find things	they can't *find things*
have fun	have fun with
find parks	to *find parks* or playgrounds
find playgrounds	to *find* parks or *playgrounds*
find clinics	You can only *find* small cleniqes [*clinics*] with a few doctors
Text Three	
find home	The actuall [actual] place where you can *find home*
affect personality	to *affect* you and your *personality* such as village
consider thing	Some might consedr [*consider*] it an old *thing* to it
villages force	Some *villages* might *force* you
drive way	to *drive* a long *way*
make call	to *make* a *call*.
need job	you will *need* a *job*
afford life	to *afford* this cheap *life* there
bring action	to *bring* you some *action*
find peace	you will *find peace*
have time	you'll *have* a lot of *time* with yourself you will *have time*

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
enjoy weather	while you are enjoing [*enjoying*] the windy *weather*
cure soul	and *cure* your *soul* with the green areas all around you
choose place	to *choose* a *place*
like rush	wether [whether] you *like* the *rush* and noisy life or the peacful [peaceful] and clear one
like life	wether [whether] you *like* the *rush* and noisy life or the peacful [peaceful] and clear one
Text Four	
have advantages	Living in village have their *advantages* and disadvantages living in village have *advantages* and disadvantages it have a lot of *advantages*
have disadvantages	Living in village have their advantages and *disadvantages* living in village have advantages and *disadvantages*
have schools	that there they havn't [*haven't*] *schools* or hospital
have hospitals	that there they havn't [*haven't*] schools or *hospital*
village have	some *villages* a lot *have*
love place	you *love* quite [quiet] please [*place*] anyway
Text Five	
gives ability	Living in a village *gives* you the *ability*
live life	to *live* a hithier [healthier] *life*
smell smoke	you can not [can't] *smell* the *smoke* of the factories, big restaurants and cars
make life	living in a village will *make* your *life* easier, more simple and more comfortable
lead people	that will *lead* the *people*
visit cities	to *visit* the big *cities* from time to another
Text Six	
has advantage	Living in a village *has* many *advantages*
need car	you won't *need* a *car*
have time	and *have* great *time*
means end	For some people living in a village *mean* the *end* of there [their] life
love city	they *love* the *city*
have malls	they will *have* a lot~~s~~ of *malls*, hospitals and big school, university
things make	this *things make* the city
make city	this things *make* the *city*
city contains	the *city* contun [*contains*] lots of pollitoned [pollution]
contains pollution	the city contun [*contains*] lots of pollitoned [*pollution*]
finish education	I *finish* my educiton [*education*] in the city
love farms	I *love* the green *farms* and the old houses
take care	I *like* teka [take] *car[e]* of the animals and farms
Text Seven	
village have	*Villages* haven't entertainment
Have/ has entertainment	Villages haven't *entertainment*

Candidate Verb-Noun/ Noun-Verb Combinations	**Citation from the Text**
	village *hasn't* many *entertainment* and emenities [amenities]
has amenities	village *hasn't* many entertainment and emenities [*amenities*]
people think	Some *people think* that
people see	Some *people* will *see* a lot of advantages
see advantages	Some people will *see* a lot of *advantages*
waste time	that can't *waste* your *time*
people have	the *people* can't *have* many options for education like, universities, global schools and educational institutes
have options	the people can't *have* many *options* for education like, universities, global schools and educational institutes
village has	a *village* hasn't good health centers [centres] *village* hasn't many entertainment and emenities [amenities]
has centres	a village *hasn't* good health centers *[centres]*
people get	the *people* in a village can't *get* the best job with good and high salaries
get job	the people in a village can't *get* the best *job* with good and high salaries
Text Eight	
live life	to *live* a simple village *life*
villages got	*villages got* alot [a lot] of trees
got trees	villages *got* alot [a lot] of *trees*
own car	you should *own* your own *car* or bike
own bike	you should *own* your own car or *bike*

Table Two

Candidate Adjective-Noun Combinations	**Citation from the Text**
Text One	
small villages	So many people still live in *small villages* there are so many disadvantages of living in a *small village* There are a few advanteges [advantages] of living in *small villages* So many people still living in *small villages* for some reson [reason]
people able	*people* will not be *able*
Open mind	to be more cefictecaded [sophisticated] or *open mind*
important facilities	there is no *important* fucllitys [*facilities*] such as hospitals with good doctors
good doctors	there is no important fucllitys [facilities] such as hospitals with *good doctors*

Candidate Adjective-Noun Combinations	**Citation from the Text**
pure air	people can have a *pure* and fresh *air*
fresh air	people can have a pure and *fresh air*
air polluted	any other thing can make the *air polluted*
calm life	people can have a very calme [*calm*] and quite [quiet] live [*life*] with city nois [noise]
quiet life	people can have a very calme [calm] and quite [*quiet*] live [*life*] with city nois [noise]
big city	to live in a *big city*
Text Two	
new name	every day you hear about a *new* village’s *name*
big city	Like if you were living in a *big city* lots of people prefare [prefer] living in a *big city*
long hours	you can forget about *long* houres [*hours*] driving
small village	to visit a *small village*
different village	that might be in a diffrint [*different*] *village*
big hospitals	there are no *big hospitals* with excellent care
excellent care	there are no big hospitals with *excellent care*
small clinics	You can only find *small* cleniqes [*clinics*] with a few doctors
people scared	*people* oftenly *scared* of what things they have never tried
Text Three	
actual place	The actuall [*actual*] *place* where you can find home
powerful thing	home is the most *powerful thing*
old thing	Some might consedr [consider] it an *old thing* to it
great thing	Living in a village can be a *great thing*
long way	to drive a *long way*
cheap life	to afford this *cheap life* there
noisy voices	you will get rid of all the *noisy voices* of the crowd city and the cars
beautiful thing	You will think about every *beautiful thing* in your life
windy weather	you are enjoing [enjoying] the *windy weather*
green areas	and cure your soul with the *green areas* all around you
same level	the advantages and disadvantages to living in a village might be on the *same level*
noisy life	wether [whether] you like the rush and *noisy life* or the peacful [peaceful] and clear one
peaceful life	wether [whether] you like the rush and noisy *life* or the peacful [*peaceful*] and clear one.
clear life	wether [whether] you like the rush and noisy *life* or the peacful [peaceful] and *clear* one.
cheap job	a *job* which is *cheap*
Text Four	
good idea	Living in village it’s a *good idea* for healty [healthy] life Living in viliage [village] a *good idea* if you can live with the disadvantages
healthy life	Living in village it’s a good idea for healty [*healthy*] *life*

Candidate Adjective-Noun Combinations	**Citation from the Text**
fresh air	as you know ~~a~~ *fresh air* and a natural place with ~~a~~ good and genrase [generous] people the fesh [*fresh*] *air* Because of forest and farmes [farms] there is no nois [noise] at all
natural air	you know fresh *air* ~~a~~ *natural*
house small	the *house* is sumall [*small*]
generous people	as you know ~~a~~ fresh air and a natural place with ~~a~~ good and genrase [*generous*] *people* and the *people* are genruse [generous] and kind
good people	as you know ~~a~~ fresh air and a natural place with ~~a~~ *good* and genrase [generous] *people*
natural place	as you know ~~a~~ fresh air and a *natural place* with ~~a~~ good and genrase [generous] people
school complete	The hospitals and *school* its not *complete* like the citey [city]
hospitals complete	The *hospitals* and school its not *complete* like the citey [city]
peaceful place	it's a peace [*peaceful*] *place*
fresh food	the *food* there *fresh* and more weather than the citey [city] Because, it's came from the farm to me
kind people	the *people* are genruse [generous] and *kind*
quiet place	you love quite [*quiet*] please [*place*] anyway
Text Five	
big city	Living in a *big city* is not similer [similar] as living in a small village to visit the *big cities* from time to another
small village	Living in a big city is not similer [similar] as living in a *small village*
good side	Each has its *good* and bad *sides*
bad side	Each has its good and *bad sides*
healthier life	to live a hithier [*healthier*] *life*
big restaurant	you can not [can't] smell the smoke of the factories, *big restaurants* and cars
life easy	living in a village will make your *life easier,* more simple and more comfortable
life simple	living in a village will make your *life* easier, more *simple* and more comfortable
hard thing	living there can be a *hard thing*
big markets	living in a village might be exhausting sometimes because of not having *big markets,* mall, book stores [bookstores], ect [etc.]
big mall	living in a village might be exhausting sometimes because of not having *big* markets, *mall,* book stores [bookstores], ect [etc.]
big bookstores	living in a village might be exhausting sometimes because of not having *big* markets, mall, *book stores [bookstores]*, ect [etc.]
bad thing	Living in a village is not a *bad thing* at all

Candidate Adjective-Noun Combinations	**Citation from the Text**
Text Six	
quiet place	Some like to be in *quite* [quiet] plaes [*places*]
good mood	you will be at *good mood* all the time
great time	and have *great time*
big schools	they will have ~~a~~ lots of malls, hospitals and *big school*, university
crowded place	The city so croded [*crowded*] pleas [*place*]
good place	that not *good* pleas [*place*]
different countries	There are millons [millions] of people from diffrent [*different*] countres [*countries*]
great thing	Living in a village is a *great thing*
green farms	I love the *green farms* and the old houses
old houses	I love the green farms and the *old houses*
Text Seven	
lifestyle expensive	the life style [*lifestyle*] won't be very *expensive* like, renting a home and shopping
villages quiet	*villages* are *quiet*
global schools	the people can't have many options for education like, universities, *global schools* and educational institutes
educational institutes	the people can't have many options for education like, universities, global schools and *educational institutes*.
good centres	a village hasn't *good* health *centers [centres]*
hospitals well-equipped	the *hospitals* aren't *well-equipped*
good clinics	there aren't *good clinics*
best job	the people in a village can't get the *best job* with good and high salaries
high salaries	the people in a village can't get the best job with good and *high salaries*
good salaries	the people in a village can't get the best job with *good* and high *salaries*
big cities	living in *big cities* is better than living in a village
good thing	Living in a village to some people is a *good thing*
Text Eight	
right place	Choosing the *right place* to live in is very important
simple life	to live a *simple* village *life*
great place	its [it’s] a *great place* for the elderly
clean environment	it's such a *clean environment* to live in
big companies	*big companies* are usually in the city
transportation well-developed	the public *transportation* isn't *well-developed*
good places	both villages and also cities are *good places* to live in
public transportation	the *public transportation* isn't well-developed

Table Three

Candidate Noun-Noun Combinations	**Citation from the Text**
Text One	
traditions customs	they still stick with their tardtions [*traditions*] and *customs*
schools universities	there is no *schools* or univircities [*universities*]
city noise	people can have a very calme [calm] and quite [quiet] live [life] with *city* nois [*noise*]
home town	they should not forget their *home town*
disadvantages advantages	Have a lot of *disadvantages* and a lot of *advantages*
Text Two	
village name	every day you hear about a new *village's name*
family neighbours	living in a village gets you closer to your *family* and *neighbours*
roads traffic	there are a few stores with a few *roads* and no ~~a~~ trafic [*traffic*]
parks playgrounds	to find *parks* or *playgrounds*
Text Three	
home place	*Home* is a *place*
home thing	*home* is the most powerful *thing*
city cars	you will get rid of all the noisy voices of the crowd *city* and the *cars*
advantages disadvantages	the *advantages* and *disadvantages* to living in a village might be on the same level
rush life	wether [whether] you like the *rush* and noisy *life* or the peacful [peaceful] and clear one
phone station	to get to the *phone station*
shining stars	to lay down watching the sky fell [full] with the *shining stars*
Crowd city	you will get rid of all the noisy voices of the *crowd city* and the cars
Text Four	
advantages disadvantages	Living in village have their *advantages* and *disadvantages* living in village have *advantages* and *disadvantages*
schools hospitals	there they havn't [haven't] *schools* or *hospital* The *hospitals* and *school* its not complete like the citey [city]
wolves bears	the woolfs [*wolves*] ands [and] *bears* living in village that isn't save Because of animals like the woolfs [*wolves*] and *bears*
air place	as you know a fresh *air* and a natural *place* with a good and genrase [generous] people
houses wood	the *houses* made of *wood*
forest farms	the fesh [fresh] air Because of *forest* and farmes [*farms*] there is no nois [noise] at all
birds animals	there is no nois [noise] at all just the *birds* and *animals*
birds noise	there is no nois [noise] at all just the *birds* and *animals*

Candidate Noun-Noun Combinations	**Citation from the Text**
Text Five	
town location	The home *town location* of the human being effects on his body, brian [brain] and heart
home location	The *home* town *location* of the human being effects on his body, brian [brain] and heart
home town	The *home town location* of the human being effects on his body, brian [brain] and heart
Human being	The home town location of the *human being* effects on his body, brian [brain] and heart
body brain	The home town location of the human being effects on his *body*, brian [*brain*] and heart
brain heart	The home town location of the human being effects on his body, brian [*brain*] and *heart*
body heart	The home town location of the human being effects on his *body*, brian [brain] and *heart*
people cars	to listen to the noice [noise] of the *people*, *cars*, buses and trains
people buses	to listen to the noice [noise] of the *people*, cars, *buses* and trains
cars buses	to listen to the noice [noise] of the people, *cars*, *buses* and trains
cars trains	to listen to the noice [noise] of the people, *cars*, buses and *trains*
buses trains	to listen to the noice [noise] of the people, cars, *buses* and *trains*
factories restaurant	you can not [can't] smell the smoke of the *factories*, big *restaurants* and cars
factories cars	you can not [can't] smell the smoke of the *factories*, big restaurants and *cars*
restaurants cars	you can not [can't] smell the smoke of the factories, big *restaurants* and *cars*
entertaining place	there are not many intertaining [*entertaining*] *places*
coffee shops	to go like *coffee shops*, malls, cenimas [cinemas] and barks [parks]
shops malls	to go like coffee *shops*, *malls*, cenimas [cinemas] and barks [parks]
shops cinemas	to go like coffee *shops*, malls, cenimas [*cinemas*] and barks [parks]
coffee shops	to go like *coffee shops*, malls, cenimas [*cinemas*] and barks [parks]
mall cinemas	to go like coffee shops, *malls*, cenimas [*cinemas*] and barks [parks]
mall parks	to go like coffee shops, *malls*, cenimas [cinemas] and barks [*parks*]
cinemas parks	to go like coffee shops, malls, cenimas [*cinemas*] and barks [*parks*]

Candidate Noun-Noun Combinations	**Citation from the Text**
markets mall	living in a village might be exhausting sometimes because of not having big *markets*, *malls*, book stores [bookstores], ect [etc.]
markets bookstores	living in a village might be exhausting sometimes because of not having big *markets*, mall, *book stores [bookstores]*, ect [etc.]
mall bookstores	living in a village might be exhausting sometimes because of not having big markets, *mall*, *book stores [bookstores]*, ect [etc.]
Text Six	
advantages disadvantages	There are a lots of *advantages* and *disadvantages* about living in a village
malls hospitals	they will have ~~a~~ lots of *malls*, *hospitals* and big school, university
malls schools	they will have ~~a~~ lots of *malls*, hospitals and big *school*, university
hospitals schools	they will have ~~a~~ lots of malls, *hospitals* and big *school*, university
schools university	they will have ~~a~~ lots of malls, hospitals and big *school*, *university*
Hospital university	they will have ~~a~~ lots of malls, hospitals and big *school*, *university*
city place	The *city* so croded [crowded] pleas [*place*]
car station	there is many *cars* gas *staican* [*station*]
gas station	there is many cars *gas staican* [*station*]
farms houses	I love the green *farms* and the old *houses*
animals farms	I like teka [take] car[e] of the *animals* and *farms*
Text Seven	
culture health	to living in a village related to *culture*, *health* and education
culture education	to living in a village related to *culture*, health and *education*
health education	to living in a village related to culture, *health* and *education*
people cars	there aren't a lot of *people* and *cars* on the streets
universities schools	the people can't have many options for education like, *universities*, global *schools* and educational institutes
schools institutes	the people can't have many options for education like, universities, global *schools* and educational *institutes*
entertainment amenities	village hasn't many *entertainment* and emenities [*amenities*]
Health centres	a village hasn't good *health centers [centres]*
Text Eight	
car bike	you should own your own *car* or *bike*
villages cities	both *villages* and also *cities* are good places to live in
cities places	both *villages* and also *cities* are good places to live in

Appendix 9: Identified collocations from pre-intermediate level written texts organized according to their types with their LogDice score and level of fixedness

Collocation	Type of Collocation	LogDice Score	Level of the collocation fixedness
has experience	Verb-noun	6.5	Medium
visited uncle	Verb-noun	4.8	Medium
left bed	Verb-noun	5.8	Medium
played football	Verb-noun	8.7	High
made sound	Verb-noun	6.7	Medium
know problem	Verb-noun	6.2	Medium
saw son	Verb-noun	5.1	Medium
saw face	Verb-noun	7.6	High
learned lesson	Verb-noun	9.6	High
visit person	Verb-noun	5.0	Medium
doing exams	Verb-noun	3.4	Low
found people	Verb-noun	7.8	High
heard knock	Verb-noun	5.2	Medium
tell family	Verb-noun	5.9	Medium
tell parents	Verb-noun	6.5	Medium
visited friend	Verb-noun	7.6	High
buy coffee	Verb-noun	5.8	Medium
told dad	Verb-noun	5.8	Medium
spent vacation	Verb-noun	4.0	Medium
become student	Verb-noun	6.2	Medium
had lunch	Verb-noun	4.5	Medium
get rest	Verb-noun	5.4	Medium
called police	Verb-noun	7.8	High
stayed day	Verb-noun	6.7	Medium
called doctor	Verb-noun	6.5	Medium
took father	Verb-noun	5.9	Medium
find taxi	Verb-noun	3.2	Low
finished exams	Verb-noun	4.6	Medium
ate lunch	Verb-noun	7.9	High
saw fountain	Verb-noun	0.9	Low
buy things	Verb-noun	6.9	Medium
take care	Verb-noun	8.7	High
Give things	Verb-noun	6.5	Medium
took sister	Verb-noun	4.6	Medium
saw dress	Verb-noun	3.5	Low
lost sister	Verb-noun	4.7	Medium
told mother	Verb-noun	7.6	High
Asked security	Verb-noun	4.1	Medium
bought tickets	Verb-noun	8.3	High
miss plane	Verb-noun	5.1	Medium

lost day	Verb-noun	6.5	Medium
took garden	Verb-noun	4.8	Medium
have space	Verb-noun	4.9	Medium
tasting fruit	Verb-noun	6.6	Medium
saw snake	Verb-noun	1.8	Low
took photos	Verb-noun	5.0	Medium
call taxi	Verb-noun	4.5	Medium
told father	Verb-noun	7.2	High
get bags	Verb-noun	5.7	Medium
found people	Verb-noun	7.8	High
get flu	Verb-noun	2.8	Low
called doctor	Verb-noun	6.5	Medium
ask people	Verb-noun	7.5	High
checked maps	Verb-noun	5.4	Medium
got information	Verb-noun	6.5	Medium
spent day	Verb-noun	8.8	High
move hotel	Verb-noun	4.6	Medium
asked service	Verb-noun	5.4	Medium
clean rooms	Verb-noun	6.0	Medium
find hotel	Verb-noun	5.0	Medium
find place	Verb-noun	7.6	High
found family	Verb-noun	6.2	Medium
have dinner	Verb-noun	4.8	Medium
gave food	Verb-noun	6.1	Medium
saw buildings	Verb-noun	5.2	Medium
took pictures	Verb-noun	6.7	Medium
have fun	Verb-noun	4.6	Medium
saw windows	Verb-noun	6.6	Medium
found clothes	Verb-noun	5.0	Medium
found bags	Verb-noun	4.8	Medium
told story	Verb-noun	9.3	High
sleep night	Verb-noun	7.8	High
life has	noun-verb	7.5	High
doctors told	noun-verb	6.8	Medium
accident made	noun-verb	3.6	Medium
son looking	noun-verb	5.2	Medium
Uncle came	noun-verb	3.0	Low
Uncle left	noun-verb	2.7	Low
mother asked	noun-verb	6.7	Medium
trip leave	noun-verb	3.9	Medium
father said	noun-verb	6.6	Medium
father came	noun-verb	6.2	Medium
things happened	noun-verb	9.0	High
mother want	noun-verb	6.0	Medium
mother took	noun-verb	6.2	Medium
sister know	noun-verb	4.8	Medium
parent bought	noun-verb	5.4	Medium
day came	noun-verb	8.1	High

brother came	noun-verb	5.2	Medium
bus came	noun-verb	4.9	Medium
garden have	noun-verb	4.8	Medium
house keep	noun-verb	7.1	High
mum started	noun-verb	3.9	Medium
dad tried	noun-verb	3.5	Low
pain came	noun-verb	4.5	Medium
sister started	noun-verb	3.7	Medium
car work	noun-verb	5.6	Medium
brother get	noun-verb	4.8	Medium
father called	noun-verb	6.1	Medium
father told	noun-verb	7.2	High
problem happened	noun-verb	4.7	Medium
family started	noun-verb	6.1	Medium
restaurant called	noun-verb	3.6	Medium
policeman told	noun-verb	4.7	Medium
something good	Adjective-noun	6.4	Medium
something bad	Adjective-noun	5.7	Medium
experience good	Adjective-noun	5.8	Medium
experience bad	Adjective-noun	6.4	Medium
big sound	Adjective-noun	4.9	Medium
sick person	Adjective-noun	5.6	Medium
horrible day	Adjective-noun	2.3	Low
final exams	Adjective-noun	5.6	Medium
next trip	Adjective-noun	5.5	Medium
good student	Adjective-noun	5.5	Medium
food delicious	Adjective-noun	6.4	Medium
restaurant small	Adjective-noun	5.7	Medium
beautiful hotel	Adjective-noun	5.9	Medium
big day	Adjective-noun	6.6	Medium
little sister	Adjective-noun	5.9	Medium
bad things	Adjective-noun	8.4	High
famous hotel	Adjective-noun	5.4	Medium
nice place	Adjective-noun	6.0	Medium
clean place	Adjective-noun	4.8	Medium
beautiful dress	Adjective-noun	6.9	Medium
big day	Adjective-noun	6.6	Medium
big brother	Adjective-noun	6.6	Medium
deep sleep	Adjective-noun	8.4	High
big garden	Adjective-noun	6.1	Medium
small space	Adjective-noun	6.6	Medium
little brother	Adjective-noun	6.2	Medium
big snake	Adjective-noun	3.6	Medium
pain strong	Adjective-noun	3.9	Medium
sick brother	Adjective-noun	4.1	Medium
weather good	Adjective-noun	6.1	Medium
long vacation	Adjective-noun	4.8	Medium
long time	Adjective-noun	9.5	High

big problem	Adjective-noun	7.4	High
good hotel	Adjective-noun	5.6	Medium
dark colours	Adjective-noun	8.2	High
meal cold	Adjective-noun	6.3	Medium
beautiful area	Adjective-noun	4.6	Medium
high buildings	Adjective-noun	6.0	Medium
modern restaurants	Adjective-noun	4.7	Medium
people nice	Adjective-noun	6.1	Medium
worst thing	Adjective-noun	8.4	High
windows open	Adjective-noun	8.8	High
whole story	Adjective-noun	7.9	High
living room	Noun-noun	7.6	High
security guard	Noun-noun	8.8	High
living room	Noun-noun	7.6	High
fruit garden	Noun-noun	6.3	Medium
room service	Noun-noun	6.1	Medium
sisters brother	Noun-noun	10.5	High
gifts shops	Noun-noun	7.5	High
breakfast area	Noun-noun	2.9	Low
police station	Noun-noun	9.7	High
Clothes bags	Noun-noun	6.5	Medium

Appendix 10: Identified collocations extracted from intermediate level written texts organized according to their types with their LogDice score and level of fixedness

Collocation	**Type of Collocation**	**LogDice Score**	**Level of the collocation fixedness**
improve life	Verb-noun	6.8	Medium
make air	Verb-noun	4.9	Medium
Have air	Verb-noun	4.9	Medium
recommend people	Verb-noun	3.5	Low
has advantages	Verb-noun	5.8	Medium
Forget home	Verb-noun	4.9	Medium
have disadvantages	Verb-noun	3.0	Low
see family	Verb-noun	6.0	Medium
visit village	Verb-noun	6.7	Medium
find things	Verb-noun	7.2	High
have fun	Verb-noun	4.6	Medium
find parks	Verb-noun	3.5	Low
find clinics	Verb-noun	2.6	Low
find home	Verb-noun	7.4	High
affect personality	Verb-noun	4.8	Medium
consider thing	Verb-noun	5.9	Medium
drive way	Verb-noun	5.5	Medium
make call	Verb-noun	6.8	Medium
need job	Verb-noun	6.6	Medium
afford life	Verb-noun	3.1	Low
bring action	Verb-noun	7.8	High
find peace	Verb-noun	4.8	Medium
have time	Verb-noun	9.1	High
enjoy weather	Verb-noun	5.9	Medium
choose place	Verb-noun	6.0	Medium
like life	Verb-noun	5.7	Medium
have advantages	Verb-noun	5.8	Medium
have disadvantages	Verb-noun	3.0	Low
have schools	Verb-noun	6.9	Medium
have hospitals	Verb-noun	5.4	Medium
love place	Verb-noun	5.3	Medium
gives ability	Verb-noun	5.1	Medium
live life	Verb-noun	8.8	High
smell smoke	Verb-noun	8.2	High
make life	Verb-noun	7.5	High
lead people	Verb-noun	6.2	Medium
visit cities	Verb-noun	6.7	Medium
has advantage	Verb-noun	5.8	Medium
need car	Verb-noun	5.7	Medium
have time	Verb-noun	9.1	High

means end	Verb-noun	6.4	Medium
love city	Verb-noun	4.7	Medium
make city	Verb-noun	4.9	Medium
finish education	Verb-noun	3.8	Medium
take care	Verb-noun	8.7	High
Have/ has entertainment	Verb-noun	1.7	Low
has amenities	Verb-noun	0.7	Low
see advantages	Verb-noun	5.1	Medium
waste time	Verb-noun	7.6	High
have options	Verb-noun	4.8	Medium
has centres	Verb-noun	5.2	Medium
get job	Verb-noun	8.2	High
live life	Verb-noun	8.8	High
got trees	Verb-noun	4.5	Medium
own car	Verb-noun	6.0	Medium
own bike	Verb-noun	4.5	Medium
people have	Noun-verb	8.5	High
thing make	Noun-verb	7.8	High
thing bother	Noun-verb	4.9	Medium
people prefer	Noun-verb	5.8	Medium
people hate	Noun-verb	5.0	Medium
village have	Noun-verb	5.0	Medium
things make	Noun-verb	7.8	High
city contains	Noun-verb	4.8	Medium
village have	Noun-verb	5.0	Medium
people think	Noun-verb	8.5	High
people see	Noun-verb	8.1	High
people have	Noun-verb	8.5	High
village has	Noun-verb	5.0	Medium
people get	Noun-verb	8.5	High
villages got	Noun-verb	3.9	Medium
small villages	Adjective-noun	7.7	High
people able	Adjective-noun	6.8	Medium
Open mind	Adjective-noun	7.2	High
important facilities	Adjective-noun	4.4	Medium
good doctors	Adjective-noun	5.4	Medium
pure air	Adjective-noun	5.7	Medium
fresh air	Adjective-noun	9.6	High
air polluted	Adjective-noun	6.0	Medium
calm life	Adjective-noun	2.1	Low
quiet life	Adjective-noun	6.2	Medium
big city	Adjective-noun	7.6	High
new name	Adjective-noun	6.7	Medium
big city	Adjective-noun	7.6	High
long hours	Adjective-noun	7.9	High
small village	Adjective-noun	7.7	High
different village	Adjective-noun	4.3	Medium

big hospitals	Adjective-noun	5.2	Medium
excellent care	Adjective-noun	4.4	Medium
actual place	Adjective-noun	4.9	Medium
powerful thing	Adjective-noun	2.6	Low
old thing	Adjective-noun	6.6	Medium
great thing	Adjective-noun	7.0	Medium
long way	Adjective-noun	8.7	High
cheap life	Adjective-noun	3.5	Low
beautiful thing	Adjective-noun	5.9	Medium
windy weather	Adjective-noun	7.4	High
green areas	Adjective-noun	4.8	Medium
same level	Adjective-noun	7.2	High
peaceful life	Adjective-noun	4.5	Medium
cheap job	Adjective-noun	4.4	Medium
good idea	Adjective-noun	8.9	High
healthy life	Adjective-noun	5.4	Medium
fresh air	Adjective-noun	9.6	High
natural air	Adjective-noun	4.7	Medium
house small	Adjective-noun	7.5	High
generous people	Adjective-noun	3.5	Low
good people	Adjective-noun	7.7	High
natural place	Adjective-noun	5.2	Medium
peaceful place	Adjective-noun	3.8	Medium
fresh food	Adjective-noun	7.5	High
kind people	Adjective-noun	3.2	Low
quiet place	Adjective-noun	5.8	Medium
big city	Adjective-noun	7.6	High
small village	Adjective-noun	7.7	High
good side	Adjective-noun	6.2	Medium
bad side	Adjective-noun	5.3	Medium
healthier life	Adjective-noun	5.4	Medium
life easy	Adjective-noun	7.2	High
life simple	Adjective-noun	6.0	Medium
hard thing	Adjective-noun	6.0	Medium
big markets	Adjective-noun	6.7	Medium
bad thing	Adjective-noun	8.4	High
quiet place	Adjective-noun	5.8	Medium
good mood	Adjective-noun	4.9	Medium
great time	Adjective-noun	7.3	High
big schools	Adjective-noun	5.4	Medium
crowded place	Adjective-noun	3.4	Low
good place	Adjective-noun	7.3	High
different countries	Adjective-noun	7.8	High
great thing	Adjective-noun	7.0	Medium
old houses	Adjective-noun	8.0	High
lifestyle expensive	Adjective-noun	4.5	Medium
villages quiet	Adjective-noun	6.5	Medium
educational institutes	Adjective-noun	4.7	Medium
good centres	Adjective-noun	4.6	Medium

best job	Adjective-noun	8.2	High
high salaries	Adjective-noun	6.1	Medium
good salaries	Adjective-noun	3.5	Low
big cities	Adjective-noun	7.6	High
good thing	Adjective-noun	8.8	High
right place	Adjective-noun	7.4	High
simple life	Adjective-noun	6.0	Medium
great place	Adjective-noun	6.4	Medium
clean environment	Adjective-noun	6.5	Medium
big companies	Adjective-noun	7.5	High
good places	Adjective-noun	7.3	High
public transportation	Adjective-noun	3.1	Low
Big restaurant	Adjective-noun	3.0	Low
traditions customs	Noun-noun	7.3	High
schools universities	Noun-noun	7.4	High
city noise	Noun-noun	3.6	Medium
home town	Noun-noun	7.4	High
disadvantages advantages	Noun-noun	9.4	High
village name	Noun-noun	5.8	Medium
family neighbours	Noun-noun	5.8	Medium
roads traffic	Noun-noun	9.1	High
home place	Noun-noun	6.6	Medium
home thing	Noun-noun	5.7	Medium
city cars	Noun-noun	5.9	Medium
advantages disadvantages	Noun-noun	9.4	High
rush life	Noun-noun	1.6	Low
advantages disadvantages	Noun-noun	9.4	High
schools hospitals	Noun-noun	6.3	Medium
wolves bears	Noun-noun	7.4	High
air place	Noun-noun	5.2	Medium
houses wood	Noun-noun	5.3	Medium
birds animals	Noun-noun	8.5	High
birds noise	Noun-noun	4.6	Medium
town location	Noun-noun	5.3	Medium
Home location	Noun-noun	4.4	Medium
home town	Noun-noun	7.4	High
Human being	Noun-noun	6.5	Medium
body brain	Noun-noun	6.0	Medium
brain heart	Noun-noun	6.5	Medium
body heart	Noun-noun	5.5	Medium
people cars	Noun-noun	6.1	Medium
people buses	Noun-noun	4.5	Medium
cars buses	Noun-noun	6.9	Medium
cars trains	Noun-noun	6.5	Medium
buses trains	Noun-noun	8.4	High

factories cars	Noun-noun	5.7	Medium
restaurants cars	Noun-noun	4.7	Medium
coffee shops	Noun-noun	7.7	High
shops cinemas	Noun-noun	5.3	Medium
coffee shops	Noun-noun	7.7	High
advantages disadvantages	Noun-noun	9.4	High
hospitals schools	Noun-noun	6.3	Medium
schools university	Noun-noun	7.4	High
Hospital university	Noun-noun	5.5	Medium
city place	Noun-noun	6.1	Medium
car station	Noun-noun	5.9	Medium
gas station	Noun-noun	7.1	High
farms houses	Noun-noun	5.8	Medium
animals farms	Noun-noun	8.1	High
culture health	Noun-noun	4.2	Medium
culture education	Noun-noun	6.8	Medium
health education	Noun-noun	9.1	High
people cars	Noun-noun	6.1	Medium
universities schools	Noun-noun	7.4	High
schools institutes	Noun-noun	2.6	Low
entertainment amenities	Noun-noun	5.8	Medium
Health centres	Noun-noun	7.3	High
car bike	Noun-noun	5.7	Medium
villages cities	Noun-noun	5.9	Medium
cities places	Noun-noun	6.1	Medium

Appendix 11: List of the non-collocations (candidate VAN combinations extracted from pre-intermediate and intermediate level written texts, but were not identified in the corpus) organized according to their types

Non-collocation	**Type of the VAN combination**	**Level of the written text it appeared in**
arranged bag	Verb-noun	Pre-intermediate
asked guard	Verb-noun	Pre-intermediate
forgot tickets	Verb-noun	Pre-intermediate
took shots	Verb-noun	Pre-intermediate
forget bags	Verb-noun	Pre-intermediate
quick order	Verb-noun	Pre-intermediate
forgot passport	Verb-noun	Pre-intermediate
Had vacation	Verb-noun	Pre-intermediate
face crash	Noun-verb	Pre-intermediate
vacation had	Noun-verb	Pre-intermediate
father surprised	Noun-verb	Pre-intermediate
father screaming	Noun-verb	Pre-intermediate
tourist guide	Noun-verb	Pre-intermediate
waiter gave	Noun-verb	Pre-intermediate
worst vacation	Adjective-noun	Pre-intermediate
sister lost	Adjective-noun	Pre-intermediate
long planeting	Adjective-noun	Pre-intermediate
beautiful photos	Adjective-noun	Pre-intermediate
sister sick	Adjective-noun	Pre-intermediate
food poisonous	Adjective-noun	Pre-intermediate
bored time	Adjective-noun	Pre-intermediate
tired brother	Adjective-noun	Pre-intermediate
huge road	Adjective-noun	Pre-intermediate
weather fantastic	Adjective-noun	Pre-intermediate
dancing fountain	Noun-noun	Pre-intermediate
poisoning snake	Noun-noun	Pre-intermediate
daddy meal	Noun-noun	Pre-intermediate
candies gifts	Noun-noun	Pre-intermediate
candies shops	Noun-noun	Pre-intermediate
buildings restaurants	Noun-noun	Pre-intermediate
shingle hotel	Noun-noun	Pre-intermediate
forget town	Verb-noun	Intermediate
find playgrounds	Verb-noun	Intermediate
cure soul	Verb-noun	Intermediate
like rush	Verb-noun	Intermediate
have malls	Verb-noun	Intermediate
contains pollution	Verb-noun	Intermediate
love farms	Verb-noun	Intermediate

villages force	Noun-verb	Intermediate
small clinics	Adjective-noun	Intermediate
people scared	Adjective-noun	Intermediate
noisy voices	Adjective-noun	Intermediate
noisy life	Adjective-noun	Intermediate
clear life	Adjective-noun	Intermediate
school complete	Adjective-noun	Intermediate
hospitals complete	Adjective-noun	Intermediate
big mall	Adjective-noun	Intermediate
big bookstores	Adjective-noun	Intermediate
green farms	Adjective-noun	Intermediate
global schools	Adjective-noun	Intermediate
hospitals well-equipped	Adjective-noun	Intermediate
good clinics	Adjective-noun	Intermediate
transportation well-developed	Adjective-noun	Intermediate
parks playgrounds	Noun-noun	Intermediate
phone station	Noun-noun	Intermediate
shining stars	Noun-noun	Intermediate
forest farms	Noun-noun	Intermediate
factories restaurant	Noun-noun	Intermediate
shops malls	Noun-noun	Intermediate
mall cinemas	Noun-noun	Intermediate
mall parks	Noun-noun	Intermediate
cinemas parks	Noun-noun	Intermediate
markets mall	Noun-noun	Intermediate
markets bookstores	Noun-noun	Intermediate
mall bookstores	Noun-noun	Intermediate
entertaining place	Noun-noun	Intermediate
malls hospitals	Noun-noun	Intermediate
malls schools	Noun-noun	Intermediate
Crowd city	Noun-noun	Intermediate

Appendix 12: List of collocations agreed on as acceptable collocations by native-speaker informants from pre-intermediate and intermediate level written texts organized according to their types

Acceptable Collocation	Type of the collocation	Level of the written text it appeared in
asked guard	Verb-noun	Pre-intermediate
forgot tickets	Verb-noun	Pre-intermediate
forgot passport	Verb-noun	Pre-intermediate
father screaming	Noun-verb	Pre-intermediate
tourist guide	Noun-verb	Pre-intermediate
waiter gave	Noun-verb	Pre-intermediate
father surprised	Noun-verb	Pre-intermediate
worst vacation	Adjective-noun	Pre-intermediate
sister sick	Adjective-noun	Pre-intermediate
weather fantastic	Adjective-noun	Pre-intermediate
sister lost	Adjective-noun	Pre-intermediate
beautiful photos	Adjective-noun	Pre-intermediate
dancing fountain	Noun-noun	Pre-intermediate
candies shops	Noun-noun	Pre-intermediate
buildings restaurants	Noun-noun	Pre-intermediate
have malls	Verb-noun	Intermediate
love farms	Verb-noun	Intermediate
people scared	Noun-verb	Intermediate
big mall	Adjective-noun	Intermediate
good clinics	Adjective-noun	Intermediate
global schools	Adjective-noun	Intermediate
hospitals well-equipped	Adjective-noun	Intermediate
noisy life	Adjective-noun	Intermediate
green farms	Adjective-noun	Intermediate
big bookstores	Adjective-noun	Intermediate
small clinics	Adjective-noun	Intermediate
entertaining place	Noun-noun	Intermediate
markets bookstores	Noun-noun	Intermediate
shops malls	Noun-noun	Intermediate
mall cinemas	Noun-noun	Intermediate
cinemas parks	Noun-noun	Intermediate
mall bookstores	Noun-noun	Intermediate
mall parks	Noun-noun	Intermediate
parks playgrounds	Noun-noun	Intermediate
phone station	Noun-noun	Intermediate
shining stars	Noun-noun	Intermediate
markets mall	Noun-noun	Intermediate

Appendix 13: The final list of the non-collocations (less idiomatic combinations) extracted from pre-intermediate and intermediate level written texts, and were not identified in the corpus nor by native-speaker informants, organized according to their types

Less idiomatic combination	**Type of the VAN combination**	**Level of the written text it appeared in**	**Suggested correction by the informants**
arranged bag	Verb-noun	Pre-intermediate	Organised bags
take shots	Verb-noun	Pre-intermediate	-Take medicine -receive shots
had vacation	Verb-noun	Pre-intermediate	Experienced a vacation
forget bags	Verb-noun	Pre-intermediate	None
quick order	Verb-noun	Pre-intermediate	-Made our order quickly -quickly ordered -placed an order in the quick order line
face crash	Noun-verb	Pre-intermediate	None
bored time	Adjective-noun	Pre-intermediate	Long time not long bored time
tired brother	Adjective-noun	Pre-intermediate	Brother too not brother also
huge road	Adjective-noun	Pre-intermediate	Long road
long planeting	Adjective-noun	Pre-intermediate	Long planning
food poisonous	Adjective-noun	Pre-intermediate	Food poisoning
candies gifts	Noun-noun	Pre-intermediate	Candies shops not candies and gifts shops
poisoning snake	Noun-noun	Pre-intermediate	Poisonous snakes
daddy meal	Noun-noun	Pre-intermediate	Daddy’s meal
shingle hotel	Noun-noun	Pre-intermediate	None
Forget town	Verb-noun	Intermediate	None
Find playgrounds	Verb-noun	Intermediate	None
like rush	Verb-noun	Intermediate	None
cure soul	Verb-noun	Intermediate	Heal soul (undecided)
contains pollution	Verb-noun	Intermediate	None
villages force	Noun-verb	Intermediate	None
clear life	Adjective-noun	Intermediate	Calm life
school complete	Adjective-noun	Intermediate	None

transportation well-developed	Adjective-noun	Intermediate	None
hospitals complete	Adjective-noun	Intermediate	Equipped hospitals
noisy voices	Adjective-noun	Intermediate	Loud voices
malls schools	Noun-noun	Intermediate	Malls shops
Crowd city	Noun-noun	Intermediate	None
forest farms	Noun-noun	Intermediate	Forests not forest
factories restaurant	Noun-noun	Intermediate	None
Malls hospitals	Noun-noun	Intermediate	None