Chapter 2
The Notion of Collocation

Collocation not only plays a crucial role in language production and comprehension, but also functions as a key indicator of L2 learners’ overall proficiency in the field of second language acquisition. This chapter briefly clarifies the notion of collocation before presenting in the next chapter the reviews of L2 collocation studies. It begins by highlighting the importance of collocation for both native speakers and L2 learners (Sect. 2.1). The second section (Sect. 2.2) proceeds to discuss the different approaches to collocation, and develops a definition adopted in this study. Finally, how collocations were previously classified and the classification of collocation used in the present study are presented.

2.1 The Importance of Collocation

2.1.1 The Pervasiveness of Phraseological Tendency

In the process of speech or text production, complete freedom of choice of a single word is rare and rather there is a phraseological tendency where meanings are created through word combinations (Sinclair 2004: 29). What Sinclair refers to by word combinations are collocations and other features of idiomaticity like fixed expressions, idioms, etc. The phraseological nature of language has long been recognised, as “language does not expect us to build everything starting with lumber, nails, and blueprint, and rather it provides us with an incredibly large number of prefabs” (Bolinger 1976: 1). Research on word combinations has accumulated extensive evidence for this phraseological tendency, either in written or spoken language (e.g. Altenberg 1998; Biber et al. 1999; Cowie 1991, 1992; Howarth 1998a; Kjellmer 1994; Nattinger and DeCarrio 1992; Pawley and Syder 1983; Renouf and Sinclair 1991; Sinclair 1991; Stubbs 2001; inter alia).
A considerable proportion of prefabs has been identified in various kinds of genres of texts. Kjellmer (1987) used a straightforward way to measure the collocational density of two short samples of texts in the Brown corpus and discovered that a large proportion of the text was made up of collocational elements. In examining journalistic writings (news stories and editorials), Cowie (1992: 1) concluded that “journalistic prose draws very heavily on verb–noun collocations that are already well-established and widely known”, as his studies revealed a collocational density as high as more than 40% (Cowie 1991, 1992). In addition to journalistic writings, this collocational density has also been found in academic writings, in which 41% of the verb–noun combinations were found to be conventional collocations (restricted collocations and idioms) (Howarth 1996). Furthermore, in general English writings, still a significant number of fixed phrases and idioms can be found, as reported by Moon (1998) in her examination of the Oxford Hector Pilot Corpus and Birmingham Collection of English Text. An interesting way of proving the strength of phraseological tendency was introduced by Stubbs (2001), who counted the frequency of attested phraseological units of the word forms beginning with the letter f in a 1000-word sample. It was found that all the 47 words with an initial f were in recognisable phrases, which confirmed the ubiquitous presence of phraseological units. In all, as Howarth (1998a: 171) summarised, “there is in native writing an identifiable core of collocational conventionality”. 

In the meantime, the phenomenon that natural language is made up of a large proportion of word clusters is not manifested in written discourse alone. It shows an even stronger tendency in the spoken language. As early as the 1980s, working on data from conversational talk, Pawley and Syder (1983: 215) estimated that “by far the largest part of the English speaker’s lexicon consists of complex lexical items including several hundred thousand lexicalised sentence stems”. Similarly, Jackendoff (1997) collected the data from an American television game show Wheel of Fortune and discovered a high ratio of collocations, idioms and prefabricated phrases. Switching to a different perspective, Altenberg investigated recurrent word combinations retrieved from the London-Lund Corpus of Spoken English, and reported that “over 80% of the words in the corpus form part of a recurrent word-combination in one way or another” (Altenberg 1998: 102). This figure shows an outstandingly high proportion of word combinations, but it encompasses a whole range of recurrent word combinations, a large proportion of which are of little phraseological interest (e.g. the the, and the, in a, out of the) (ibid.). The inclusion of these word sequences is owing to the automatic retrieval method adopted by Altenberg, whose calculation of the percentage of word clusters is based on the

---

1Kjellmer’s recognition of collocation is based on his definition of collocation as a grammatically well-structured sequence occurring more than once (1987: 133). So, more collocations were counted than collocations defined in the present study (cf. Sect. 2.2.2).

2The 1000-word sample was compiled from a 10,000-headword database, which recorded the most frequent content words in the Cobuild (1995) database. So there were no function words beginning with an f (e.g. for).
inclusion of any continuous string of words occurring more than once in identical form. In a similar vein, Biber et al. (1999) identified many lexical bundles (recurrent expressions) in a large corpus. Unlike Altenberg (1998), they set a fairly high threshold level for what qualifies as a lexical bundle—lexical sequences occurring at least ten times per million words and at the same time across at least five different texts in a register. Even with such a high cut-off point between lexical bundles and casual lexical co-occurrences, they discovered a large proportion of lexical bundles: 45% in conversation and 21% in academic prose.

Both the written and spoken languages of native speakers thus exhibit a strong phraseological tendency. The spoken language has been found to consist of a greater proportion of recurrent word combinations than the written language. One reason provided by Biber et al. (1999) is that the spoken language involves a considerable amount of repetitions, which increases the potential proportion of clusters. Another underlying reason explaining why the spoken language is more formulaic might be the time constraints imposed on speakers. Speakers usually do not have enough time to coin novel expressions as they do in writing. This is the case with journalistic reporting, where the intense pressures and time constraints on reporters require them to use a great many familiar ready-made expressions (Cowie 1992). Hence, there is an unavoidably larger occurrence of formulaic language use in spoken than written production. In all, word combinations make up a very high proportion in both the written and spoken performance of native speakers. This phenomenon demonstrates the block-like nature of language and facilitates the inference that “when we speak or write it is therefore often more apposite to say that we move from one cluster to the next than to say that we move from one word to the next” (Kjellmer 1994: ix). The clusters, or multiple-word units, are stored in the psychological lexicon and are believed by Kelly and Stone (1975) to be at least as numerous as single words.

Therefore, as to learners of a second or foreign language, the existence of a large number of word combinations underlying proficient performance requires them to be empowered with this phraseological competence. Phraseological knowledge is naturally of central importance to fluent and idiomatic control of the language for L2 learners as well. The next section moves on to discuss the significance of this phraseological competence for L2 learners.

2.1.2 The Importance of Collocation for L2 Learners

The importance of collocational knowledge for L2 learners has been long and widely recognised (e.g. Cowie 1992; Fox 1998; Kjellmer 1991; Lee and Liu 2009; Lewis 2000; Meara 1984; Palmer 1933; Pawley and Syder 1983; Wray 2002; Yorio 2002). In this research, the terms second and foreign language are used interchangeably, referring to any language learned after one’s native language, although they are differentiated by Richards and Schmidt (2010: 224f) in terms of whether the language is used as a medium of instruction in schools or widely used in a country as a medium of communication by the government, media, etc.
1989). In this section, its significance is briefly summarised from two perspectives: for native-like production and for efficient comprehension.

### 2.1.2.1 Phraseological Knowledge Is Important for Native-like Production

Knowledge of collocations is of the same importance as knowledge of grammar. It is considered key to native-like production, as is claimed by Fox (1998: 33):

> when even very good learners of the language speak or write English, the effect is often slightly odd. There is nothing that is obviously wrong, but somehow native speakers know that they would not express themselves in quite that way. … The problem is often one of collocation.

Here, the oddness of expressions produced by learners is not concerned with the inappropriateness of grammar, but with the co-selected word combinations. To know a language not only requires the knowledge of appropriate rules to generate grammatically well-formed utterances of that language, but also knowledge of which of these grammatical utterances are native-like (Biber et al. 1999; Wray 2002: 143). Failing to appropriately use these lexicalised expressions, as has been pointed in Chap. 1, may even divert the reader’s attention from content to form (Howarth 1998a: 174). As Cowie (1992: 10) acknowledged, it is impossible to perform at a native-like level without knowledge of an appropriate range of multiword units. Therefore, the significance of phraseological knowledge for L2 learners should in no way be downplayed.

A good command of phraseological knowledge helps attain the goal of native-like production through promoting fluency. A store of formulaic units in the mental lexicon plays a key role in reducing the processing effort en route to language production (cf. Hunston and Francis 2000: 271). Unlike the creative side of language production, in which individual words are combined one by one according to grammatical rules, the agglomeration of words into clusters constitutes one single choice and thus saves much processing time (cf. Sinclair 1987: 320). Jackendoff’s analogy between fixed word combinations and chunking in music well illustrates the role of prefabricated units in promoting fluency, as he maintained that:

> any musician can attest the fact that one of the tricks to playing fast is to make larger and larger passages form simplex units from the point view of awareness—to “chunk” the input and output. This suggests that processing speed is linked not so much to the gross measure of information processed as to the number of highest-level units that must be treated serially. Otherwise, chunking wouldn’t help. (Jackendoff 1983: 125)

### 2.1.2.2 Phraseological Knowledge Is Beneficial for Efficient Comprehension

Knowing a wide range of multiword units not only facilitates native-like production, but also contributes to efficient comprehension on the part of L2 learners.
2.1 The Importance of Collocation

Hunston and Francis (2000: 270–271) argued that storing a large number of multiword units in the mental lexicon, learners can understand the meaning of text without having to pay attention to every word. This is beneficial for enhancing both the reading and listening efficiency. They further pointed out that knowledge of phraseological patterns can help L2 learners reconstruct the meanings even if they mishear some words in speech. At a micro-level, knowledge of co-occurring word combinations contributes to successful comprehension of the semantics of each constituent. For example, through a corpus-based analysis of the collocations with affect/influence, Lee and Liu (2009) exemplified how the use of collocations provides a solid conceptual grounding of the target word for L2 learners in grasping the lexical semantics of the two words.

In sum, in the process of striving for native-like language production, phraseological knowledge is, on the one hand, important for L2 learners’ idiomatic and fluent production; on the other, it helps promote the efficiency of language comprehension in general and the comprehension of lexical semantics of individual words. Collocation is thus recognised by Lewis (2000: 45) as “the most powerful force in the creation and comprehension of all naturally-occurring texts”.

2.1.3 Summary

In this section, we have placed collocation within the context of formulaic language and reviewed its importance for both native speakers and L2 learners. First, the ubiquitousness of formulaic language in either spoken or written language has long been acknowledged and verified in previous studies. Given the pervasiveness of conventionalised word combinations, it follows that L2 learners have to gain a good command of them in order to achieve native-like proficiency. A good control of formulaic language not only facilitates idiomatic production, but also promotes efficient language comprehension. Collocation is one of the most important and frequent aspects of formulaic language and constitutes the target of this study. The next section will be devoted to a deeper discussion of the nature of collocations.

2.2 The Notion of Collocation

Given the abundance of terminology in the field of phraseology (cf. the various expressions mentioned in Sect. 2.1, e.g. collocations, fixed expressions, idioms, prefabs, complex lexical items, multiword units, etc.), a clarification of which of

---

4Kjellmer (1994: xi) listed various terms referring to clusters of words: expressions, fixed combinations, formula units, formulas, larger-than-word units, lexical phrases, lexicalised sentence stems, multiword lexical units (MLU), multiple-word units, patterned speech, patterns, phrases,
these aspects of formulaic language forms the object of this study is in order. The present study will focus on the most common manifestations of formulaic language—collocation. Yet as Bahns (1993: 57) admitted, “regrettably, collocation is a term which is used and understood in many different ways”. So the primary aim of this section is to summarise previous definitions and classifications and develop a definition and classification of collocation in order to identify those word combinations in learner English.

2.2.1 Collocation Previously Approached

Collocation, which refers to syntagmatic lexical relations in a language, can be traced back to as early as the 1930s. Palmer (1933: title page) defined the collocation as “a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts”. Examples of such a definition by Palmer are to strike while the iron’s hot, thank you and to commit suicide. Though Palmer used “collocation” as an umbrella term to generally refer to all “comings-together-of-words”, he is believed to be the first to use collocation in its present-day sense. Yet collocation approached by Palmer is mainly pedagogically oriented, and it is not clear from his definition what kind of co-selecting relationship between two or more words can qualify them as a collocation (and thus be learnt as an integral whole). These classifying criteria were later developed by Russian phraseologists like Vinogradov. Based on an analytical framework of descriptive categories and regarding collocations as a type of word combinations with a degree of inseparability or fixedness, collocation approached this way is termed the phraseological approach (Nesselhauf 2004) [or “significance oriented approach” termed by Herbst (1996: 380)], which shall be discussed in detail in this section.

The notion of collocation formally came into being in the 1950s when Firth, commonly accredited as the father of collocation, viewed collocation from a purely linguistic standpoint and put forward the notion of collocation through the celebrated dictum: “you shall know a word by the company it keeps” (Firth 1957: 179). This statement has been endorsed by many linguists and also been scientifically validated through corpus-based studies as we shall see below. Definitions of collocation in various forms following Firth are called the Firthian approach [or “statistically oriented approach” called by Herbst (1996: 380); the “frequency-based approach” by Nesselhauf (2004)].

(Footnote 4 continued)

prefabricated speech, ready-made utterances, recurrent combinations, stock phrases and word-like units. See also the terms to describe the phraseological phenomenon in Wray (2002: 9).

5 Fellbaum (2007: 8) distinguished collocation, a linguistic phenomenon, from collocations, specific lexical instances resulting from collocation that are part of the lexicon. No differentiation is attempted in this study.
Collocation has also been psychologically envisaged (Aitchison 2003). In what follows, we shall present three main approaches to collocation: the psychological approach; the Firthian approach and the phraseological approach. The psychological approach will be briefly discussed and the latter two will be more elaborated since the Firthian approach sets a trend for lexical studies in corpus-based research and the phraseological approach concentrates primarily on the classifying criteria of collocations, which is particularly useful for collocation studies in the field of second language acquisition.6

2.2 The Notion of Collocation

2.2.1.1 The Psychological Approach

Collocation involves strong associations between words. This association can be frozen into one type of the meanings of a word, defined as the collocative meaning, which “consists of the associations a word acquires on account of the meanings of words which tend to occur in its environment” (Leech 1974: 20). Leech (ibid: 20) gave the example of pretty and handsome, which have the similar meaning of “good looking”, but can be differentiated by the range of nouns with which they take, e.g. (handsome) man, and (pretty) woman. This definition of collocation, concerned with the (collocative) meaning of a word through association with its likely-to-occur collocates, is viewed as a “psychological” or “associative” definition (Partington 1998: 15). The associative tendency of words is so strong that in the mental lexicon the number of collocations is inferred by Kelly and Stone (1975) and Pawley and Syder (1983) to be as many as single words. The claim that a word strongly associates with other words is not only evidenced through the large existence of clusters as discussed in Sect. 2.1.1, but also verified through word association tests in the field of psycholinguistics. Aitchison (2003: 86) reported that the second commonest type of response to stimulus words in a test is collocation.7 For example, water, sea, shaker and lake were among the top ten commonest responses to the word salt, which shows that words are stored in the mental lexicon in connection with their collocates. Tongue slips, according to Aitchison, constitute another interesting form of evidence that words are linked with their collocates in the mental lexicon, as there are cases when “people sometimes start out with one phrase and then get ‘derailed’ on to a familiar routine, as in Hungarian restaurant

---

6This classification of the different approaches to collocation is similar to previous collocation reviews. For example, definitions of collocation have been neatly summarised by Partington (1998) into “textual”, “psychological” or “associative”, and “statistical” ones, whilst Handl (2008) classifies previous definitions into four categories: text-oriented, association-oriented, statistically oriented and semantically oriented. Herbst (1996) distinguishes three approaches in collocation: “statistically oriented approach”, “significance oriented approach” and “text-oriented approach”. Nesselhauf (2004) summarises the approaches of collocation as the “frequency-based” and “phraseological” approach.

7According to Aitchison (2003: 86), the commonest type of response to stimulus words is coordination, e.g. salt with pepper, butterfly with moth.
for ‘Hungarian rhapsody’” (Aitchison 2003: 91). Words in a collocational relationship are believed to be stored in a single remembered set from which they can be retrieved (cf. Greenbaum 1974: 80).

Therefore, the collocating relationship of words is psychologically real and takes a major position in the mental lexicon of language users. Yet these associating bonds between words stored in the mental lexicon of native speakers might be quite different from those in a L2 learner, whose mental lexicon, as Meara (1984: 232) put it, is “in general more loosely organised than the native speaker’s lexicon”. When it comes to the actual use of collocation, the associative bond in the mental lexicon may work well for native speakers, but probably poses difficulty for non-native speakers. For example, when both a native speaker and an L2 English learner are asked to express strong coffee, the collocate strong can be easily associated with coffee by native speakers, but learners might use other collocates like powerful rather than strong. It thus requires consciousness and efforts for English learners to build up that associative bond between words.

The associative power between words in a syntagmatic relation helps the prediction of the co-occurring words to a greater or lesser extent, as the word bonsai has a strong prediction for tree and spick for span, but pill-box cannot be forecast by letter (Crystal 1997; Stubbs 2001: 29). The realisation of predictable word combinations in texts is collocation—two or more words that tend to co-occur (Lewis 2000: 73). So combining the psycholinguistic phenomenon of collocation and its realisation in texts, collocation is defined by Hoey (2005: 5) as a psychological association between words and “evidenced by their more frequent occurrence together in corpora more often than is explicable in terms of random distribution”. Next follows a discussion of this text-based study of collocation.

2.2.1.2 The Firthian Approach

Firth’s statement that “you shall know a word by the company it keeps” is well exemplified by the co-occurring words dark night, where he claimed “one of the meanings of night is its collocability with dark, and of dark, …, collocation with night” (1957: 196). The meaning by collocation, as Firth argued, is “an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words” (ibid: 196). Firth put forward a significant conception as to the realisation of meaning by its instantiations with co-occurring words. At a time when “the idea the language is based on a system of rules determining the interpretation of its infinitely many sentences is by no means novel” (Chomsky 1965: v), Firth’s meaning by collocation was fresh. This conception has become a substantial and new impetus in observable-text-based and later computer-assisted studies on collocation and established the British traditions in text analysis (Stubbs 1996). Taking inspiration from Firth’s definition of collocation, the Firthians have conducted studies of word co-occurrences based on real language in use, and proposed other definitions.
Sinclair is the main inheritor and innovator of the Firthian approach, along with other linguists who followed the British tradition and viewed collocations primarily as a syntagmatic relation between words in texts (Halliday 1966; Hoey 1991, 2005; Kjellmer 1987, 1994; Lewis 2000; Moon 1998; Sinclair 1966, 1987, 1991, 2004; Stubbs 1996, 2001; etc.). Given the abundance of studies on collocation in this trend, these studies are summarised in two subsections: one discusses the notion of collocation in word sense recognition and differentiation, and the expansion of the notion of collocation to other aspects, such as colligation, semantic prosody and semantic preference; the other discusses frequency-based studies of collocation focusing on defining and recognising collocations based on word frequencies. These two lines of studies are however not mutually exclusive and only discussed separately for the purpose of stressing their differences.

Text-Oriented Studies on Collocation

Collocation refers to a patterning of language with tendencies of lexical items to co-occur (Sinclair 1966). These co-occurring tendencies of words are instantiated in texts, a fact which induces a text-oriented definition of collocation by Sinclair (1991: 170) as “the occurrence of two or more words within a short space of each other in a text”. Based on this definition, word combinations are considered to be in collocational relationships as long as they are within a short space of each other. Thus, the strong associative bond between words is not retained in Sinclair’s definition, but this definition constitutes the precondition to the study of collocation in texts, since the occurrence of two or more words specifies the forms of all collocations and a text forms the basic medium where collocations are homed and recognised. Further specification to what counts as a short space of each other is developed:

We may use the term node to refer to an item whose collocations we are studying, and we may define a span as the number of lexical items on each side of a node that we consider relevant to that node. Items in the environment set by the span we will call collocates. (Sinclair 1966: 415)8

Through a computer-based study, Jones and Sinclair (1974) discovered that significant collocates usually fall in a span of 4:4, that is, four words to the left and four words to the right of the node (cf. Sinclair 1991: 170). Based on this text-oriented study of collocation, the notion of collocation has been utilised by Sinclair for the study of lexis, lexis and grammar and the expansion of the concept of collocation to other aspects. Collocational information is useful for word sense recognition. For instance, with the evidence of collocates like per, average, population, economic profitability, gradual, sharp, slowly, the sense of the node word decline is recognised as a reduction in size, whilst another sense of decline—deterioration—is supported by words like sad, suffered (Sinclair 1991). This sense

---

8Words in boldface are quoted in their original forms.
differentiation method based on actual language use has revolutionised lexical research and has been widely followed in lexicography, with the Collins COBUILD English Dictionary as a typical example.

Collocational evidence also contributes to the combining of lexis and grammar, as illustrated by Sinclair in the word *yield* (1991: 56). He found 33 corpus instances of *yield* showing the sense of “give way”, realised by *yield* used as an intransitive verb; 30 cases meant “produce”, realised as a noun. In 15 cases, *yield* was used as a transitive verb, meaning “lead to”. One might argue that words of different parts of speech (structure) naturally have different meanings (sense), as the senses of the polysemous word *drink* are quite different depending on its use as a verb (*take in liquids*) and a noun (*any liquid suitable for drinking*). Yet what is significant in Sinclair’s demonstration is a bottom-up approach, i.e. word sense is recognised through word co-occurrences by using authentic language data (this method is widely embraced by Hoey 2005; Renouf 1987; Stubbs 1996, 2001; Teubert 2010; inter alia). So essentially he approaches the study of lexis in a data-driven fashion.

The close correlation between sense and structure as revealed through collocational information is also strengthened by the fact that even different word forms of the same lemma have quite different collocational behaviour. In a 130-million corpus, Stubbs (1996: 172) found that for the lemma *educate*, the most frequent word form is *education* and it collocates primarily with terms denoting institutions (e.g. *further, higher, secondary, university*). The base form *educate*, on the other hand, collocates with synonymous verbs such as *enlighten, entertain, inform*, etc. Moreover, different word forms can enter similar collocations and this therefore induces the definition of collocation as a relationship between lexemes (Halliday 1966; Sinclair 1991). For example, according to Halliday (1966: 151), “he argued strongly, I don’t deny the strength of his argument, his argument was strengthened by other factors” would all be considered instances of the same collocation as *strong argument* (cf. Greenbaum 1974: 80).

The notion of collocation has further been expanded to more abstract levels, such as colligation, semantic prosody and preference. There are syntactic constraints on a word’s selection of its co-occurring words: these constraints are called *colligation* (Firth 1957). Colligation refers to the co-occurrence of grammatical choices (Hoey 2005; Sinclair 1996, 1998, 2004). Compared with collocations, which are directly observable in texts, colligations are not so directly observable and involve abstractions based on generalisations about the behaviour of the word in question (Stubbs 2001: 88). *Consequence*, for example, tends to co-occur with the preposition *of* (Hoey 2005). Besides the grammatical constraints on the collocates of a word, a word can also co-occur with positive or negative groupings of words and as a result it is presented with a certain semantic prosody, defined by Louw (1993: 157) as “the consistent aura of meaning with which the form is imbued by its

---

9Explanations of *drink* are quoted from WordNet.

10Lemma refers to the composite set of word forms. For example, the lemma *give* refers to the forms of *give, gives, given, gave* and *giving* (Sinclair 1991: 41–42).
collocates”. For example, the phrase *set in* primarily co-occurs with an unpleasant state of affairs and has a negative prosody (Sinclair 1991: 68, for semantic prosody, cf. Sinclair 2003; Stubbs 1995a, b, 1996, 2001). Other words would usually collocate with a certain semantic preference, as *the naked eye* collocates with verbs and adjectives indicating visibility and the word *unemployment* usually collocates with the semantic set of statistics (Sinclair 1991: 33; Stubbs 1995b: 254). Semantic preference is therefore an abstraction of the semantic orientations over the collocates of the node word.

**Frequency-Based Studies on Collocation**

Besides the loose treatment of collocation as co-occurring words within a set span, other researchers reserve the notion of collocation for statistically significant co-occurring words and define collocation as “the relationship a lexical item has with items that appear with greater than random probability in its (textual) context” (Hoey 1991: 7, 2005; cf. Greenbaum 1974; Moon 1998; Sinclair et al. 2004; Stubbs 2001). The higher the probability is, the more likely for a word combination to be a collocation. Significant collocations are quantitatively identified by using statistical formulae (cf. Church et al. 1991; Church and Hanks 1990; Church and Hindle 1990; McEnery and Wilson 1996; McEnery et al. 2006; Stubbs 1995a). Within the field of frequency-based definition of collocation, some definitions purely rely on frequency, as Moon (1998: 26) considered a collocation as that which “typically denotes frequently repeated or statistically significant co-occurrences, whether or not there are any special semantic bonds between collocating items”. Yet frequency alone is not a reliable criterion for identifying meaningful collocations. Other researchers add a grammatical standard as well as with frequency and define collocation as recurring sequences of items that are grammatically well formed (cf. Johansson and Hofland 1989: 95; Kjellmer 1987: 133, 1994: xiv). According to Kjellmer (1994: xv), sequences that have no or only a very distant grammatical relationship are excluded. For example, instances like *but too, day but, however in the, night he* would not be considered as collocations even if the frequency criterion was satisfied. Instead, *by me, in April, of the Government* all qualify as collocations (ibid: xiv). However, even though the definition incorporates grammatical well-formedness, it is not sufficient to distinguish between combinations formed on the basis of grammatical rules (e.g. *by me, in April*) and collocations of phraseological value (e.g. *make a decision, strong argument*). The approach to identifying collocations that are of phraseological value is the phraseological approach, which will be illustrated in the following section.

**2.2.1.3 The Phraseological Approach**

In *Aspects of the Theory of Syntax*, Chomsky (1965: 190f) distinguished two types of word relations: a close construction (as *decide on a boat* in the sense of *choose*...
the boat) and a loose association (as decide on a boat meaning decide while on a boat). This distinction is much the same as collocations and free combinations, where close construction refers to collocation (e.g. the verb decide occurring together with the particle on to mean choose), which represents a unit, and loose association resembles free word combinations, which are constructed on the basis of grammatical rules. The phraseological approach is concerned with the defining criteria of collocation and demarcating it from other types of word combinations.

The phraseological approach, in contrast with the psychological and Firthian approaches, concerns itself with classifying schemes of phraseological units according to their varying degrees of fixedness. Russian phraseologists such as Vinogradov (1947, cited in Cowie 1998: 4f) established three categories of word combinations: “phraseological fusions” (e.g. spill the beans), “phraseological unities” (e.g. blow off steam) and “phraseological combinations” (e.g. meet the demand). Different phraseologists adopt slightly different classifications with different terminology, as summarised in the table below:

As Table 2.1 shows, word combinations are generally divided into idioms, collocations and free combinations, which are on a spectrum from the most fixed to the most free. What is also revealed through the above table is that generally there are two separate directions of interests distinguishing the Russian phraseologists (Vinogradov and Amosova) from other phraseologists like Aisenstadt, Cowie and Howarth. The Russian scholars start from the idiomatic spectrum to delineate the specific phraseological zone and are preoccupied with the distinction between idioms and collocations. Their classifying criteria will not be elaborated here, since what is more challenging and significant for L2 learning is the distinction between collocations and free word combinations, and to “identify at what point language users are manipulating expressions as wholes rather than composing them

Table 2.1 Classifications of English word combinations

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Classifications of word combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinogradov (1947)</td>
<td>Phraseological fusion Phraseological unity Phraseological combination</td>
</tr>
<tr>
<td>Amosova (1963)</td>
<td>Idiom Idiom (not differentiated) Phraseme, or phraseoloid</td>
</tr>
<tr>
<td>Aisenstadt (1979)</td>
<td>Idiom Restrict ed collocation Free word combination</td>
</tr>
<tr>
<td>Cowie (1981)</td>
<td>Pure idiom Figurative idiom Restricted collocation Open collocation</td>
</tr>
<tr>
<td>Nattinger and DeCarrico (1992)</td>
<td>Idiom Collocation Free combination</td>
</tr>
<tr>
<td>Howarth (1996)</td>
<td>Pure idiom Figurative idiom Restricted collocation Free collocation</td>
</tr>
<tr>
<td>Howarth (1998a, b)</td>
<td>Pure idiom Figurative idiom Restricted collocation Free combination</td>
</tr>
</tbody>
</table>

These classifications are partly cited in Cowie (1998: 7), and largely modified in this research. See also Howarth (1996: 34) for a summary of the terminologies.
According to generative rules” (Howarth 1996: 31). Therefore, special attention is paid to the distinction between collocations and free combinations.

As is acknowledged by Cowie (1998: 5), phraseological combination, or restricted collocation, is the most interesting and yet most difficult to delimit. It is difficult to delineate because collocation is located in the fuzzy zone between free combinations and idioms. Previous researchers distinguish the three types of word combinations: idioms, collocations and free combinations in terms of semantic transparency, semantic specialisation of one element in the combination and commutability/substitution of one of the elements.

Semantic Transparency/Opacity

Semantic transparency/opacity is measured in terms of whether the meaning of the whole combination can be deduced from the meaning of the individual elements. This criterion is well suited to the differentiation of idioms and non-idiomatic expressions (Hausmann 1989). For the former, the semantics of the whole combination is opaque in that their meanings are not made up of the sum of their constituents (e.g. *kick the bucket*, *spill the beans*). Yet the meanings of non-idiomatic word combinations, namely collocations and free combinations, are easily derivable from their constituents. For example, the meaning is transparent in both *commit a crime* as a collocation and *control the crime* as a free combination). So collocations can be differentiated from idioms by applying the criterion of semantic opacity. But this criterion fails to demarcate collocations from free word combinations, given that both types are transparent in meaning. Criteria central to the distinction of collocations and free combinations are: one element used in its specialised sense and the degree of commutability of either of the constituents (cf. Aisenstadt 1979; Cowie 1981, 1992, 1998; Howarth 1996; 1998a).

Specialised Senses of One Element

Phraseologists distinguish collocations from free combinations in terms of the senses/meanings of the constituents, and claim that for a word combination to qualify as a collocation, either of the elements must have a specialised meaning. What they mean by specialised meaning are figurative senses (as *pay in pay one’s respects*, *adopt in adopt a policy*), technical senses (as *obtain in obtain a warrant*) and delexical senses (as *make in make a decision*) (Aisenstadt 1979; Cowie 1991, 1992, 1998; Howarth 1998a, b; Moon 1998). The requirement for either of the elements to have a specialised meaning is meant to exclude free combinations, for which both elements are used in their literal senses (e.g. *bake bread*, *cut cheese*). However, it is not always easy to discern whether the sense of one element is specialised. Take the collocations used by Howarth (1998a: 170) as an example:
(1) Figurative: require qualifications
(2) Delexical: give evidence of
(3) Technical: obtain a warrant.

A note of caution is due here concerning the senses of the three verbs (require, give, and obtain). In example (1), the verb require in require qualifications may not be used in the claimed figurative sense; rather, it is in its literal sense “to ask”/“to request”; so is the verb give in example (2), which is a delexical verb but means provide in the context of give evidence; In example (3), though the whole combination is used in a technical text, obtain is used in its original meaning—get.

So the figurative, delexical and technical senses complicate the categorisation process and are not universally reliable. Meanwhile, the application of this criterion in delimiting collocations in turn excludes a large number of real collocations, such as commit a crime, for which no specialised senses are involved and both of the elements are used in their literal senses. So the criterion that either of the constituents must have a specialised sense does not qualify as a defining criterion in the definition proposed later in this study.

Commutability/Substitutability

Unlike free combinations which are subject to free substitution of either element without a consequent alteration in the meaning of the other, collocations are restricted in the commutability of either element (Aisenstadt 1979; Cowie 1992; Howarth 1996, 1998a). Aisenstadt (1979: 73) illustrated restricted commutability in the following two examples:

(4) 

\begin{align*}
\text{shrug one’s shoulders} & \\
\text{shrug something off} & \\
\text{shrug something away} & \\
\text{shrug one’s shoulders} & \\
\text{square one’s shoulders} & \\
\text{hunch one’s shoulders} & \\
\end{align*}

(5) 

\begin{align*}
\text{make a decision} & \\
\text{take a decision} & \\
\text{have a look} & \\
\text{give a look} & \\
\text{take a look} & \\
\end{align*}

In example (4), both shrug and shoulders are restricted to a number of co-occurring words and neither of them can be substituted; In example (5), there is a restricted commutability on the verbs, as decision is limited in alternative verb collocates: make/take, and look in verbs such as have/give/take. Aisenstadt attempted to demarcate collocations according to the restricted substitutability of word constituents. Yet on the one hand, commutability itself is a vague criterion and depends much on the conceivability of a human mind. With shrug one’s
shoulders for example, shoulders can have a rather wide set of verbs to go with, as in straighten one’s shoulders, wash one’s shoulders, look at one’s shoulders, rub one’s shoulders, scratch one’s shoulders (Nesselhauf 2005: 27). This is also the case with decision, which can co-occur with a variety of verbs, such as reach a decision, come to a decision, postpone a decision, criticise a decision, explain a decision (ibid: 27). On the other hand, commutability can also be restricted in free combinations like wash the glass, since substituting the verb clean for wash slightly alters the original sense and the same applies to replacing the noun glass with cup. So what qualifies the two combinations as collocations is the fact that the word shoulders has a rather restricted set of co-occurring words with the sense of “shrug” in shrug one’s shoulders (probably only the verb, i.e. shrug) and decision has a restricted number of verbs with the sense of “make” in make a decision (make/take/reach, etc.). The notion of the given sense was adopted by Cowie (1992: 5f) in his commutation tests to demarcate restricted collocations. The commutability of the verb is tested through whether it is the only verb or one of a set of synonymous verbs used in the appropriate sense in relation to a given noun (e.g. verbs are commutable in abandon/give up a cherished principle, but verbs are not commutable in run a deficit).

A comprehensive classification of collocations on the basis of commutability was established by Howarth (1996: 102) in his categorisation of verb + noun collocations from the most free to the most restricted (from L1 to L5), as summarised in Table 2.2.

From L1 to L5, restrictedness of collocations is scaled from a slight degree of restriction of one element to complete restriction of both elements and this restriction is explained by the number of synonyms either element can take. For example, for L1 collocations, nouns are subject to free substitution whilst restriction is placed on the verbs because of the limited number of synonymous verbs. When neither element permits substitution, i.e. with no synonym in the given sense, the word combination is the most restricted collocation (L5), such as curry favour.

However, this classifying scheme complicates the differentiation between collocations and free combinations once the notion of synonyms is introduced. Like

<table>
<thead>
<tr>
<th>Verb</th>
<th>Noun</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Some restriction</td>
<td>Free substitution</td>
<td>Adopt/accept/agree to a proposal/suggestion, etc.</td>
</tr>
<tr>
<td>L2 Some substitution</td>
<td>Some substitution</td>
<td>Introduce/table/bring forward a bill/an amendment</td>
</tr>
<tr>
<td>L3 Some substitution</td>
<td>Complete restriction</td>
<td>Pay/take heed</td>
</tr>
<tr>
<td>L4 Complete restriction</td>
<td>Some substitution</td>
<td>Give the appearance/impression</td>
</tr>
<tr>
<td>L5 Complete restriction</td>
<td>Complete restriction</td>
<td>Curry favour</td>
</tr>
</tbody>
</table>
the notion of commutability, the number of synonyms is also subject to the con-
ceptualizability of a human mind. With the examples in L3 for an example, the com-
bination pay heed is considered as a restricted collocation in the sense that heed is
completely restricted in its substitution. Yet according to the Oxford Dictionary of
Synonyms and Antonyms, attention is in a synonymous relationship with heed and
pay attention is an acceptable English collocation. The example of give appearance/impression classified in L4 has the same problem, as the verb give can be replaced by make/leave given that make appearance/impression or leave appearance/impression are expressions with similar meanings. So the judging on
the number of synonyms requires a good deal of subjectivity. As with Cowie’s
commutation test in which verbs are measured in terms of the number of synonyms
they have, it is hard to find synonyms for verbs even in free combinations such as
open the door (?unblock, ?unlock). In cases where no synonyms are found, it can
just as well be a free combination rather than a restricted collocation, e.g. drink
one’s tea (Nesselhauf 2005). Commutability is not a clear criterion for differenti-
ating collocations from free combinations.

In this section, the notion of collocation has been first introduced in the domain
of psychological studies, with collocation viewed as psycholinguistic lexical
associations. Another field in which collocation has been researched is the
text/frequency-based studies of collocation. Much text/frequency-based research
focuses on the collocational relationship between words, the extension of the notion
of collocation to more abstract levels, such as colligation, semantic prosody,
semantic preference and the identification of significant collocations. However, the
Firthian approach is based on linear co-occurrence of items and takes little account
of the syntactic and semantic statements that are essential in treating collocations
(Greenbaum 1970: 10). In addition, the span established for identifying colloca-
tions—four words each side—is insufficient to account for certain common collo-
cations (e.g. collect stamps in the following examples):

(6) They collect many things, but chiefly stamps.
(7) They collect many things, though their chief interest is in collecting coins. We,
however, are only interested in stamps (Greenbaum 1970: 11).

So the frequency-based approach, although it can identify significant colloc-
tions of statistical value, cannot incorporate all the collocations of phraseological
value (like collecting stamps in the above examples). Moreover, the Firthian tra-
dition is preoccupied with collocation as a linguistic phenomenon per se and is not
concerned with demarcating collocations from other types of word combinations.
As discussed above, the notion of restriction inevitably forms part of accounting for
what a collocation is and this restriction distinguishes it from other forms of lexical
coop-occurrences (e.g. free word combinations and idioms). Measured frequency
of co-occurring words is not a significant measure of collocational restriction

---
11The verb collocates of impression—make and leave—are listed with reference to a collocation
the phraseologists on the other hand have proposed categorisation frameworks of word combinations. The separation of collocations from free combinations is of essential importance in the investigation of collocations used by L2 learners, since that constitutes the first step in examining what is phraseological rather than what is free (cf. Howarth 1996). However, even with the widely adopted defining features in demarcating collocations within the phraseological approach, a clear borderline between free combinations and collocations still cannot be set. The next sections, then, continue to examine the definition of collocations within the phraseological approach, attempting to develop a usable categorisation of collocation and discussing previous classifications of collocations.

2.2.2 Collocation Defined in This Study

Since this study is situated in the field of second language acquisition, aiming at measuring nonnative phraseological performance, the approach taken to collocation is mainly phraseological, in order to delimit collocations from idioms and free combinations in learners’ English writings. Collocation within this approach has been defined by previous researchers in more or less the same way, adopting the criteria of semantic transparency/opacity, specialised sense of one element and commutability (see definitions summarised in Table 2.3).

As demonstrated in the previous section, collocations can be distinguished from idioms by applying the criterion of semantic transparency, i.e. the former are relatively transparent in meaning (e.g. make a decision) and not as opaque as idioms (e.g. kick the bucket). Another criterion—the specialised senses required of at least one element of a word combination—is weaved, since certain collocations with both elements used in their literal senses are excluded otherwise (e.g. commit a crime, answer questions). As for the criterion of commutability measured in terms of the number of synonyms an element can take, though it contributes to the identification of the restrictedness in collocations, it operates more or less at an intuitive level.

Therefore, there is a need for a clear definition using terminology which avoids blurring the notion of collocations. What is commonly acknowledged is that there is restricted commutability in either of the constituent words in a collocation. In other words, either of the two elements has a limited set of words with which to co-occur (Cowie 1981, 1998; Howarth 1996). For example, the noun stir in the given context of “make a stir” has a limited set of verbs: cause/create/make a stir (Cowie 1981: 228). Or a verb in a given context has a limited set of nouns (e.g. pay one’s respects/a compliment/court) (Cowie 1998: 216). The limited set of verbs/nouns is termed collocational range, referring to the number of co-occurring words a word can take (see also Cowie 1981, 1998; Granger 1998; Handl 2008; Leech 1974: 20; Nesselhauf 2003; Philip 2007). In Greenbaum’s (1974: 80) words, the notion of collocational range is exemplified by turn on, which “collocates with (among other
items) light, gas, radio, and TV…These items and others we might add to them constitute the COLLOCATIONAL RANGE of turn on”.

Collocational range is used as a criterion for distinguishing phraseological units in that elements in collocations have a restricted range of co-occurring words. With the example of commit a crime, commit has a restricted range of nouns, such as crime, wrongdoing, murder, and thus commit a crime qualifies as a collocation. Combinations with both elements having a wide/unrestricted range of co-occurring words are free word combinations, e.g. want a book, for which the verb want can occur with, a car, money, peace, etc., and the noun book can occur with have, buy, read, take, etc.
Therefore, this study utilises two essential defining criteria in defining colloca-
tions, namely, semantic transparency and the range of co-occurring words.
Collocations are then defined as combinations of two or more words which are
classified as having a restricted range of co-occurrence in at least one of their con-
stituent words and by relative transparency in meaning.

Based on this definition, we propose that word combinations with both elements
taking a wide range of co-occurring words are classified as free combinations;
combinations in which either or both elements have a restricted range of co-
occurring words, and also are transparent in meaning, are categorised as col-
locations; combinations with both elements having a very restricted range of co-
occurring words and being opaque in meaning are viewed as idioms (see
Table 2.4 for a detailed illustration).

According to this framework, want a book is a free combination since both the
verb and noun have an unlimited range of co-occurring words. Call the shots is an
idiom with both the verb and noun having a very restricted range of words and
being semantically opaque. Both free combinations and idioms are disregarded in
this study. The focus is on collocations such as pay heed, commit a crime and curry
favour. This framework is a simplified version of Howarth’s categorisation of
collocations into five levels of restrictedness and Nesselhauf’s five groups of
combinatory possibilities of verbs in verb–noun combinations (cf. Howarth 1996:
102; Nesselhauf 2005: 30).

### 2.2.3 Collocations Classified in This Study

Different approaches to collocations result in different classifications. This section
concerns itself with a brief presentation of previous classifications, especially those
that are relevant to the present study.

Based on the strength of associations between words, Aitchison (2003: 91)
distinguished three types of collocations from the loosely to the most strongly
associated: words that are optionally, yet commonly associated (e.g. fresh-faced

<table>
<thead>
<tr>
<th>Verb</th>
<th>Noun</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Free combination | − | − | Want a book (car, money, peace, etc.)  
Want/have/buy, etc. a book |
| Collocation | + | − | Pay/take heed; make/take decision  
Commit a crime/murder; shrug one’s shoulders |
| − | + | Curry favour; curry/court favour |
| Idiom | ++ | ++ | Call the shots; face the music |

Note “−” means the word in question has a wide range of co-occurring words; “+” represents a restricted range of words; underlined words in the Examples column are those words whose ranges are considered.
youths), words with habitual connections or clichés (wide awake) and words frozen into a fixed order or “freezes” (knife and fork). This framework of collocation classification resembles that of Howarth’s categorisation of collocations from the least to the most restricted. The difference lies in the criteria they adopted, namely the strength of association by Aitchison, as opposed to the analytical method of semantic commutability by Howarth. The common denominator is that both acknowledge the degree of fixedness in collocations. If words are strongly associated, they tend to co-occur more often than would be expected in texts. This leads to the classification of collocations in frequency-based studies, where collocations are classified into significant and casual ones (cf. Jones and Sinclair 1974; Sinclair 1987, 1991; Sinclair et al. 2004). Moon (1998: 27) made a distinction among collocations based on the constraints where collocation arises. The simplest collocations are semantically constrained and represent co-occurrence of the referents in the real world (strawberry jam); the second kind is constrained both lexi-grammatically and semantically and “arises where one word requires association with a member of a certain class or category of item” (rancid butter); the third type is syntactically constrained and arises where a word requires complementation with a specified particle (too—to).

Collocations to be classified in this study are neither based on the psychological approach or frequency-based approach. As discussed in Sect. 2.2.2, the definition of collocation is phraseological, and thus its classification is not meant to be based on restrictedness of combinations (cf. Howarth 1996, 1998a, b; Nesselhauf 2005); rather it is broadly based on the word classes of its constituents, since the study aims at investigating L2 learners’ performance with regard to certain types of collocations and its relationship with vocabulary growth.

According to the syntactic structures of collocations, Hausmann (1989: 1010) divided collocations into the following six types:

- adjective + noun (heavy smoker)
- (subject-) noun + verb (storm rage)
- noun + noun (lemon tree)
- adverb + adjective (deeply disappointed)
- verb + adverb (criticise severely)
- verb + (object-) noun (stand a chance).

Similar classifications were also proposed by Benson (1985) and Benson et al. (2010), in whose classifications, collocations were further divided into grammatical and lexical collocations. A grammatical collocation, according to Benson et al. (2010), is a phrase consisting of a dominant word and a preposition or a grammatical structure; lexical collocations resemble those in Hausmann’s classifications, which consist of nouns, adjectives, verbs and adverbs. In the classification put forward by Benson et al. (2010), verb + noun collocations were further divided into CA collocations (collocations containing a verb denoting creation/activation with a noun) and EN collocations (collocations containing a verb denoting eradication/nullification with a noun) (see Table 2.5 for classifications of lexical collocations).
Among the lexical collocations categorised by Hausmann (1989) and Benson et al. (2010), verb + noun, adjective + noun and noun + noun collocations fall into the domain of this study, with the exception that verb + noun collocations are not divided into CA and EN ones. As part of the aim of this research is to investigate the growth of vocabulary produced by Chinese EFL learners, i.e. the growth of verbs from delexical to lexical ones, verb + noun collocations are accordingly further divided into delexical verb + noun and lexical verb + noun collocations.

Delexical verbs, also known as light verbs, such as make, have, or take, are commonly defined as those verbs “whose semantic content is ‘light’ (or has little lexical meaning), as opposed to ‘heavy’ (or lexically more specified), and much of the semantic content is obtained from its arguments” (Miyamoto 2000: 12). Although the total number of delexical verbs in the English language is small, they have very high frequency of occurrence and are the commonest words (Sinclair and Fox 1990: 147). Examples of light verb + noun constructions are make progress, have a discussion and take a bath, where the main semantic content is provided not by the verbs, but by the following nouns. The verbs are semantically general and the object nouns are semantically specific (Algeo 1995). Most delexical structures (a delexical verb followed by a noun group) can be replaced by an analogous single word verb (e.g. give advice = advise), though some cannot be replaced by a single verb, e.g. give evidence, give birth.

In this study, delexical verbs are used in a broad sense and no differentiation is made regarding the semantic contents they carry in constructions like give advice and give evidence (cf. Wang 2011). The six most common delexical verbs targeted are do, give, have, make, take and get (Chi Man-lai et al. 1994; Kaszubski 2000; Sinclair and Fox 1990; Wang 2011). All verb + noun collocations with the above six verbs are considered as delexical verb + noun collocations.

12Reasons for focusing the three types of collocations will be given in Chap. 4.
2.2.4 Summary

This chapter has concerned itself with developing the definition of collocation and introducing its classifications. We have reviewed three different approaches to collocation: the psychological approach, which views collocation as psychological association in the mental lexicon, the Firthian approach, regarding collocation as words in syntagmatic relations in texts and the phraseological approach, aiming at demarcating collocation and distinguishing it from other types of word co-occurrences like free combinations and idioms. The phraseological approach is mainly followed in this study, since an empirical study on collocations in learner language requires a categorisation framework allowing them to be separated from idioms and free combinations. Based on the criteria of semantic transparency, specialised senses of words and commutability commonly adopted by phraseologists in demarcating collocation, a slightly refined definition of collocation is proposed: collocations are combinations of two or more words which are characterised by a restricted range of co-occurrence in at least one of their constituent words and by relative transparency in meaning. In addition, based on collocation classifications proposed by Hausmann (1989) and Benson et al. (2010), three types of lexical collocations: verb + noun, adjective + noun and noun + noun collocations will be examined in this study. Having defined what is meant by a collocation, the next chapter moves on to discuss previous studies on collocation learning by L2 learners.

References


Vocabulary Increase and Collocation Learning
A Corpus-Based Cross-sectional Study of Chinese Learners of English
Men, H.
2018, XIV, 206 p. 15 illus., Hardcover