VOCABULARY INCREASE AND COLLOCATION LEARNING: A CORPUS-BASED CROSS-SECTIONAL STUDY OF CHINESE EFL LEARNERS

HAIYAN MEN

A Thesis Submitted to Birmingham City University

in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

March 2015

The Faculty of Arts, Design and Media

Acknowledgements

I would like to extend my sincere gratitude to all those who have directed me into the field of English linguistics and provided me with so much guidance and encouragement throughout the period of my doctoral study.

First and foremost, my deepest gratitude goes to my supervisor, Professor Richard Ingham, whose rigorous academic training, thought-provoking guidance, and invaluable spur in various viewpoints have always been treasures for me in the writing process and beyond. I would also like to thank him for his endless patience in revising the draft chapters, and my conference abstracts. Apart from being my supervisor, he is also a wonderful mentor. I have also benefited much from him for his humour and unique ways of thinking. Thanks to his valuable instruction and warm support, I am able to pursue this doctoral study and accomplish it in time.

I am very much indebted to two other supervisors, Dr. Ursula Lutzky and Prof. Antoinette Renouf, for their valuable comments and assistance, and for their conscientious proofreading of the drafts.

I am particularly grateful to Prof. Yang Huizhong for not only his lively and stimulating instruction throughout my MA study, but also for his continuous monitoring and constant critical comments on my doctoral thesis. I also owe my debt to him for his constant care of both my academic study and life overseas.

At an institutional level my thanks are due to Shanghai Sanda University for providing financial support for this doctoral study, and for providing funds for me to attend conferences of this field.

Finally, my gratitude is reserved for my parents and friends. Special thanks are given to my fiancé and now husband, Mr. Wang Shuai for his unselfish help and unfailing support.

Though immense help has been received from those whose contributions have been acknowledged above, I take sole responsibility for any errors and failings I have not been able to correct.

Abstract

Collocation learning has long been recognised as a problematic domain for even high-level learners and acknowledged to lag far behind acquisition of other SLA aspects. This thesis explores the role of vocabulary growth in the learning of L2 collocations. It addresses the relationship between vocabulary increase and L2 collocation learning, aiming to identify whether increasing lexemes in a syn(onym)set (Fellbaum, 1998) is the main factor responsible. A corpus-based cross-sectional study was undertaken on Chinese EFL learners' written production of three types of collocations: verb + noun collocations (the main research target), adjective + noun and noun + noun collocations. Lexical verbs in verb + noun collocations were classified into synsets and analyses were performed on collocations within these synsets.

It finds a lag in L2 learners' verb + noun collocational knowledge with rising proficiency. This lag in collocational knowledge was found to be associated with an increase in lexical verbs learnt at higher levels. Collocation errors were seldom made where there was no increase in verb synsets. However, for synsets in which there was a verb increase, collocation errors involving new verbs were significantly more likely than errors with old verbs. The occurrence of collocation errors became increasingly limited to synsets with a verb increase as learners proceeded to more advanced levels. An alternative explanation was attempted to see if newly acquired nouns were also a factor responsible for the collocation lag. Results showed that in the majority of new nouns produced by higher levels of learners, collocations were target-like, and the percentages of new nouns in erroneous verb + noun collocations remained constant at both higher levels.

In contrast, it finds an improvement in L2 learners' knowledge of adjective + noun and noun + noun collocations. The thesis attempted to account for such differing performance from the perspective of vocabulary growth within synsets. The decreasing synonym density of verbs, adjectives and nouns may account for the relatively poorer performance on verb + noun collocations, and better performance on adjective + noun and noun + noun collocations. These findings are discussed with a view to a clearer understanding of the process of second language collocation learning, and to pedagogical implications.

Abbreviations

Corpora:	
BNC	British National Corpus
CLEC	Chinese Learner English Corpus
ICLE	International Corpus of Learner English
Dictionaries:	
Bilingual dictionaries:	
NCCED	*New Century Chinese-English Dictionary*
OALECD	*Oxford Advanced Learner's English-Chinese Dictionary* (7th Edition)
English dictionaries:	
BBI	*The BBI Combinatory Dictionary of English: Your Guide to Collocations and Grammar* (3rd Edition)
EVCA	*English Verb Classes and Alternations*
OCDSE	*Oxford Collocations Dictionary for Students of English* (2nd Edition)
ODSA	*Oxford Dictionary of Synonyms and Antonyms*
COBUILD	*Collins COBUILD English Dictionary* (2nd Edition)
Chinese dictionaries:	
CCD	*Contemporary Chinese Dictionary* (5th Edition)
Other abbreviations:	
AN	adjective + noun
DeLexVN	delexical verb + noun
EFL	English as a foreign language
ELT	English language teaching
ESL	English as a second language
FL	foreign language
FLT	foreign language teaching
L1	first language/mother tongue
L2	second language/foreign/target language
LexVN	lexical verb + noun
NN	noun + noun
NNSs	non-native speakers
NSs	native speakers
POS	Part of Speech
SL	second language
SLA	second language acquisition
VN	verb + noun

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.0 General background

The past decades have seen a dramatic increase in studies on collocations in second language acquisition. Several reasons account for such an increase. The first is a general one, associated with a growing body of research on collocation as a linguistic phenomenon per se in native speaker language. Initiated from the Firthian tradition of looking for word meanings through syntagmatic relations between words and a search for a lexical theory complementary to grammatical theory (Firth, 1957; Halliday, 1966; Sinclair, 1966), collocation has been a thriving and on-going field of linguistic enquiry (cf. Hoey, 2005; Moon, 1998; Renouf and Sinclair, 1991; Sinclair, 1991; 2004; Stubbs, 1996; 2001). Linguistic investigations into collocations have provided extensive evidence that native speaker texts are on the most part formulaic (e.g. Altenberg, 1998; Biber et al. 1999; Cowie, 1991; 1992; Howarth, 1998a). It follows that this phraseological tendency for meanings to be created through conventionalised word combinations underlying proficient performance requires L2 learners to have a good command of collocations.

The observation that collocation learning is of central importance for L2 learners' idiomatic control of that language constitutes another motive for a growing interest in L2 collocation learning. The importance of collocation knowledge for L2 learners has been long and widely recognised. Idiomaticity is identified as key to the attainment of native-like proficiency. As Pawley and Syder (1983: 191) acknowledged, "fluent and idiomatic control of a language rests to a considerable extent on knowledge of a body of 'sentence stems' which are 'institutionalised' or 'lexicalised'". Mastery of collocations not only facilitates idiomatic production, but also promotes the efficiency of language comprehension in general and the comprehension of lexical semantics of individual words. Failing to use native-like expressions may not only "divert the reader's attention from content to form" (Howarth, 1998a: 174), but also cause "an impression of brusqueness, disrespect or arrogance (Wray, 2002: 143), and "may sound rather bookish and pedantic to a native speaker" (Channell, 1994: 21).

In the meantime, arbitrarily restricted co-occurrence of word combinations abound in the English language. For example, *blond* is perfect for modifying *hair*, but not *door* or *dress* (Palmer, 1981: 76f), and we '*do* the cooking' but '*make* dinner' (Fox, 1998: 33). Learning to construct word combinations

that are customarily used by native speakers is one of the most difficult tasks for even the most proficient non-natives (Pawley and Syder, 1983). This phraseological deficiency in second language learners was realised as early as in the 1930s and has been extensively discussed ever since. Palmer (1933) noted that when forming such combinations (e.g. *to ask a question, to do a favour*), which are not-rule-governed word combinations, learners may produce expressions such as **to make a question, *to perform a favour*. A great deal of previous research has found that collocation learning constitutes a problematic domain for non-natives even at fairly high proficiency levels. Studies in this field are generally devoted to a description of L2 learners' difficulties with collocations. The overall picture that emerges from previous L2 collocation research is that apart from learners' better receptive knowledge of collocations (e.g. Biskup, 1990; Gyllstad, 2005; Marton, 1977), collocation production poses great problems. Overall, L2 learners' "building material is individual bricks rather than prefabricated sections" (Kjellmer, 1991:124). They are found to operate more on the 'open choice principle' than the 'idiom principle' and use fewer collocations compared with native-speaker counterparts. In addition to insufficient uses of collocation, overuse, underuse and misuse of certain collocations are frequently reported in learners' writings (e.g. Granger, 1998a; Howarth, 1996; Laufer and Waldman, 2011).

Apart from the difficulties L2 learners encounter in the production of collocations, phraseological knowledge is believed to lag behind grammar and lexis and constitutes the "last and most challenging hurdle in attaining near native-like fluency" (Spottl and McCarthy, 2004: 191). In comparison with learners' general vocabulary knowledge, knowledge of collocations is rather weak as Bahns and Eldaw (1993) found that collocation errors were more than twice than errors with lexical words. When collocation knowledge is compared among learners at different levels, it was reported that collocation performance did not improve as the advanced and the intermediate learners produced significantly more erroneous collocations than the basic learners (Laufer and Waldman, 2011). Similarly, in another more recent study exploring the collocational competence of two groups of Nigerian advanced speakers of English as a second language, Obukadeta (2014) discovered that the participants who had been living/studying in the UK for up to 15 years were less proficient in terms of their knowledge of collocations than the other group which had never lived or studied outside Nigeria. These studies indicate a collocation lag, which means that collocational knowledge does not develop alongside learners' general level of English proficiency.

Given the difficulties learners are confronted with and the lag in collocational knowledge, investigating collocations in an L2 is a continuing concern within the field of second language learning and teaching. However, research in this field is still in its early stage since there is not an overall theory accounting for how collocations are acquired by L2 learners (Gitsaki, 1999). Knowing how collocations are acquired and produced can provide valuable insight into how they are best taught. Although extensive research has been carried out on the learning and production of L2 collocations, much of the research up to now has been descriptive in nature. It remains unclear what factor(s) are associated with the lag in collocation knowledge.

Therefore, the importance of phraseological knowledge in both language production and comprehension, and its acquisition as a problematic territory for L2 learners provide sufficient justification for further research in this area. The aim of our study is to fill the gap by examining factors associated with collocation lag.

1.1 Aims of the thesis

The major task of second language collocation research is to discover what it means for L2 learners to acquire a collocation, how they learn it, and what problems they encounter in acquiring a collocation. Previous research has provided a comprehensive description of how L2 learners use collocations and what problems they encounter in using them. Yet little is known with regard to factors responsible for the stagnant development of collocation knowledge. Hence, this study is intended to investigate factors that are associated with this collocation lag. By examining the factor(s) that are responsible for the lag in collocation knowledge, we can better understand the process of collocation learning. Furthermore, more knowledge on how collocation is acquired can further shed light on how collocations are best learnt and taught.

As regards which factor(s) may be responsible for collocation lag, we hypothesise that vocabulary growth is an inhibiting force in collocation learning. So a general question is asked: is vocabulary growth an inhibiting factor in the learning of collocations by L2 learners? This question needs to be further divided into detailed questions by taking particular types of collocations as examples. For this purpose, one most important and frequent type of collocation – verb + noun (henceforth: VN) collocations is first selected.

The relationship between verb increase and the production of VN collocations is examined among verb + noun collocations produced by different levels of learners. At a macro level, verb increase is measured in terms of the development from delexical to lexical verbs. Verbs are divided into two categories according to the semantic contents they take: delexical verbs (*do, make, take, have, give and get*) and lexical verbs (*acquire, fulfil, perform*, etc.). Accordingly, VN collocations are divided into delexical verb + noun (DeLexVN) collocations and lexical verb + noun (LexVN) collocations. At a micro level, the growth of verbs is measured in synonym sets (Fellbaum, 1998) in specific VN collocations. The following developmental patterns with regard to the increase of verbs from delexical to lexical verbs are hypothesised, such that lower levels of L2 learners make more errors with delexical verbs and higher levels make more errors with lexical verbs in VN collocations.

The hypothesis on the developmental patterns is closely linked with the general hypothesis that verb increase is a hindrance in collocation acquisition. More precisely, at lower stages of L2 development, due to their limited mastery of verbs, learners resort to delexical verbs to collocate with a noun instead of a specific lexical verb. As their verb vocabulary grows, they have more access to lexical verbs and tend to make more collocation errors with lexical verbs, because the increase in synonymous verbs allows more chances of incorrect verb choices. The increase in lexical verbs and the subsequent occurrences of errors with lexical verbs suggest that vocabulary growth impedes collocation acquisition. To test whether the growth of verb vocabulary constitutes an inhibiting force in collocation learning, the relationship between vocabulary growth and collocation development has to be viewed locally in specific VN collocations, through locating the semantic domains of verbs in collocations where there is an increase in verbs and examining whether the increase in these verbs subsequently leads to collocation errors.

Based on the two hypotheses, the following research questions will be addressed in our study:

1. What developmental patterns appear in the verb + noun collocations produced by L2 learners, in terms of delexical verb and lexical verb + noun collocations?

a. Is there a tendency towards increasing use of lexical verb + noun collocations with rising proficiency?

b. Is there a tendency towards increasing errors with lexical verb + noun collocations and decreasing errors with delexical verb + noun collocations with rising proficiency?

2. Within specific semantic domains of the verbs in verb + noun collocations used by all levels of learners, is there a tendency for these verbs, as they increasingly occur at the higher levels, to be associated with collocation errors?

Research questions 1b and 2 are interrelated as they bear the relation of the whole and a part. They are both concerned with the increase of lexical verbs and the production of verb + noun collocations at the higher levels. Research question 1b addresses the relationship between the overall increase of lexical verbs at the higher levels on the whole and the increasing/decreasing trend of verb + noun collocation errors associated with lexical verbs; the scope of research concerning verb increase and collocation errors is further narrowed down in research question 2, which is aimed at a concrete investigation of the increase of lexical verbs within particular semantic domains. Through confining the verb increase into semantic domains, our study sets out to examine if verb increase in a semantic domain is a factor associated with the lag in verb + noun collocational knowledge. The particular focus on verb increase in semantic domains in research question 2 is built on the belief that learners may be confused with semantically related words (e.g. *acquire* and *obtain*) rather than words falling in different semantic domains (e.g. *acquire* and *change*) in producing verb + noun collocations.

Furthermore, two other common types of collocations – adjective + noun and noun + noun collocations – will be examined, so as to compare the results with verb + noun collocations. Similar research questions on verb + noun collocations will be addressed with regard to adjective + noun and noun + noun collocations. Specifically, we focus on the following questions:

3. Are adjective + noun and noun + noun collocations produced by Chinese L2 learners at the same accuracy level as verb + noun collocations? If not, what patterns do they follow?

4. Within specific semantic domains of the adjectives in adjective + noun collocations and nouns in noun + noun collocations used by all levels of learners, is there a tendency for these adjectives/nouns, as they increasingly occur at the higher levels, to be associated with collocation errors?

As another field of enquiry, this thesis will investigate the role of L1 in the production of congruent and non-congruent L2 collocations. Congruent collocations refer to collocations whose word elements in one language have direct word-for-word translational equivalence in another language; if word elements in one collocation do not share direct word-for-word translational equivalence between two languages, then it is considered as a non-congruent collocation (Nesselhauf, 2005; Wolter

and Gyllstad, 2011). Previous studies show that congruent collocations are much easier than non-congruent ones (e.g. Bahns, 1993; Nesselhauf, 2005), and non-congruent collocations once acquired, are processed independently of the learners' L1 (Yamashita and Jiang, 2010; Wolter and Gyllstad, 2011). In light of these findings, we set out to examine Chinese L2 learners' performance on congruent and non-congruent collocations, with the aim to test whether congruent collocations are easier than non-congruent ones for them, and whether non-congruent collocations once acquired, are less prone to errors.

The present research is therefore an empirical study of the phraseological performance (verb + noun collocations in particular) of Chinese learners of English across different proficiency levels. This investigation will provide insight into how collocations are acquired by EFL learners, a question which has not yet been addressed but is of central importance for a comprehensive understanding of second language collocation learning. The study will shed light on the key question that researchers in SLA attempt to answer, expressed as "What is acquired? What is not acquired? Why so?" by Gries (2008: 407).

The objective of the research project was an empirical study of the use of collocations by Chinese L2 learners. However, in order to describe non-native phraseological competence, it is first necessary to establish native-speaker norms in this regard. The native speaker and non-native speaker dichotomy is contentious. Davies (2003: 1) introduces the common-sense concept of native speakers, referring to "people who have a special control over a language, insider knowledge about 'their' language" and are the "models we appeal to for the 'truth' about the language". However, as Davies (2003) acknowledges, he can more easily define what a non-native speaker is than a native speaker, and he even argues that the native speaker is a myth. Those who have two native speaking parents, both preferably monolingual, and are raised in a native speaking community, can still not be definitely defined as native speakers of that language, since other social factors like mobility and the rise of new Englishes are at play (Davies, 2003). As English is becoming a lingua franca, and an increasing number of proficient academics whose first language is not English enter English academia (Hyland, 2006), it is even harder to define what a native speaker is. The theoretical aspects of the native speaker construct will not be addressed in this study.

Nevertheless, for the investigation and description of learner interlanguage, language learning goals in terms of native-speaker norms, need to be set. Two widely-used English collocation

dictionaries and the British National Corpus, a collection of the texts in British English, were taken as a kind of target norm for L2 English learners. Language forms produced by L2 learners that conform to the norm were regarded as well-formed, and those that deviate from the norm were viewed as erroneous.

1.2 The shape of the thesis

The thesis is divided into ten chapters. Chapters 2 and 3 discuss previous theoretical and empirical studies on (L2) collocations. Chapter 2 highlights the importance of collocation and clarifies the notion of collocation. The significance of collocations is discussed in terms of their prevalence in native-speaker texts and their importance for a fluent and idiomatic control of English for L2 learners. Then the notion of collocation is examined on the basis of previous different approaches, and a definition and classification applied in our study are presented. Chapter 3 reviews previous collocation studies in second language learner English. In this chapter, the methodologies commonly adopted in L2 collocation studies are firstly addressed, with a view to introducing the methodology that has been more and more widely used in the analysis of collocations in learner corpora; then major findings of previous L2 collocation studies are discussed. Chapter 4 presents the detailed design of the present cross-sectional study of Chinese EFL learners' collocation performance. The learner corpus chosen for such an investigation, the types of collocations targeted, the sources of reference in extracting these collocations, and the procedures for collocation extraction and analyses are introduced. Chapters 5, 6, 7, 8 and 9 contain a detailed analysis of the data. In Chapter 5, the overall picture of Chinese L2 learners' performance in verb + noun collocations is depicted, with the main focus on the developmental patterns of collocation production from delexical verb + noun to lexical verb + noun collocations. Chapter 6 is devoted to an investigation of the relationship between verb increase in specific synonym sets and collocation uses associated with verbs in these synsets. Moreover, an alternative explanation, i.e. the learning of new nouns in collocation production, is made in order to see whether the acquisition of new nouns is responsible for a lag in collocation. Chapter 7 goes on to explore learners' performance on two other important and frequent types of collocations, i.e. adjective + noun and noun + noun collocations. It aims at a corroboration of findings from verb + noun collocations. Chapter 8 presents and compares learners' performance on verb + noun, adjective + noun and noun + noun collocations. In Chapter 9,

cross-linguistic influence in the production and learning of L2 collocations is investigated. It addresses the role of L1 in learners' performance on congruent and non-congurent collocations, and the role of L1 in different word-class collocations such as verb + noun and adjective + noun collocations. Chapter 10, finally, concludes the whole thesis by summarising the findings of this study, discussing both theoretical and pedagogical implications for effective collocation learning, acknowledging the limitations of the present study and putting forward proposals for future research.

Chapter 2: Literature review (1): The notion of collocation

2.0 Introduction

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 (Section 2.1). The second section (Section 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 DeCarrico, 1992; Pawley and Syder, 1983; Renouf and Sinclair, 1991; Sinclair, 1991; Stubbs, 2001; inter alia).

A considerable proportion of prefabs have 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.[1] 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.[2] 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 inclusion of any continuous string of words

[1] Kjellmer'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. Section 2.2.2).

[2] The 1000-word sample was compiled from a 10,000-headword data-base, which recorded the most frequent content words in the Cobuild (1995) data-base. So there were no function words beginning with an *f* (e.g. *for*).

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 language 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,[3] 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.

[3] 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.

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, 1989). In this section, its significance is briefly summarised from two perspectives: for native-like production and for efficient comprehension.

a. 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 Chapter 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)

b. 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. 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 mis-hear 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. Firstly, 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 non-native speakers 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 our 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 Section 2.1, e.g. *collocations, fixed expressions, idioms, prefabs, complex lexical items, multiword units,* etc.),[4] a clarification of which of these aspects of formulaic language forms the object of our study is in order. The present study will focus on the most common manifestations of formulaic language – collocation.[5] 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

[4] Kjellmer (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, multi-word lexical units (MLU), multiple-word units, patterned speech, patterns, phrases, prefabricated speech, ready-made utterances, recurrent combinations, stock phrases, 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.

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)).

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]

2.2.1.1 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

[6] This 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.

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 Section 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* 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 a non-native speaker are asked to express *strong coffee*, the collocate *strong* can be easily associated with *coffee* by native speakers, but non-native speakers might use other collocates like *powerful* rather than *strong*. It thus requires consciousness and efforts for non-native speakers 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

[7] According to Aitchison (2003: 86), the commonest type of response to stimulus words is co-ordination, e.g. *salt* with *pepper*, *butterfly* with *moth*.

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 sub-sections: 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.

a. 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 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*

[8] Words in boldface are quoted in their original forms.

suitable for drinking).[9] 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.[10] 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 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

[9] Explanations of *drink* are quoted from WordNet.

[10] Lemma 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).

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.

b. 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; cf. Greenbaum, 1974; Hoey, 2005; 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:

Table 2-1 Classifications of English word combinations[11]

Vinogradov (1947)	Phraseological fusion	Phraseological unity	Phraseological combination	
Amosova (1963)	Idiom	Idiom (not differentiated)	Phraseme, or Phraseoloid	
Aisenstadt (1979)	Idiom		Restricted collocation	Free word-combination
Cowie (1981)	Pure idiom	Figurative idiom	Restricted collocation	Open collocation
Nattinger and DeCarrico (1992)	Idiom		Collocation	Free combination
Howarth (1996)	Pure idiom	Figurative idiom	Restricted collocation	Free collocation
Howarth (1998a; b)	Pure idiom	Figurative idiom	Restricted collocation	Free combination

[11] 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.

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 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.

a. 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. Aisensdadt, 1979; Cowie, 1981; 1992; 1998; Howarth, 1996; 1998a).

b. 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, 1998; 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.

c. 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) *shrug one's shoulders*
shrug something off
shrug something away

shrug one's shoulders
square one's shoulders
hunch one's shoulders

(5) *make a decision*
take a decision
have a look
give a look
take a look

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:

Table 2-2 Howarth's categorisation of collocations into five levels of restrictedness

	Verb	Noun	Examples
L1	Some restriction	Free substitution	*adopt/accept/agree to a proposal/suggestion, etc.*
L2	Some substitution	Some substitution	*introduce/table/bring forward a bill/an amendment*
L3	Some substitution	Complete restriction	*pay/take heed*
L4	Complete restriction	Some substitution	*give the appearance/impression*
L5	Complete restriction	Complete restriction	*curry favor*

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 the notion of commutability, the number of synonyms is also subject to the conceivability of a human mind. With the examples in L3 for an example, the combination *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.[12] So the judging on the number

[12] The verb collocates of *impression* – *make* and *leave* – are listed with reference to a collocation dictionary – *Oxford Collocations Dictionary for Students of English* (2nd Edition).

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 differentiating 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 collocations –four words each side – is insufficient to account for certain common collocations (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 collocations of statistical value, cannot incorporate all the collocations of phraseological value (like *collecting stamps* in the above examples). Moreover, the Firthian tradition 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 co-occurrences (e.g. free word combinations and idioms). Measured frequency of co-occuring words is not a significant measure of collocational restriction (Cowie, 1998: 226, Greenbaum, 1974: 83); 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 non-natives, 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 non-native 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 below).

Table 2-3 Previous definitions of collocations and criteria adopted

Author	Definitions	Criteria
Aisenstadt (1979: 71)	"combinations of two or more words used in one of their regular, non-idiomatic meanings, following certain structural patterns, and restricted in their commutability not only by grammatical and semantic valency (like the components of so-called free word-combinations), but also by usage"	semantic transparency; commutability
Aisenstadt (1981: 54)	"a type of word combination consisting of two or more words, unidiomatic in meaning, following certain structural patterns, restricted in commutability not only by semantics, but also by usage, belonging to the sphere of collocations"	semantic transparency; commutability
Van Roey (1990: 46)	"the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its "synonyms" because of constraints which are not on the level of syntax or conceptual meaning but on that of usage"	commutability
Howarth (1996: 47)	"combinations in which one component is used in its literal meaning, while the other is used in a specialised sense. The specialised meaning of one element can be figurative, delexical or in some way technical and is an important determinant of limited collocability at the other. These combinations are, however, fully motivated"	specialised sense of one element; commutability (collocability); semantic transparency (semantically motivated)
Nesselhauf (2005:25)	"combinations in which at least one element has a non-literal meaning (and at least one a literal one) and in which commutability is arbitrarily restricted, but some commutability is possible"	specialised sense of one element; commutability
Laufer and Waldman (2011: 648)	"habitually occurring lexical combinations that are characterised by restricted co-occurrence of elements and relative transparency in meaning"	semantic transparency; commutability

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 a collocational range, referring to the number of co-occurring words a word can take (see also Cowie, 1981; 1998; Granger, 1998a; 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 collocations, namely, semantic transparency and the range of co-occuring words. Collocations are then defined as 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.

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 of or both elements have a restricted range of co-occurring words, and also are transparent in meaning, are categorised as collocations; 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)

Table 2-4 A framework for demarcating collocations

	Verb	Noun	Examples
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)

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 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 lexico-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 Section 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; 1998; Nesselhauf, 2005); rather it is broadly based on the word classes of its constituents, since the study aims at investigating non-natives' 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 (*criticize 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).

Table 2-5 Classifications of lexical collocations by Benson et al. (2010)

Types	Examples
verb + noun/pronoun (or prepositional phrase); with the verb denoting *creation* and/or *activation*	*come to an agreement, make an impression, compose music*
verb + noun; with the verb denoting *eradication* and/or *nullification*	*reject an appeal, lift a blockade, break a code*
adjective + noun	*strong tea, warm regards, reckless abandon*
noun + verb	*adjectives modify, alarms go off, bees buzz*
noun + of +noun	*a herd of buffalo, a pack of dogs, a bouquet of flowers*
adverb + adjective	*deeply absorbed, strictly accurate, sound asleep*
verb + adverb	*affect deeply, amuse thoroughly, argue heatedly*

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.[13] As part of the aim of this research is to investigate the growth of vocabulary produced by non-natives, 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.

[13] The reasons for focusing the three types of collocations will be given in Chapter 4.

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.

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 words 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.

Chapter 3: Literature review (2): Collocation studies in second language learner English

3.0 Introduction

Similar to collocation studies discussed in the previous chapter, the past decades have also seen a large volume of studies on collocation learning in an L2. The purpose of this chapter is to review previous collocation studies in learner English to date. It begins by addressing the methodologies commonly adopted in L2 collocation studies, with a view to introducing the methodology employed in our study; Section 3.2 presents major findings of previous research in this field.

A review of the studies on L2 collocation learning is inevitably combined with studies on other prefabricated forms of language (cf. Granger's (1998a) study on collocations and formulae), since collocation as a linguistic phenomenon belongs to a larger umbrella term – 'formulaic language' – and in practice collocations are not always carefully delimited from other types of word combinations (Nesselhauf, 2005: 3). Therefore, in this literature survey, research on L2 learners' knowledge of restricted combinations of words (e.g. formulae, formulaic sequences, routines, etc.) is briefly reviewed, with the main focus on studies of collocations produced by L2 learners.

3.1 Methodologies adopted in L2 collocation studies

In the investigation of L2 learners' collocation knowledge, studies are generally based on two categories of data: elicitation data and production data (Fan, 2009: 112; Nesselhauf, 2005: 4). Production data in this sense stands in contrast to the elicited data type, referring exclusively to naturally occurring data or spontaneous data (Penke and Rosenbach, 2007: 10). Yet confusion may arise out of the overlapping between the two types, since elicitation data encompass the type of production data elicited from L2 learners in translation tasks (compared with learners' introspective data elicited in intuition judgment tasks); at the same time production data also include production data of the elicited type (cf. Ellis, 1994: 670; Granger, 1998b: 4). In order to avoid confusion, Penke and Rosenbach's distinction

between elicited and spontaneous data is adopted and previous L2 collocation studies are considered as either elicitation data- or spontaneous data-based.

3.1.1 Elicitation data-based collocation studies

Elicitation tasks designed to assess second language learners' phraseological production/comprehension include translation tasks (e.g. Bahns and Eldaw, 1993; Biskup, 1990; 1992; Farghal and Obeidat, 1995, Hasselgren, 1994; Irujo, 1993; Marton, 1977), blank filling tasks (Farghal and Obeidat, 1995; Hoffman and Lehman, 2000; Scarcella, 1979; Zhang, 1993), cloze (Al-Zahrani, 1998; Bahns and Eldaw, 1993; Schmitt et al., 2004c) and word-combination tests (Bonk, 2001; Channell, 1981; Granger, 1998a; Gyllstad, 2005; Siyanova and Schmitt, 2008; Wolter and Gyllstad, 2011; Yamashita and Jiang, 2010).

A great advantage of this approach is that researchers can directly observe and analyse L2 learners' collocation production/comprehension of a set of pre-selected collocations. For instance, in eliciting L2 learners' production data of a particular type, Bahns and Eldaw (1993) chose 15 English verb – noun collocations whose German equivalents would be likely to be translated into the respective English collocations and designed both a translation test and a cloze test to measure German L2 learners' active knowledge of these collocations. Farghal and Obeidat (1995) administered both blank filling and translation tests to two groups of Arabic learners of English at two proficiency levels, aiming to investigate their productive knowledge of 22 common English collocations on topics such as food, clothes, colour and weather. Irujo (1993) confined the study of the production of English idioms in translation tests by 12 bilingual native speakers of Spanish to three types of idioms according to their similarity between Spanish and English: exact equivalents, similar ones and totally different idioms. Hoffman and Lehmann (2000) concentrated on 55 adjective-noun and noun-noun collocations strongly associated in the BNC and designed a gap filling task to investigate native and non-native speakers' familiarity with these collocations. With the collocations pre-selected, researchers can directly observe learners' performance on them and eliminate the risk of obtaining unnecessary data.

Moreover, in testing L2 learners' receptive knowledge through eliciting their introspective data, elicitation techniques such as acceptability judgments possess a distinct advantage which cannot be outweighed by other methods (e.g. spontaneous data). Word-combination tests designed to tap into L2

learners' intuition about possible collocations are usually in this form. In these tasks, learners were asked to choose possible collocates of pre-determined words (Channell, 1981; Granger, 1998a; Gyllstad, 2005) or judge the acceptability of certain collocations (Siyanova and Schmitt, 2008). The two kinds of recognition tests correspond to the test formats for measuring learners' receptive knowledge of English verb + noun phrase (NP) collocations by Gyllstad (2005), namely, COLLMATCH and COLLEX. In COLLMATCH, learners were presented with a number of grids consisting of verbs and NP objects, and then asked to indicate which verbs can combine with which nouns. The other test (COLLEX) involved testees to choose the correct collocation among two lexical combinations: one correct and one pseudo-collocation (e.g. *pay a visit* and *do a visit*). Unlike translation, blank filling and cloze tasks, whose main advantage is to test the productive collocational knowledge of L2 learners, these elicitation tasks afford a direct assessment of L2 learners' receptive collocational knowledge.

A tight control over what is elicited from L2 learners enables direct comparisons of collocational evidence based on unified criteria. Comparisons can be made between different elicitation tasks, between collocation performance of different participants and between L2 learners' receptive and productive collocation knowledge. For example, different elicitation tasks administered to the same group of learners can reveal different strategies L2 learners adopt in producing collocations. Bahns and Eldaw (1993) compared the collocation production of the same levels of learners in two tasks and reported that subjects did not perform significantly better in a translation task than a cloze task, even though they were able to paraphrase the target collocations in translation sentences but not in cloze sentences. Findings showed that the difference between the number of correctly translated collocations and the number of correct collocates in the cloze task did not reach statistical significance (ibid: 106). The explanation proposed by Bahns and Eldaw (1993) was that collocations were not easy to be paraphrased.

By controlling the set of collocations to be elicited, comparisons can also be made between performances of different groups of participants, i.e. between learners at different proficiency levels, learners of different L1 backgrounds and learners in contrast with native speakers. Different collocation performances and strategies in producing the same collocations were observed by Farghal and Obeidat (1995) in their two groups of subjects: junior and senior English major L2 learners (Group A) and language teachers of English (non-native speakers) (Group B). Both groups were found to be seriously deficient in collocations (ibid: 315). With regard to different strategies adopted by the two groups, they

noted that Group B participants resorted to paraphrasing as a strategy (25.1%) in the translation task more than Group A (3.8%) in the blank filling test. The two groups are not comparable, however, since in the translation task administered to Group B, more freedom of paraphrasing is allowed than in the blank filling test given by Group A, so it could have been more effective if the two groups had been given similar elicitation tests. Additionally, though there was a higher percentage of paraphrasing strategy adopted by Group B, the target translations of collocations were not found to be satisfactory. Examples of such paraphrases of collocations are *food little fat* for *light food*, *does not change* for *fast color* (ibid: 325). The finding of this unnaturalness in the produced collocations caused by paraphrasing in translation tests is consistent with Bahns and Eldaw's (1993) finding that some collocations cannot be readily paraphrased.

Besides offering a comparison between learners of different proficiencies, collocation performances of learners with different L1 backgrounds can be compared in elicitation based studies. In a translation task, Biskup (1992) reported that the German learners were risk-takers and produced more variant collocations whereas Polish learners produced more restricted collocations. Meanwhile, in terms of the comparison between learners and native speakers, L2 learners' receptive knowledge was found to be poorer than native speakers as in the judgment tasks conducted by Siyanova and Schmitt (2008) and Granger (1998a).

Elicitation tasks also enable a comparison between L2 learners' receptive and productive knowledge over predefined test materials. Biskup (1990) discovered a striking difference in Polish learners' performances on L2 - L1 and L1 - L2 translation tasks: their answers were 100% correct in the former but they had great difficulty in the latter. Similar results were obtained by Marton (1977) in a pre-treatment and post treatment Polish-English translation test. The studies conducted by both Biskup and Marton suggest that learners' productive knowledge of collocation lags far behind their receptive knowledge. On the one hand, their apparent ease with collocation comprehension rests on the high degree of semantic transparency in collocations; on the other, translation from L2 to L1 is always easier than the other way round because L2 words are initially associated to L1 translation equivalent to access meaning whereas translation from L1 to L2 words requires concept mediation, which leads to a stronger lexical association from L2 to L1 than that from L1 to L2 (Kroll and Stewart, 1994).

It is evident from the above discussion that elicitation data-based studies on L2 learners' collocation performance possess several advantages difficult to obtain with other types of data.

Variables affecting subjects' production can be clearly and systematically controlled with the result that certain collocations happening "to occur very rarely or not at all unless specifically elicited" (Yip, 1995: 9) are elicited by the researchers. Meanwhile, by targeting the same set of collocations in elicitation tasks to different learners, comparisons can be performed between variables researchers aim to investigate. However, one of the limitations of elicitation data-based studies lies in generalisability of the research findings to the broader language proficiency of the participants. As Bahns and Eldaw (1993: 108) acknowledged, 15 collocations tested in the translation tasks were too small a sample from which a hypothesis is generalised. Generalisability is also affected by the artificiality of an experimental situation that "may lead learners to produce language which differs widely from the type of language they would use naturally" (Granger, 1998b: 5). Take for example the following sentences in the blank-filling and cloze tests designed by Schmitt et al. (2004) for tapping into learners' knowledge of formulaic sequences:

(1) With reg_ to giving directions, you must know phrases like 'Turn right at the corner'. (*concerning this certain thing*) (answer: *regard*)

(2) The economy is sure to improve ___c__
a. in the long period
b. over a long time
c. in the long run
d. over a long space
e. I DON'T KNOW

Learners may produce the right kind of answer but avoid using the target expression (e.g. *with regard to*) in real language production. In sentence (1), they may well use other expressions on which they are more confident, such as *about*. In (2), they may choose C but produce *over a long time* in their writing, given the close semantic similarity of these expressions. So there may be a gap between learners' elicited performance and their natural production. There are also cases where learners "tend to evaluate non-standard forms as bad and distance themselves from them when asked explicitly, while still using such forms actively" (Penke and Rosenbach, 2007: 12). In other words, L2 learners may recognise the co-occurrence of the two or more words as an appropriate collocation, but may neglect its use in actual production. That is because "language use involves choices, and each time we construct an utterance we have to select from our available resources in such a way as to convey what we want to say" (Ellis, 1987: 5). The discrepancy between learners' linguistic knowledge and communicative use is acknowledged by Widdowson (1979: 197), who pointed out that in the process of converting the former to the latter errors take place. For instance, the past tense forms may be 'known' but not 'used' (Willis, 2010: 10). As for collocations, the adjectival collocates of amplifiers

such as *highly* were reported to be better known than actually produced (Granger, 1998a). In Granger's study, learners marked more adjectives with the adverb *highly* in an elicitation test than they actually produced in writings, a phenomenon which was "somewhat paradoxical when considered in the light of evidence that learners underuse *highly* in their writing" (ibid: 153). The discrepancy between receptive and productive collocation knowledge is also endorsed by Gyllstad (2005: 27), who despite finding that the most advanced Swedish university learners performed almost as well as native speakers on receptive tasks, recognised that the two groups would behave differently through a test of productive knowledge.

Thus, it is insufficient to use a few elicitation tasks such as questionnaires in the hope of contributing to a comprehensive and systematic linguistic description (Pu, 2010). The criticisms levelled against elicited data–based studies are addressed by the analysis of learner language on the basis of natural language use data, or third-person observed data (in the terminology of Stubbs (2001) and Widdowson (2000)). Basing the description of L2 learners' collocation acquisition on spontaneous data is gaining acceptance as it contributes to a fuller picture of non-natives' collocation performance and learning.

3.1.2 Spontaneous data-based collocation studies

Unlike L2 collocation studies based on elicitation techniques, in which learners are asked to produce or recognise a specific set of collocations pre-selected for investigation, spontaneous data-based L2 collocation studies focus on L2 learners' natural production of collocations either in conversational or written texts. Whether a set of collocations is pre-selected and then elicited from learners or not is a radical difference between the two data types. Spontaneous data collected in second language acquisition in most cases involves elicitation with a very limited degree of control in essays (where only the topic is given) or oral interviews (where the interviewer introduces one or a few topics) (Nesselhauf, 2005: 40). Such control affects the topics or the time limit: otherwise the oral or written production of the learners is spontaneous and free (cf. Granger, 2002: 8). Based on observations of learners' authentic use of English, much concrete evidence on learners' phraseological performance has been accumulated, especially on their written performance (Ädel and Erman, 2012; Durrant and Schmitt, 2009; Fan, 2009; Granger, 1998a; Howarth, 1996; 1998a; b; Hsu, 2007; Kaszubski, 2000;

Laufer and Waldman, 2011; Li and Schmitt, 2009; Lorenz, 1999; Martelli, 2006; Men, 2010; Nesselhauf, 2005; Siyanova and Schmitt, 2008; Yorio, 1989).[14]

One advantage of spontaneous data–based L2 collocation studies is that learners' free production performance can be examined in large quantities, which elicitation data does not provide. These large quantities of data are compiled into a corpus, defined by Sinclair (1991: 171) as "a collection of naturally-occurring language text, chosen to characterise a state or variety of a language". In the context of SLA, the data collected are freely produced learner language, so the corpus in use is a learner corpus – defined by Granger (2002: 7) as: "electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance".

The availability and sophistication of computers greatly facilitates learner corpus research. Studies of L2 collocations based on spontaneous data have almost exclusively utilised learner corpora, whether self-assembled corpora (usually of smaller size and more specific to research requirements) or public ones (of larger size and fulfiling more varied research purposes). Studies using self-assembled corpora include those by Yorio (1989), Howarth (1996; 1998a; b), Hsu (2007), Li and Schmitt (2009). The merits of these learner corpora are that they are less time- and energy- consuming to compile than public learner corpora, and they enable researchers to focus on a particular type of learners. The disadvantages of self-assembled data include questions over the generalisability of their results based on a limited number of participants and their replicability given the unavailability of these corpora to the public.

Another trend in investigating L2 collocation uses is to carry out studies based on public learner corpora (e.g. Ädel and Erman, 2012; Granger, 1998a; Laufer and Waldman, 2011; Lorenz, 1999; Martelli, 2006; Nesselhauf, 2003; 2005; Siyanova and Schmitt, 2008; Zhang and Gao, 2006).[15] Many learner corpus-based collocation studies use the sub-corpora of the well-known learner corpus, the International Corpus of Learner English (ICLE). For example, Granger (1998a) investigated the uses of collocations and formulae by French learners of English in the French sub-corpus of ICLE; the

[14] Studies of collocations in L2 speech are very rare, although there are studies of other forms of phraseological performance in speech (for example, Aijmer, 2009; Crossley and Salsbury, 2011; De Cock, 2011; De Cock et al., 1998; Foster, 2001). Lexical bundles, formulaic sequences and routines are their main study foci.

[15] The corpora in the studies of Laufer and Waldman (2011) contain 300, 000 words, composed of argumentative and descriptive essays by native speakers of Hebrew of different levels. The corpora compiled by Lorenz (1999) record German learners' argumentative writing totaling 200, 000 words. Though it is not known whether these two corpora are publicly available, the large quantities of data from multiple learners manifest great advantages compared with a corpus of several essays of a limited number of L2 learners.

study of verb-noun collocations produced by German-speaking learners was undertaken by Nesselhauf (2005) by using the German sub-corpus of ICLE, and Siyanova and Schmitt (2008) focused on the production of adjective–noun collocations by Russian learners using the Russian sub-corpus of ICLE. Other non-ICLE-based collocation studies include those by Zhang and Gao (2006), Laufer and Waldman (2011) and Ädel and Erman (2012). The large amount of learner data recorded in these large-scale learner corpora allows for a description of L2 collocation production that is as comprehensive as possible. What is more, utilising public corpora enables studies of the same corpora to be easily comparable and replicable (Penke and Rosenbach, 2007: 11).

With the development of corpus analysis techniques, another advantage of learner corpus based collocation studies is that data can be (semi-) automatically extracted and processed. As in Howarth's (1996; 1998a; b) study, the machine-readable corpus is useful for a rapid check of the original context for pre-extracted collocations. Nesselhauf (2005) performed a similar automatic analysis to check whether all instances of verbs that were found to be restricted had actually been spotted by the manual analysis. Granger (1998a) used text-retrieval software (TACT) to automatically retrieve all the words ending in *ly* from the NS and NNS corpora and then manually sorted them according to pre-defined semantic and syntactic criteria. In the process of extraction of verb + noun collocations, Laufer and Waldman (2011) created concordances of the 220 most frequent pre-generated nouns in a NS corpus, and manually identified the verbs to go with these nouns in the NNS corpus. Yet whether these 220 most frequent nouns are also frequent in the NNS corpus is in question. So a full picture of learners' collocation production would be neglected by selecting a set of predetermined words for retrieval. To get a fairly comprehensive picture of L2 collocation uses, collocations in this study are not confined to certain predefined node words. Furthermore, differing from some of the collocation studies where collocations are manually identified (Howarth, 1996; 1998a; b; Li and Schmitt, 2009; Nesselhauf, 2005; Siyanova and Schmitt, 2008), collocations will be semi-automatically extracted in our study with the aid of text retrieval software.

Spontaneous data-based collocation studies also have certain disadvantages, insofar as only productive knowledge rather than receptive knowledge can be investigated; infrequent features are hard to examine even in fairly large corpora since they occur rarely unless specifically elicited. In other words, only the performance of learners is investigated but not their competence (Granger, 1998b; Nesselhauf, 2005; Yip, 1995). However, learners' performance can be taken as indicating their

phraseological competence, since as acknowledged by Ellis (1994: 13): “learners’ mental knowledge is not open to direct inspection; it can only be inferred by examining samples of their performance”.

To summarise the methods used in investigation of L2 collocation knowledge, previous research commonly explores two types of data: elicited and spontaneous data, each possessing distinctive advantages in answering particular research questions. Though elicitation-based studies enable direct observation and analysis of L2 learners’ collocation production/comprehension of a set of pre-selected collocations, criticisms are levelled in terms of their naturalness and generalisability. Spontaneous data-based studies examine L2 learners’ natural production of collocations through using large learner corpora. On the one hand, the use of a large learner corpus enables a comprehensive description of real language use; on the other, the development of computer software greatly enhances efficiency in retrieving and analysing collocations in learner corpora. The point of departure of this investigation is thus learner corpus-based, using the publicly available Chinese Learner English Corpus (CLEC). It aims to examine Chinese English learners’ productive knowledge rather than receptive knowledge. It also sets out to (semi-)automatically extract all the collocations within a syntactic category (e.g. verb + noun collocations) instead of focusing only on a number of pre-determined collocations. In this way some disadvantages of spontaneous data-based studies can be overcome.

3.2 Previous findings from L2 collocation research

Based on the two data types discussed in the precious section, collocation in an L2 has been extensively studied. Previous studies are varied in nature, as seen from a wide range of differing task types, learner types and collocation types. Their heterogeneity makes it difficult to compare the results of past studies (Paquot and Granger, 2012: 131). However, the overall picture that emerges through previous L2 collocation research is that collocation production constitutes a particular problematic domain in SLA, even for learners at an advanced level, compared with their better receptive collocation knowledge (e.g. Biskup, 1990; Gyllstad, 2005; Marton, 1977). Most importantly, L2 collocation studies indicate a collocation lag, where collocation knowledge lags far behind the development of syntax and lexis. This deficiency in collocation learning was recognised as early as the 1930s by Palmer (1933) and is strongly upheld in later studies. In this section, major findings of prior studies are presented.

3.2.1 Forms of collocation deficiency: overuse, underuse and misuse[16]

3.2.1.1 Overuse and underuse

Explorations of L2 learners' collocational performance through quantitative comparisons of native and non-native uses (a methodological approach called Contrastive Interlanguage Analysis by Granger (2002: 11f)) reveal that L2 learners operate more on the 'open choice principle' than the 'idiom principle', using fewer collocations than their native-speaker counterparts. In addition to insufficient collocation uses, they are found to overuse and underuse certain collocations (e.g. Ädel and Erman, 2012; Cobb, 2003; De Cock et al., 1998; Durrant and Schmitt, 2009; Foster, 2001; Granger, 1998a; Howarth, 1996; 1998a; b; Laufer and Waldman, 2011; Lorenz, 1999; Kaszubski, 2000; Yorio, 1989). In an investigation of the verb-noun collocations produced by both native and non-native speakers of English, Laufer and Waldman (2011) found a far lower number of verb - noun collocations produced by L2 learners (5.9%) compared with their native-speaker counterparts (10%). Similar results were reported by Howarth (1998a) and Granger (1998a). In Howarth's (1998a) study, the percentage of conventional verb – noun collocations was 25% among the writings by L2 learners, compared with 38% for native-speakers. Likewise, Granger (1998a) observed that NNSs used significantly fewer intensifying adverbs ending in *ly* (e.g. *completely, highly*) in terms of both types and tokens. As to the use of lexical bundles (e.g. *this can be seen, there seems to be*), L2 learners also manifested much lower use (60) compared with native speaker peers, whose writing showed a considerably large number of lexical bundles (130) (Ädel and Erman, 2012). Failing to make a wide use of native-like expressions often results in a lack of diversity in writing, which leads to a sense of foreignness and even oddness in NNSs' writings.

The lack of diversified use of collocations is also characterised by overuse and underuse of certain collocations. In the spoken production of formulaic sequences and routines by non-natives, both De Cock et al. (1998) and Foster (2001) discovered an overuse of some vagueness tags (e.g. *and so on*) and a highly significant underuse of other vagueness tags (e.g. *sort of thing, stuff like that*). Turning to written performance, Yorio (1989) recorded that in the writings of 25 ESL students, non-natives produced a far lower proportion of 'idiomatic' phrasal verbs (e.g. *bring up*) than natives.

[16] The findings of L2 collocation studies were neatly summarised into overuse, underuse and misuse by Laufer and Waldman (2011) and Paquot and Granger (2012) in their review of L2 collocation studies. This broad summarisation is employed in the present study.

Through analysing the adjective - intensifier combinations amongst the writings of intermediate and advanced learners with L1 German, Lorenz (1999) concluded that learners underused more restricted collocations and overused collocations that are less restricted. The type of overused collocations are always linked to lexical combinations in learners' L1 (e.g. Granger, 1998a; Kaszubski, 2000) In Granger's (1998a) study, the widely used combinations (e.g. *closely linked, deeply rooted*) typically had a close translation equivalent in learners' L1 French, but combinations non-congruent with their L1 were underused (e.g. combinations with *highly*, which is relatively much less frequent in French) (1998a: 148f). The same pattern is discerned in the use of discourse frames by French learners of English. They were reported to massively overuse the active voice frames which correspond to the uses of sentence introductory phrases in French (e.g. *We can see that …*).

As Cobb (2003: 408) pointed out, what distinguishes L2 learners from NSs "is the small number of precasts[17] advanced learners have at their disposal, and the extent to which these are used and overused". The underlying reason for the overuse and underuse phenomena that emerge in L2 learners' collocation uses is that learners tend to "'cling on' to certain fixed phrases and expressions which they feel confident in using" (Granger, 1998a: 156). These fixed phrases and expressions become their 'safe bets' (ibid: 148), 'islands of reliability' (Dechert, 1983: 184), even referred to cutely as 'lexical teddy bears' (Hasselgren, 1994: 237) or 'collocational teddy bears' (Nesselhauf, 2005: 69). Therefore, learners' heavy reliance on familiar collocations leads to overuse and avoidance of those which they are unsure in using leads to underuse. These non-native features of L2 collocation production are in fact not surprising since in the process of interlanguage development, overuses and underuses of collocations are unavoidable phenomena, as is the case with the use of grammatical structures or lexis. Additionally, the comparison of collocational uses between NSs and NNSs will inevitably reveal less diversified uses in non-natives since L2 learners not attaining native-like proficiency naturally cannot reach a level on a par with NSs. This is where Contrastive Interlanguage Analysis encounters criticism, to the effect that there tends to be an oversimplified generalisation of learners' overuse and underuse when their language is in direct comparison with native speakers' (Li, 2009: 16). In other words, overuse and underuse is hardly a specific problem of collocation. What is more important in L2 collocation studies is to investigate the forms of misuses and find the underlying difficulties confronted with collocation learning.

[17] 'Precast' refers to prefabricated chunks by Cobb (2003).

3.2.1.2 Collocation misuse

Previous L2 collocation studies report a large proportion of inappropriate uses of collocations. Nesselhauf (2005) investigated the verb – noun collocations in a corpus of writings by advanced German-speaking learners of English, and one of her principal findings was that approximately one third of the collocations were unacceptable or questionable. She concluded that advanced learners had considerable difficulties in selecting the correct verbs in verb-noun collocations. The proportion of erroneous collocations is supported by Laufer and Waldman (2011), in whose study learners at three proficiency levels produced about a third of erroneous verb – noun collocations. An even larger proportion of errors were identified in a gap filling task where learners were asked to supply the correct collocates of frequent adjective-noun or noun-noun combinations, and non-native speakers achieved an average accuracy of only 34% (Hoffman and Lehman, 2000).

Although L2 learners produce a large proportion of collocation errors, past studies indicate that not all types of collocations pose equal problems to L2 learners. They experience greater difficulty in producing verb + noun collocations by L2 learners than with other types of collocations, e.g. adjective + noun collocations. In a cross-sectional study on the development of Greek ESL learners' collocation knowledge, Gitsaki (1999) discovered a developmentally determined acquisition order: adjective – noun collocations were "easy" and "early acquired" type of collocations and verb – noun collocations, were the "difficult" and "late acquired" ones. In Siyanova and Schmitt's (2008) study on the production of adjective – noun collocations by Russian advanced learners of English, they noted that a large percentage of learners' collocations were appropriate (75.3% were attested at least once in the BNC). Among the appropriate collocations, around 45% of the collocations were not only appropriate, but also frequent and strongly associated English word combinations. Only one quarter of the combinations were not attested in the BNC. However, the absence of these collocations in the BNC does not mean these collocations are erroneous since they "found evidence that many of these were in fact appropriate as well" (ibid: 437). When L2 learners' performance in AN collocations was compared with that of native speakers, very little difference was found between the use of appropriate collocations. Moreover, not only good performance on adjective + noun collocations has been observed in previous research, an improvement of collocational knowledge with rising proficiency is identified. Gitsaki's (1999) study

showed that learners' accuracy on adjective – noun collocations (e.g. *sore throat, marine life, heavy drinker* in Gitsaki's study) increased as they became more and more proficient. Likewise, the better command of adjective – noun collocations with rising proficiency was also found in Zhang and Chen's (2006) study. In a cross-sectional study to test both the receptive and productive knowledge of English adjective-noun collocations among three groups of EFL learners at different proficiency levels, higher level subjects were found to have obviously better command of AN collocations than lower levels in the acceptability judgment tasks and translation exercises. Results indicate learners' adjective-noun collocational knowledge develops along with the rise in language proficiency.

In spite of learners' good performance on adjective + noun collocations, collocations in general undoubtedly pose great learning difficulties even for proficient L2 learners. Most researchers are in agreement that this learning difficulty involves the arbitrary restrictions in word combinations. Findings arising out of studies investigating L2 learners' erroneous collocations and the degree of restriction in a word combination suggest that combinations with a medium degree of restriction are more prone to errors than more restricted combinations (Martelli, 2006; Nesselhauf, 2003; 2005). Nesselhauf (2003) for example, found that more restricted collocations like *pay attention* and *run a risk* were less prone to errors than combinations where a verb takes a wider range of nouns (verbs like *exert, perform, reach*). It was further suggested that more restricted collocations were learnt as wholes whereas the less restricted ones were used creatively (ibid: 233).

Some researchers point out that the relative infrequency of individual collocations in input is a problem for L2 collocation learning. Henriksen (2013: 49) argues that "collocations are more low-frequent than the words that make up the collocations, and learners therefore mostly lack sufficient exposure to collocations". Exposure to collocations is good for the learning of a second language and for L2 collocations as well, and the lab-based study of collocation learning by Durrant and Schmitt (2010) has confirmed that frequent input helps the learning of collocations. A large amount of collocation input is a contributor in collocation learning, as language input is beneficial for the learning of other L2 aspects, but it is not sufficient. L2 learners do not pay attention to collocational relationships between words even when they encounter collocations (Wray, 2002). Unlike collocation acquisition by native speakers, L2 learners are influenced by their mother tongue in both collocation learning and production, and the influence of learners' L1 is a significant factor commonly identified as linked to (mostly erroneous) collocation production in L2 collocation studies.

The next section will discuss the role of learners' mother tongue in the learning of L2 collocations.

3.2.2 The role of learners' L1

Past studies have demonstrated that L1 plays an important role in L2 collocation learning, though no empirical evidence has been gathered to compare L1 influence on collocations with its influence on other aspects of acquisition (e.g. phonology, syntax, and morphology) through quantifying and comparing the percentage of interference errors.[18] Studies of the influence of L1 on L2 collocation acquisition generally fall into two major areas: those focusing on L2 learners' collocations in terms of L1-induced inappropriate uses and those exploring the potential influence of L1 in terms of L1 and L2 congruent and non-congruent collocations.

3.2.2.1 L1 influence in terms of L1-induced inappropriate collocation uses

Many collocation studies into L2 learners' collocation performance discovered traces of L1 in erroneous collocations (e.g. Biskup, 1992; Farghal and Obeidat, 1995; Martelli, 2006). Biskup (1992) observed the translation performance on collocations by Polish and German learners and compared their collocation errors in terms of cross-linguistic influence. She found that for Polish learners of English, the errors were loan translations or extension of L2 meaning on the basis of the L1 words, whereas German learners tended to produce errors resulting from assumed formal similarity, e.g. *to crack nuts* as **to crunch nuts*. Biskup (1992) interpreted these two types of L1 influence as the perceived differences between languages on the part of learners: the Polish learners saw a distance between Polish and English and thus did not assume much formal similarity, whilst German learners assumed more formal similarity between their mother tongue and English. Farghal and Obeidat (1995) analysed the tendencies of lexical simplification that learners followed in two elicitation tasks: blank filling and a translation task. Four strategies that learners adopted in producing collocations were distinguished: synonymy, avoidance, transfer and paraphrasing, among which transfer took up 9.9% and 12.9% of all attempted collocations among two groups of learners. However, some caution

[18] There is evidence that for advanced L2 learners, L1 influence plays a marginal role in the acquisition of word formation devices (Olshtain, 1987), but L1 is believed to play a larger role in lexis.

is needed here since 'avoidance' as a strategy is a complex phenomenon, and it is not clear whether subjects in Farghal and Obeidat's study knew the target collocations but preferred avoiding them and used other forms instead. One example given is *light food,*[19] for which learners produced *soft food, little food, quick meal*, etc. To call this strategy avoidance rather than transfer is questionable since avoidance is one manifestation of language transfer (Ellis, 1994). So the percentage of L1 transfer might make up an even larger proportion in Farghal and Obeidat's data. L1 influence was also confirmed in Martelli's (2006) study in which it had a relevant role in the generation of wrong lexical collocations. However, unlike Farghal and Obeidat (1995), the proportion of L1-induced errors was not quantified in Martelli's study.

The traces of L1 in erroneous collocation uses have been investigated and quantified, with findings showing that L1-influenced errors make up a large amount of errors even for learners at advanced levels. In the erroneous uses of verb – noun collocations by Chinese learners of English with different proficiency levels, Zhang and Gao (2006) noticed a varied proportion of L1-influenced errors, from nearly one third to more than a half. Likewise, L1 influenced errors were most frequent in Nesselhauf's studies, where L1 influence occurred in about half of the non-native collocations (2003; 2005). A higher percentage of L1 induced errors – over 60% of those produced by intermediate and advanced learners were identified by Laufer and Waldman (2011) and the number of L1-induced errors was not found to decrease over time. Apart from L1-transfer errors, another consequence of heavy reliance on their mother tongue in collocation production is the overuse of certain collocations that are similar between two languages and underuse of patterns that are mismatched in two languages (cf. Section 3.2.1.1).

3.2.2.2 The role of L1 in terms of L1 and L2 collocation (non)congruence

Another line taken by past studies of the role of L1 in L2 collocation acquisition examines the potential influence of L1 lexical combinations on the learning of L2 collocations from the perspective of (non)congruence. Collocations in an L2 are either congruent, with direct translation equivalents, or non-congruent, without direct translation equivalents with learners' L1. It is maintained that L1 congruency may facilitate, and non-congruency hinder, the acquisition of L2 collocations (Bahns, 1993;

[19] *Light food* is actually not a target-like collocation. Rather *light meal* is a native-like one.

Philip, 2007; Wolter, 2006; Wolter and Gyllstad, 2011; Yamashita and Jiang, 2010). Wolter (2006) provided a theoretical account for how learners' already-established L1 lexical and conceptual knowledge might influence the building of word connections in the L2 lexicon, and argued that for L2 learners, the assimilation of new L2 words into paradigmatic hierarchical connections (e.g. *dog-animal, dog-terrier, dog-cat*) is easier than the process of building syntagmatic connections between L2 words (e.g. *small room*), since the latter requires restructuring of the existing L1 network. For example, Japanese learners of English have to restructure their network when encountering a *small room* since in their L1 the concept of 'a small room' is expressed as *a narrow room*. Learners' L1 lexical network "acts as an integrated set of 'placeholders' for L2 lexical items" (Wolter, 2006: 743) and this L1 knowledge can be useful for acquisition of L2 collocations that are similar to the L1 and be a hindrance if discrepancies occur. Bahns (1993) holds the same view as Wolter (2006). Illustrating two types of congruent and non-congruent collocations in German and English, he made a strong argument that among the tens of thousands of collocations over which L2 learners should have command, only the ones which are non-congruent with learners' L1 should be taught. The validity of this argument which ignores congruent collocations and focuses exclusively on non-congruent ones is not flawless and will be discussed below in Chapter 9.

The potential influence of L1 on the learning of L2 collocations at a psycholinguistic level is not only theoretically predicted but also explored in experiments through psycholinguistic techniques. Yamashita and Jiang (2010) administered a real-time (online) phrase-acceptability judgment task to a group of native speakers of English, Japanese English as a second language (ESL) users, and Japanese English as a foreign language (EFL) learners. They were tested on both congruent and non-congruent collocations. Subjects were asked to read a stimulus presented on a computer screen and make a judgment about its acceptability by pressing a Yes or No button on a keyboard as quickly as possible. Results showed that both EFL and ESL learners made more errors with non-congruent collocations than congruent collocations, and EFL learners reacted more slowly to non-congruent collocations than to congruent ones. Another principal finding is the indication that once non-congruent collocations are stored in memory, they are processed autonomously without word-by-word mediation of the L1. In a similar vein, Wolter and Gyllstad (2011) administered a primed Lexical Decision Task to a group of L1 Swedish learners of English and a group of English native speakers serving as controls on collocations in three conditions: congruent collocations, non-congruent collocations and unrelated

items for baseline data. Their aim was to investigate whether collocational priming[20] occurs, and whether L1 knowledge influences how L2 collocations are processed. Results demonstrated that words prime their collocates, as evidenced by the significant differences of reaction times between collocations and unrelated items for the native speaker group. For NNSs, interesting findings arise with regard to responses to congruent and non-congruent collocations. Firstly, congruent collocations received more primings than non-congruent collocations and the latter were responded to more slowly than congruent ones. Secondly, for non-congruent collocations, no significant difference of error rates was found with those of congruent collocations, and it was suggested that once non-congruent collocation "is recognised as a legitimate collocation in the L2, it becomes stored as such psychologically and when the first word in the collocation is observed the second word of the collocation is anticipatorily activated" (Wolter and Gyllstad, 2011: 442). Furthermore, another principal finding suggesting the active role of L1 is that a 'dual-activation' was found for learners performing tasks entirely in an L2: when an L2 word was activated, it stimulated not only the L2 word's collocates, but also the L1 translation equivalent and its L1 collocate.

Both the studies conducted by Yamashita and Jiang (2010), Wolter and Gyllstad (2011) indicate a considerable influence of learners' L1 on the processing of L2 collocations in the mental lexicon. It is widely acknowledged that L1 plays a role in L2 acquisition, since late learners have already acquired a well-developed L1 lexical and conceptual network. Yet collocation acquisition seems highly susceptible to L1 influence (Paquot and Granger, 2012: 140). This can be accounted for with the existing models of the structure of bilingual memory. One of the influential models about the bilingual memory of late learners is proposed by Kroll and Stewart (1994). According to their Revised Hierarchical Model (see Figure 3-1 below), there are asymmetric links between lexical representations and between lexical representations and concepts. Though the lexical and conceptual links in both L1 and L2 lexica and the lexical links between L1 and L2 are bidirectional, they differ in strength, i.e. the lexical link from L2 to L1 is assumed to be stronger than the lexical link from L1 to L2, and the link from L1 to conceptual memory is assumed to be stronger than the link from L2 to conceptual memory (Kroll and Stewart, 1994: 158). The RHM reflects the consequences of L2 acquisition in late learners, who, on the one hand, possess a fully developed L1 lexicon and their associated concepts (so the stronger link between L1 words and concepts) and on the other, L2 words access meaning via L1 translation equivalents to (so a

[20] Collocation priming is "the tendency for an activated word to accelerate the subsequent recognition of a collocate" (Wolter and Gyllstad, 2011: 431).

strong lexical link from L2 to L1). For words that have no corresponding translations, a very different process may be involved (Jiang, 2002).

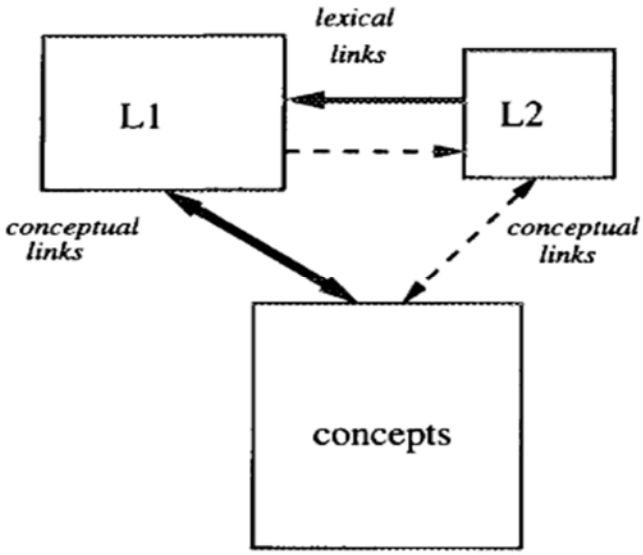


Figure 3-1 Kroll and Stewart's (1994) Revised Hierarchical Model

In the meantime, the RHM proposes that the organisation of bilingual memory changes with rising L2 proficiency, in the form of developing the ability to conceptually process L2 words directly, without mediation of L1 translation equivalents. However, studies show that even for proficient learners, the L1 translation equivalent is activated when processing the L2 word for meaning access (Thierry and Wu, 2007).[21] Thus, if the acquisition of L2 lexical knowledge is initially clinging onto L1 lexical/conceptual networks as the RHM predicts, it seems highly likely that in the production process, L1 is firstly activated prior to the production of L2 words (cf. the 'dual-activation' in Wolter and Gyllystad (2011) discussed above). That is where L1 transfer begins and yet not all features of L1 are activated and transferred to the L2. According to Jiang's (2000) psychological model of L2 vocabulary acquisition for late bilinguals (see Figure 3-2 below), a majority of L2 words fossilise at the first language lemma mediation stage, when the lemma information of the L1 (containing semantic and syntactic information) is copied into the L2 lexical entry whilst lexical information at the lexeme level (containing morphological and phonological/orthographic specifications) is stored in the L2 lexical entry.

[21] Kroll et al. (2010) argue that proficient bilinguals may access the translation equivalent after they understand the meaning of the L2 word. The exploration of intricacies of this debate on whether highly proficient bilinguals access meaning for L2 words through the mediation of their L1 is beyond the scope of the present study and won't be discussed further.

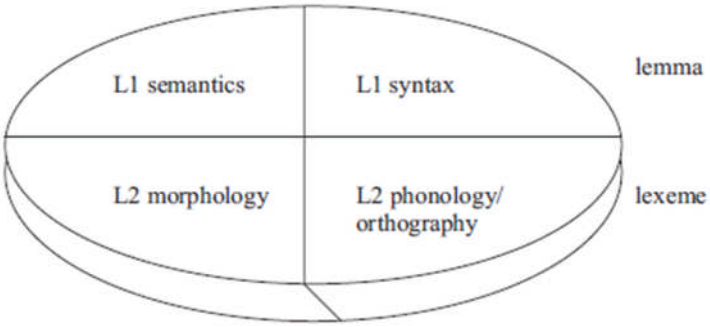


Figure 3-2 Jiang's model of fossilised L2 lexical knowledge (cited in Wolter and Gyllstad, 2011: 446)

The semantics and syntax of an L2 are quite likely to be influenced by learners' L1. Considering the nature of collocations – the arbitrary co-selection of word combinations, collocations are primarily word combinations representing syntactic and semantic relationships between lexical items. In other words, collocation is positioned at the part of the shadowy area between grammar and meaning (Nation, 1990: 38). Therefore, collocations are highly susceptible to L1 influence.

However, the influence of learners' L1 in the learning of collocations in terms of (non)congruence has not been fully examined through the production of collocations. One common feature of the studies by Yamashita and Jiang (2010) and Wolter and Gyllstad (2011) is that they tap L2 learners' receptive collocation knowledge by conducting psycholinguistic experiments. As discussed in Section 3.1.1, there is always a gap between learners' receptive knowledge and their natural production. In their studies, L2 learners were more likely to accept collocations similar to their L1 as legitimate collocations than those that are different from their L1 patterns, but it is not clear yet whether they at the same time produce congruent collocations with more accuracy than non-congruent ones. If so, Bahns's (1993) claim that only non-congruent collocations need to be taught appears valid. However, Nesselhauf's (2003; 2005) studies of L2 learners' production of congruent and non-congruent collocations challenge Bahns's claim. 11% of the congruent collocations produced by learners in her studies were found to be erroneous and the percentage of erroneous non-congruent collocations among all the non-congruent ones was 42% (Nesselhauf, 2003). Similar ratios were obtained in another study: 17% for the erroneous congruent and 42% for the erroneous non-congruent collocations (Nesselhauf, 2005). These figures show that although congruent collocations are easier than non-congruent ones for L2 learners, they also pose problems; on the other hand, not all non-congruent ones pose problems,

since nearly a half of them are not particularly problematic for L2 learners. However, these figures are to be treated with caution since, according to Nesselhauf (2003; 2005), congruence is not only measured on the level of content words, but also on the level of grammatical words. For example, a combination such as *participate in an event* is considered non-congruent since for the German equivalent *an einem Ereignis teilnehmen,* German *an* and English *in* are not considered translation equivalents. Including grammatical items in defining (non-)congruence would naturally lead to more non-congruent collocations since "items in closed grammatical classes normally behave differently across languages" (Salkie, 2002: 56). In addition, it seems paradoxical that phrasal verbs were treated as a single word and differences of particles (e.g. *up, over*) were disregarded. So if congruence is only measured on content-word-level, more non-congruent collocations in her studies would be classified into congruent ones, which leads to a modification in the interpretation of her findings.

3.2.3 Collocation lag

Apart from the deficiencies in collocation uses uncovered by previous studies, collocation knowledge has also been measured against knowledge of lexis and learners' proficiency, with the general findings showing a collocation lag in SLA acquisition. Schmitt and Carter (2004: 13) pointed out that L2 learners' formulaic language tends to lag behind other aspects. Wray (2002: 182) conveyed the same impression that "by the time the learner has achieved a reasonable command of the L2 lexicon and grammar, the formulaic sequences appear to lag behind". Empirical evidence has been gathered so as to measure learners' collocation knowledge by comparison with other aspects of language learning. For example, Bahns and Eldaw (1993) analysed learners' translations in terms of collocation errors and errors with lexis in order to determine whether learners' knowledge of collocations was on the same level with their knowledge of general vocabulary. Collocation errors were reported to be more than twice as frequent as general lexical errors. They concluded that learners' "knowledge of general vocabulary far outstrips their knowledge of collocations" (ibid: 108). In a different vein, learners in Barfield (2007) were asked to report on their knowledge of individual words and on combinations of these words, and it was found that learners reported better knowledge of the former than the latter (Barfield, 2007, cited in Laufer and Waldman, 2011). Irujo (1993) examined the use of idioms produced by 12 bilingual native speakers of Spanish, whose English was virtually indistinguishable from a native

speaker's, and discovered a deficiency of their knowledge of idioms relative to their general production with few grammatical or lexical errors.

A deficiency in collocation knowledge among learners at very advanced levels is also indicative of a lag in collocation knowledge. Farghal and Obeidat (1995) observed that language teachers of English who had a minimum of five to ten years' teaching experience were seriously deficient in collocations. In Nesselhauf's (2005) study, students with 10 to 17 years' English learning produced a similar proportion of erroneous collocations as students who had only studied English for 5 to 10 years. It was summarised that the "length of a learner's exposure to English in English-speaking countries was shown to probably have a slight effect on collocational accuracy, whereas the number of years a learner had undergone classroom teaching was shown to have no effect" (2005: 237). A comparative study on collocation performance across learners at three proficiency levels conducted by Laufer and Waldman (2011) showed a much stronger collocation lag, as the advanced and the intermediate learners produced significantly more erroneous collocations than the basic learners. Similar results were obtained in Obukadeta's study as discussed in Chapter 1. Despite the heterogeneity in L2 collocation studies, these results clearly indicate a noteworthy lack of positive correlation between general language proficiency and collocation knowledge.

The question whether collocation knowledge can be related with general proficiency is not of crucial importance here, since on the one hand, it is difficult to establish a clear link between language proficiency and phraseological competence, the former of which is usually loosely measured in terms of the number of years of English instruction for research purposes (Paquot and Granger, 2012: 137); on the other, it is evident that collocation lags behind other aspects of L2 knowledge and 'may floor even the proficient non-native' (Wray, 2000: 463). Thus, what is centrally important is to investigate what factors are associated with collocation lag.

The poor phraseological performance even for learners at an advanced level is explained in terms of lacking awareness of collocational relationship between words, i.e. learners do not pay attention to collocational relationships, and collocations "are initially seen as compositional combinations of words rather than as a phenomenon of co-selection" (Philip, 2007: 3). Studies testing learners' intuition about collocations that are frequent in the L2 show a weak sense of collocational relationships (Channell, 1981; Granger, 1998a; Siyanova and Schmitt, 2008). For example, in examining the collocational competence of a group of eight advanced learners who were asked to mark the acceptable collocates of adjectives

from a list of nouns, Channell (1981: 120) found that "learners fail to realise the potential even of words they know well, because they only use them in a limited number of collocations of which they are sure". To test whether French learners of English had an underdeveloped sense of what constitutes a significant collocation, Granger (1998a) used a word-combination test in which subjects were asked to choose all the adjectives which collocated with 11 amplifiers ending in *ly* and functioning as modifiers (e.g. *highly, bitterly*). Learners marked over 100 fewer frequent collocations than the native speakers, providing clear evidence of learners' weak sense of collocations compared with that of native-speakers (Granger, 1998a: 152). In a similar vein, participants in Siyanova and Schmitt (2008) rated native-like collocations as far less frequent, and atypical collocations as more frequent than those by NSs. The ignorance of collocating relationships in language input naturally leads to production which is "subject to whatever interlanguage rules the learner is operating under" (Yorio, 1989: 62). A typical illustration of this process from inability to recognise a collocation to utilising interlanguage rules is given by Wray (2002: 209):

> … the adult language learner, on encountering *major catastrophe*, would break it down into a word meaning 'big' and a word meaning 'disaster' and store the words separately, without any information about the fact they went together. When the need arose in the future to express the idea again, they would have no memory of *major catastrophe* as the pairing originally encountered, and any pairing of words with the right meaning would seem equally possible: *major, big, large, important, considerable,* and so on, with *catastrophe, disaster, calamity, mishap, tragedy,* and the like.

Wray's explanation of the way learners treat and produce collocations is consistent with Wolter's (2006: 746) claim that "the process of building syntagmatic connections between words in an L2 appears to be considerably harder than the process for building paradigmatic connections". Furthermore, the acquisition of new words may interfere with the production of collocations in the selection of appropriate collocates from a set of related words. One illustration of semantic relatedness is synonymy and the use of synonyms has been identified as the most frequent strategy adopted in producing collocations (Farghal and Obeidat, 1995; Irujo, 1993). Thus it seems that L2 learners' vocabulary size is closely linked with their collocation learning, though most probably in a negative way. The study conducted by Gyllstad (2005: 1), for example, suggests that "learners with large vocabularies have a better receptive command of verb + NP collocations than learners with smaller vocabularies". Yet there is still unclarity about the relationship between vocabulary size and the production of collocations, which is important for an understanding of L2 collocation learning. There is to date still a paucity of research into the relationship between vocabulary increase and collocation production.

Therefore, in order to uncover the underlying factors inhibiting L2 learners' collocation learning, this study seeks to investigate the relationship between vocabulary growth and collocation learning, particularly the increase in vocabulary in a set of semantically related words.

3.3 Summary

This chapter has reviewed the growing body of research that has been undertaken in the past several decades into collocation learning by L2 learners. A wide range of data types has been utilised in past research, including elicitation and spontaneous data, or a mixture of the two types. With each data type possessing unique advantages, learner corpora are gaining more and more popularity in terms of either naturalness of learner language or large quantities. These distinctive advantages of learner corpora will be further explored, as the present study will utilise a corpus of written English produced by Chinese EFL learners, in order to investigate their collocation performance and vocabulary increase.

Past research into L2 collocation studies has covered many types of learners, including learners of different mother tongues or different proficiency groups. Their research foci are different types of collocations, i.e. verb + noun collocations, adjective + noun collocations, lexical phrase, lexical bundles, routines, etc. Given this heterogeneity in L2 collocation research, direct comparisons of research findings is difficult, but there emerges a general picture for the learning of collocations by L2 learners, i.e. collocation learning poses special difficulty for L2 learners, as evidenced by deficiencies of collocation overuse, underuse and misuse even for learners at advanced levels. As pointed out in Chapter 1, collocation knowledge is believed to lag behind grammar and lexis and constitutes the 'last and most challenging hurdle in attaining near native-like fluency' (Spottl and McCarthy, 2004: 191).

In general, past studies of L2 collocation acquisition have been concerned with a description of learners' collocation performance and in providing evidence of the problems confronted with L2 learners. In this sense, research into second language collocation learning is still at an early stage. What brings about, or contributes to, the collocation lag still remains unclear. The point of departure of this research is postulating that vocabulary increase can be associated with collocation lag, and the study thus aims to test this prediction through a developmental study of collocation learning by Chinese EFL learners.

Chapter 4: Research design

4.0 Introduction

The main focus of this corpus-based research is to examine the relationship between vocabulary increase and collocation uses by Chinese learners of English, with the aim to test whether vocabulary increase is associated with the collocation lag in the field of second language acquisition. This chapter outlines the design of such a cross-sectional study. Section 4.1 briefly mentions research purpose and research questions; Section 4.2 presents the types of collocations to be targeted in learners' writings and justifies why verb + noun, adjective + noun and noun + noun collocations are chosen rather than other types of collocations; Sections 4.3 introduces the learner corpora used: Chinese Learner English Corpus; Section 4.4 explains the selection of two collocation dictionaries to be referenced; Section 4.5 briefly introduces the British National Corpus as a native speaker reference corpus to check the acceptability and appropriateness of combinations; Section 4.6 lists the software adopted for automatic data collection, for the creation of databases, and for statistical analyses. The main procedure of the study is presented in Section 4.7, followed by a summary of this study design (Section 4.8).

4.1 Research purpose and questions

The aim of this study is to examine the relationship between vocabulary growth and the production of L2 collocations. As was discussed in Chapter 1, verb + noun collocations are targeted and the growth of verbs is examined from two perspectives: from delexical verbs to lexical verbs and the increase in verbs within a synonym set. The thesis sets out to answer the following questions regarding VN collocations:

1. What developmental patterns appear in the verb + noun collocations produced by L2 learners, in terms of delexical verb and lexical verb + noun collocations?

a. Is there a tendency towards increasing use of lexical verb + noun collocations with rising

proficiency?

b. Is there a tendency towards increasing errors with lexical verb + noun collocations, and decreasing delexical verb + noun collocation errors with rising proficiency?

2. Within specific semantic domains of the verbs in verb + noun collocations used by all levels of learners, is there a tendency for these verbs, as they increasingly occur at the higher levels, to be associated with collocation errors?

In addition to targeting the most frequent type of collocations (verb + noun collocations), two other frequent types of collocations, i.e. adjective + noun (AN) and noun + noun (NN) collocations will be further investigated in the same way as VN collocations, in order to compare with the findings from VN collocations. Questions concerning these two types of collocations are:

3. Are adjective + noun and noun + noun collocations produced by Chinese L2 learners at the same accuracy level as verb + noun collocations? If not, what patterns do they follow?
4. Within specific semantic domains of the adjectives in adjective + noun collocations and nouns in noun + noun collocations used by all levels of learners, is there a tendency for these adjectives/nouns, as they increasingly occur at the higher levels, to be associated with collocation errors?

As stated in Chapter 1, the role of L1 will also be examined in the learning of L2 collocations by Chinese learners. This thesis will test whether congruent collocations are easier than non-congruent ones, and whether non-congruent collocations once acquired, are less prone to errors.

Thus the design of this study covers Chinese learners' production of VN collocations, as well as their production of AN and NN collocations. The hypotheses concerning verb + noun collocations, adjective + noun and noun + noun collocations are interlinked as they contribute to testing the vocabulary growth factor in collocation learning, whilst the hypothesis regarding L1 influence is a separate enquiry. For this reason, the following sections present the design for investigation of VN, AN and NN collocations, whilst the detailed procedure for analysing L1 influence is given separately in Chapter 9.

4.2 The selection of verb + noun, adjective + noun and noun + noun collocations

As was presented in Section 2.2.3, lexical collocations are divided into six types: adjective + noun; (subject-) noun + verb; noun + noun; adverb + adjective; verb + adverb; verb + (object-) noun (Benson et al., 2010; Hausmann, 1989). Among these, verb + noun collocations were primarily targeted because they are the most frequent and important (Benson et al., 2010; Howarth, 1996; 1998a) and at the same time constitute a frequent source of difficulty for L2 learners (Bahns and Eldaw, 1993; Benson, 1985; Biskup, 1990; Cowie, 1991; 1992; Gitsaki, 1999; Howarth, 1998a; b; Nesselhauf, 2005; Palmer, 1933). Furthermore, verb + noun collocations are the most frequent type of collocation errors in the learner corpus we are going to investigate (1, 572 out of 2, 940 tokens of collocation errors made by six levels of Chinese learners of English).

However, in VN collocations, verbs and nouns were not given equal weight in this study. The focus was on verbs, based on the assumption that it is the nouns (nodes) that determine the verbs to go with.[22] In both speech and writing, words are produced and arranged in a linear sequence, which makes the production process seem as if the preceding words select the following ones (verbs/adjectives select the following nouns). In fact it is the noun where language users generally start from when forming ideas (Cowie, 1998; McIntosh et al., 2009) and "the most important kind of collocations sought by a writer or translator is the one based on the noun, for it is the noun that sets the semantic context of the sentence" (Kozlowska and Dzierzanowska, 1988: 8). With the case of *lingering doubt* as an example, *doubt* selects an acceptable adjective — *lingering* but not *loitering* (Cowie, 1998: 222-223). The selection of the verb or the adjective by the noun is reflected in the organisation of collocation dictionaries like *Selected English Collocations* (Kozlowska and Dzierzanowska, 1988), where nouns are the headwords. Therefore, verbs in VN collocations produced by L2 learners will be investigated since the appropriate verb has to be chosen to collocate with the noun previously selected (e.g. *acquire knowledge* but not **grasp knowledge*, *play a role* but not **occupy a role*).[23]

[22] Similar approach is taken by Bahns and Eldaw (1993: 103), by whom the noun was viewed as the node and a verb as collocate. Howarth (1998a) on the contrary cast doubts on the direction of selection from the noun to the verb through the erroneous collocations **the contrast is drawn* and **place weight on*, and speculated it is the other way around. However, the two examples are just indicative that nouns (*contrast* and *weight*) are preselected, but the collocating verbs went wrong (*draw* and *place*).

[23] In fact, in the process of the extraction of verb + noun collocations, we found that wrong choices of the nouns were rather rare and there were only with a few instances such as **solve the question*. It will be seen in Section 4.7.2.

In addition to verb + noun collocations, two other types of collocations were briefly examined in the same way as VN collocations to compare with our findings from analyses of VN collocations. They are adjective + noun and noun + noun collocations, which are the top two most frequent word combinations used by native speakers of English (Johansson and Hofland, 1989). They are not only frequent, but also susceptible to error in L2 collocation performance. Adjective + noun and noun + noun collocation errors are the second and third most frequent types of collocation errors according to the error analysis of Chinese Learner English Corpus (henceforth referred to as CLEC) (Gui and Yang, 2003). Yet these error types have scarcely been focused on in previous L2 collocation studies (except in the studies on AN collocations conducted by Martelli 2006; Siyanova and Schmitt 2008). The investigation of these two types of collocations was aimed at finding out whether the findings with leaners' use of these two types are in line with the findings on verb + noun collocations.

4.3 The learner corpus – CLEC

The present study focuses on the developmental patterns of Chinese EFL learners' use of collocations and thus requires a longitudinal corpus of the performance of Chinese learners over an extended period of time, or else a corpus of the performance of Chinese learners at different proficiency levels in English, i.e. an apparent-time approach to the study of language development. A longitudinal corpus would be much preferred, but it is unfortunately unavailable at the time of the research. However, a corpus of the written performance of learners at different proficiency levels makes an apparent-time study possible. Our study made use of the one-million-word Chinese Learner English Corpus, a computerised textual database of writings by Chinese learners at five different levels of proficiency. CLEC is homogenous in the sense that all learners are Chinese learners of English; at the same time it is heterogeneous because it represents learners at different developmental stages. The apparent-time design assumes that the performance of different age groups of learners at different proficiency levels is indicative of successive stages of development. The inclusion of learners of five learning stages is a distinct advantage of CLEC, which cannot be outweighed by other published learner corpora. Other widely used and more recent corpora recording the written performance of Chinese learners include Spoken and Written English Corpus of Chinese Learners (SWECCL) (Wen et al., 2008), the Chinese sub-corpus of ICLE (Granger et al., 2009) and the British Academic Written English Corpus (BAWE)

(Nesi, 2011). SWECCL documents the spoken and written data of English major university students and it records the writings of English majors of four grades. The Chinese sub-corpus of ICLE contains argumentative essays written by higher intermediate to advanced Chinese learners of English at universities, and BAWE contains texts from proficient Chinese undergraduate students studying in several UK universities. Though these learner corpora are newer and some of them are larger than CLEC, they only cover learners of a fixed or limited range of proficiency levels. Learners below the university level are not targeted in these corpora. CLEC, therefore, is chosen for study.

We next provide a brief introduction to CLEC. The Chinese Learner English Corpus is a one-million word collection of compositions produced by Chinese learners of English at five developmental phases: high school students (coded by the developers as ST2, approximately corresponding to the upper levels of secondary school students in the UK), non-English major university students of lower grades (ST3) and higher grades (ST4), and English majors of lower grades (the first and second years, ST5) and higher grades (the third and fourth years, ST6) (Gui and Yang, 2003). The writings of the five groups of learners were recorded in separate files, each approximating 200, 000 words. For the convenience of discussion about the five sub-corpora, it is preferable to achieve uniformity between learner types and the corresponding files of their writings. Therefore, the files were named as follows:

a) ST2: high school students; the written performance by ST2 learners;

b) ST3: first and second year non-English major university students; the written performance by ST3 learners;

c) ST4: third and fourth year non-English major university students; the written performance by ST4 learners;

d) ST5: first and second year English majors; the written performance by ST5 learners;

e) ST6: third and fourth year English majors; the written performance by ST6 learners

The project was undertaken by teachers of English in various universities across three cities (Guangzhou, Shanghai, Xinxiang) in China, so the learners targeted in CLEC were from several middle schools/universities rather than from one particular institution. The sub-corpora of ST2, ST5 and ST6 are made up of learners' free compositions, whereas ST3 and ST4 consist of timed writings for

tests (the national general English proficiency tests: Band 4 and Band 6). In terms of text types, the assignments for the ST2 group were mainly narrative (probably it is still early for Chinese middle school students to develop argumentative writing skills), and their writings were not confined to one topic. The ST3 and ST4 sub-corpora contain argumentative essays. For the texts produced by ST5 learners, they are not confined to one or two topics and are partly argumentative and partly narrative. The ST6 sub-corpora contain argumentative essays, and most of the topics are what the ICLE suggested in their data collection process, which are "should euthanasia be legalised in China?", "crime does not pay", "the abolition of prison systems", "the value of university degrees" and "should a man/woman's financial reward be commensurate with their contribution to the society they live in", etc.

Not all the five sub-corpora were examined in this research. As noted earlier, the sub-corpora of ST2, ST5 and ST6 are made up of learners' free compositions, whereas ST3 and ST4 consist of timed writings for tests. Only the ST2, ST5 and ST6 learner files were used, as they contain the same data type – free compositions. So it is assumed they did not undergo the time and mental pressure in timed writing for tests and they could turn to referencing tools for help in the writing process. The three groups of learner data are thus homogenous. It is furthermore generally assumed that the quantity of formal English instruction learners receive is indicative of their proficiency in English. Thus, it is possible to conduct an apparent-time study on Chinese EFL learners' collocation performance on the basis of the years of instruction they get. In light of the components of CLEC, a clear dividing line in English proficiency is observable, i.e. pre-university Chinese EFL learners (ST2) and university-level learners (ST3, ST4, ST5 and ST6). Likewise, university-level students can be further divided into non-English majors (ST3 and ST4) and English majors (ST5 and ST6). However, it cannot be claimed that the proficiency of ST3, ST4, ST5 and ST6 learners is in a continuum because of the difference in majors (non-English major vs. English major). It is possible that some non-English majors are better than English majors in the overall English performance. Consequently, the intensity of English instruction is not used as a criterion to distinguish the proficiency level of non-English and English majors, even though the latter might be better than the former in general.

For these reasons, three groups of learners were examined in our study: ST2 – pre-university high school students, categorised as the "basic" level, ST5 – English majors of lower grades as the "intermediate" level, and ST6 – English majors of higher grades as the "advanced" level. The

classification of the levels, as discussed above, is based on the years of English instruction learners receive, and is mainly adopted for straightforward comparison. The ST2 learners had at least 3 to 5 years' classroom English instruction in China; the ST5 group had at least 6 to 7 years and the ST6 learners had English instruction for at least 8 to 9 years.

4.4 Collocation dictionaries for reference

As stated in Chapter 2, the approach taken to collocation in this study is mainly phraseological, and also frequency-based. For this reason, two collocation dictionaries were selected so as to check on well-formed and erroneous collocations produced by Chinese learners. One is non-corpus-based – *the BBI Combinatory Dictionary of English* (*3rd edition*) (Benson et al., 2010) (henceforth: the *BBI*), while the other is corpus-based – *Oxford Collocations Dictionary for Students of English (2nd edition)* (McIntosh et al., 2009) (henceforth: the *OCDSE*).[24] The *BBI* as a collocation dictionary represents the phraseological approach and the *OCDSE* the frequency-based approach. The *BBI* includes both grammatical and lexical collocations and has a neatly presented organisation for the noun entries (e.g. verb + object noun, adjective + noun and noun + noun collocations). It has been widely consulted in studies conducted by Cowie (1992), Gitsaki (1999), Howarth (1996), Laufer and Waldman (2011), and Nesselhauf (2004). However, for the *BBI*, some common collocations have escaped the intuitions of the authors (Klotz, 2003: 58) (e.g. *do sport, give a comment* which are absent in the *BBI*, but recorded in the *OCDSE*). Collocations from the *OCDSE* are retrieved from a large reliable database of actual language use – the Oxford English Corpus of over two billion words. It is preferable to other corpus-based lists of word combinations, e.g. the *Frequency Analysis of English Vocabulary and Grammar (Vol. 2)* (Johansson and Hofland, 1989) and *A Dictionary of English Collocations* (Kjellmer, 1994). Word combinations in these two books are respectively based on one million words of the LOB and Brown corpora whose limited size "poses problems, particular for the study for word combinations" (Johansson and Hofland, 1989: 14) and is often inadequate for the lexicologist (Kjeller, 1994: xiii).[25] So the *OCDSE* has been chosen for its wide coverage. Therefore, both dictionaries are consulted in the process of extracting collocations. If a collocation is included in either of the two dictionaries, it is considered as a

[24] The use of the two dictionaries in attesting collocations was also endorsed by Siyanova and Schmitt (2008).
[25] For example, *commit a crime*, a commonly accepted collocation, is not included in Johansson and Hofland's book.

well-formed collocation.

4.5 The reference corpus – BNC

The British National Corpus was chosen to serve as a benchmark for measuring the appropriateness of learners' production of collocations which failed to be attested in collocation dictionaries. As a general corpus representing as wide a range of modern British English as possible, the 100-million-word corpus contains over 4, 000 written texts and transcripts of speech in British English (McEnery et al., 2006). It has to be noted that the BNC only covers British English of the late twentieth century. However, creativity and productivity are two of the design features of human language, by which it means that humans are able to construct understandable linguistic forms, some of which have even not been used before. Language is constantly changing, with the gradual emergence of neologisms. Some of these new constructions and interpretation of words and expressions are accepted by the language community and acquire status in the language stock. By the same token, new combinations of words, i.e. collocations, are constantly coined and gaining acceptance. Yet the BNC is not timely updated to include new forms of language use, and its limited size means that it fails to cover a wider range of English language uses. Considering the ever-changing nature of language, few corpora can include all collocations, which makes the recognition of appropriate/inappropriate word combinations difficult. However, relative to the creative use of language, there is always a conventional core in any language, which remains stable and usually becomes the learning target of language learners. Attested in a large corpus, conventional expressions have the most frequent occurrences and creative expressions are usually on the bottom of the frequency list. In published dictionaries, conventional language uses are prioritised and recorded. Therefore, in this study collocation dictionaries were taken as the criterion for locating conventional English collocations. As was discussed in Chapter One, conventional language uses are set both as the norm for L2 learners and as the criterion for judging the appropriateness of learners' interlanguage. If a collocation is listed in the two dictionaries, it was recognised as correct (cf. Section 4.4). If it is not included in the dictionaries, this research adopted the on-line version of the BNC — BYU-BNC,[26] with the aim of checking whether it is acceptable or not. The BNC was also used to locate the target

[26] http://corpus.byu.edu/bnc/ [Accessed 10 March 2012]

collocating word if an appropriate one was not found in collocation dictionaries. For example, in our learner database, create + poem, was viewed as incorrect as it was not recognised as a conventional collocation in the dictionaries, nor was recorded in the BNC, though it is understandable in the English language. The detailed procedure for identifying well-formed and erroneous collocations by using collocation dictionaries and the BNC will be shown in Section 4.7.2.

4.6 Software for retrieval and analysis

Data extraction has been greatly facilitated with the increasing sophistication and availability of computers. To allow highly efficient and labour–saving data collection and analyses, the following types of software were used: AntConc 3.2.4w and Wordsmith 5.0 to perform the function of concordancing and word-list generation; EditPad Pro 7 and PowerGREP 4 to automatically collect words of a particular part of speech and word combinations through regular expressions; Microsoft Office Excel 2010 to help create databases of collocations and perform the function of computing and graphing; and finally GraphPad Prism (Version 6.04) was used to carry out statistical analyses.

Verb + noun collocations were semi-automatically collected, i.e. via an automatic generation of all the verbs in concordances and a manual identification of verb + noun collocations (cf. Section 4.7.2 for detailed explanation). For the retrieval of other words or word combinations, the following regular expressions were used in PowerGREP 4:

a. For the retrieval of verbs: (\w+)_VV\w+
b. For the retrieval of nouns: (\w+)_NN[12]\s|(\w+)_NN\s
c. For the retrieval of adjectives: (\w+)_J\w+
d. For the retrieval of AN combinations: (\w+_J\w+\s)((\w+_NN[12]\s)|(\w+_NN\s))
e. For the retrieval of NN combinations: (\w+(_NN[12]\s)|(_NN\s))(\w+(_NN[12]\s)|(_NN\s))

4.7 Procedure

4.7.1 Tagging and reliability check

When originally compiled, CLEC was error tagged into 61 types of error. However, it was decided not to base the present study on the error-tagged version, which was found to be inadequate in the following respects: firstly, well-formed collocations relevant to the present study were not identified and tagged in CLEC; secondly, error-tagging in CLEC was faulty as some erroneous collocations were missed out while some well-formed ones were included; thirdly, the error tagging targeted erroneous word combinations of all types (including problematic collocations, erroneous free combinations, colligation errors, etc.) rather than exclusively collocation errors (cf. Zhang and Gao, 2006). For the purposes of the present research, we rectified the above problems by applying the following procedure: all the error tags were firstly removed and then the clean corpus was part-of-speech tagged, followed by a reliability check. Collocations were finally semi-automatically extracted with reference to the two widely used collocation dictionaries discussed above.

4.7.1.1 POS Tagging

The total size of the three sub-corpora (ST2, ST5 and ST6) amounts to over 600,000 words. It would have been an impossibly large undertaking to extract collocations manually by looking through the corpora word by word. For verb + noun collocation extraction in this research, therefore, the starting point was to locate all the verbs and then manually sorted out VN collocations (the justification of this collection method will be given in Section 4.7.2). For this purpose, the corpora were first automatically part-of-speech tagged using the online tagging service developed by University Centre for Computer Corpus Research on Language at Lancaster University.[27] The current standard tagset – CLAWS 7 was used for its richness in detailed subdivisions of word types.

[27] http://ucrel.lancs.ac.uk/claws/trial.html [Accessed 1 March 2013]

4.7.1.2 Reliability check

After the POS tagging, a reliability check was performed. CLAWS is thought to achieve a consistent accuracy of 96-97% and even 98.3% for the tagging of some portions of the BNC (Garside, 1987; 1996). Those figures are obtained through tagging the texts of native speakers, although the accuracy rate varies according to text types. For the tagging of learner language, a lower degree of accuracy is generally believed to be achieved, since tagging learner language is complicated by instances of grammatical and morphological errors. Prior to the retrieval of word combinations, a reliability check was carried out on a sample of over 1, 000 words in the ST6 file.

A straightforward and commonly used way of checking tagging validity is to locate how many tagging errors occur in a sample of texts. Thus two pieces of writings with a total of 1,311 words from the tagged ST6 corpus were randomly selected. After word-by-word examination, 29 words were found to be incorrectly tagged. So the tagging reliability was 97.8% ((1311-29)/1311*%), a fairly high accuracy rate and very much in line with the accuracy rate on native speaker texts.

Additionally, a further check was conducted on an approximately equivalent sample of texts in the ST2, with the aim to find out whether CLAWS gave an equally high reliability rate for texts produced by much lower proficiency levels. So a sample with a total of 1,336 words was chosen. The accuracy rate was – perhaps surprisingly – as high as that found in ST6: 97.9% ((1336-28)/1336*%).

Therefore, CLAWS achieved a very high rate of accuracy in our learner texts. Even within the wrongly tagged words, most of the errors concerned other word classes rather than verbs. Examples of the very few tagging errors for verbs are:

(1) Euthanasia_NP1 ,_, or_CC mercy_NN1 killing_NN1 ,_, **means_NN** helping_VVG to_TO hasten_VVI the_AT death_NN1 of_IO a_AT1 person_NN1 who_PNQS is_VBZ badly_RR suffering_VVG ._.

(2) In_II China_NP1 ,_, suicide_NN1 is_VBZ legal_JJ ,_, which_DDQ means_VVZ ,_, people_NN are_VBR legal_JJ to_TO kill_VVI themselves_PPX2 in_II a_AT1 helpless_JJ condition_NN1 ,_, so_RR what_DDQ we_PPIS2 **conside_NN1** is_VBZ only_RR whether_CSW it_PPH1 is_VBZ legal_JJ to_TO end_VVI the_AT life_NN1 of_IO a_AT1 incurable_JJ patient_NN1 ._.

In both cases, *means* and *conside* (a spelling error for *consider*) were tagged as nouns rather than verbs. Altogether among the 29 tagging errors in the ST6 sample, 5 verbs were wrongly tagged for other word classes instead of as verbs, making up only 0.38% (5/1311*%) of all the tags. Calculated

against the total number of verbs (276 in total) in the sample, the rate of wrongly tagged verbs was 1.8% (5/276*%). Therefore, the POS tagging was taken to be at least as reliable as CLAWS generally is, and the fact that a very small number of verbs were missed out was not problematic for the overall analyses.

4.7.2 Investigation of verb + noun collocations

In the semi-automatic extraction of VN collocations in the ST2, ST5 and ST6 files, only collocations of a verb and a noun as its object were counted (e.g. *make a contribution*, *acquire knowledge*). VN combinations can be easily retrieved with regular expressions performed by PowerGREP. However, there are varied positions of the nouns as the objects of the verbs. Verbs and nouns are not confined to the immediate linear sequence: verb + (modifiers) + noun (e.g. *make a plan*). As in the examples given by Greenbaum (1970: 10) (cf. Section 2.2.1), a collocational relationship can even transcend a sentence. The following categories of the noun's varied positions relative to concordanced verbs were also examined, e.g. the noun used before the verb in a passive voice (*great progress has been made; such problems would be solved*) and in attributive clauses (*life pays everyone in different ways for the contribution he makes to the society, *she can use the knowledge she had learned in the new job*). In these examples above, VN collocations were subsequently retrieved: *make a plan, made progress, solve problems, makes contribution* and **learned knowledge*. So an automatic extraction of verb + noun combinations within a specified span would not only leave out some combinations, but also "yield a great deal of unusable material, the sifting of which would probably be even more time-consuming than the manual extraction of all verb-noun combinations from the corpus" (Nesselhauf, 2005; 43).

Therefore, taking the different proximities of verb + noun collocations into consideration, VN combinations were not automatically retrieved but rather a semi-automatic approach was applied: all the verb tags (except the copular *be* and modal verbs) were searched, followed by a manual extraction of the nouns as the collocates of the verbs (phrasal verb + noun patterns, e.g. *put on weight* were disregarded).[28] Although nouns select other lexical words, they were not first searched because a large number of irrelevant information would be extracted (e.g. adjective + noun, (subject) noun + verb,

[28] This method for identifying verb + noun collocations has also been adopted by Howarth (1996; 1998a; b).

preposition + noun). As is acknowledged by Howarth (1996: 78), "searches based on the verb would more sharply focus the searches on the desired patterns".[29]

Verbs were classified into eight main categories in CLAWS: VV0 (base form, e.g. *work*), VVD (past tense, e.g. *worked*), VVG (-ing participle, e.g. *working*), VVGK (-ing participle catenative, e.g. *going in be going to*), VVI (the infinitive form, e.g. *it will work…*), VVN (past participle, e.g. *worked*), VVNK ((past participle catenative, e.g. *bound in be bound to*) and VVZ (the present tense form, e.g. *works*).[30] Considering catenative verbs are followed by *to* infinitives rather than nouns, they were disregarded. The remaining 6 forms of lexical verbs were examined. In addition, verbs of *do* and *have* were separately tagged in the CLAWS and they fell into the category of delexical verbs to be searched. Altogether 18 verb taggers (six forms of verbs + six forms of *do* and six forms of *have*), totaling 87,957 tokens, were examined in concordances generated by AntConc. Next came the manual extraction of well-formed and erroneous VN collocations.

a. Extracting well-formed and erroneous collocations

A collocation was taken to be well-formed when it was found either in the *BBI* or in the *OCDSE*.[31] Collocations which were not listed in the two dictionaries, but were attested in the BNC, were disregarded, for the association was too loose to be viewed as a collocation (e.g. *?give pressure, ?eat tea*). A VN collocation was viewed as erroneous (e.g. **do + problem*) when it was neither listed in the two dictionaries, nor found in the BNC. Wider contexts of the concordanced verbs were checked in cases of ambiguity. The target verbs for the erroneous collocations were supplied by consulting the *BBI*, the *OCDSE*, or the BNC (e.g. *acquire knowledge* but not **learn knowledge*, *seize time* but not **grasp time*). In only a few cases where neither source supplied the target verb for an erroneous collocation, a native speaker was consulted. As stated above, the following cases were not considered when identifying verb + noun collocation errors:

(1) Colligation errors with verb + noun collocations were disregarded, considering what is of central importance in this study is to examine the (in)correct choices of verbs, not the (in)correct choices of

[29] This method of taking verbs as the starting point in the extraction is different from the study by Laufer and Waldman (2011), in which nouns were searched as node words. They started from a set of 220 pre-selected nouns and proceeded with the identification of verb collocates. The way of choosing a limited number of frequent nouns would inevitably leave out many verb + noun collocations. Our study performed an exclusive extraction.

[30] http://ucrel.lancs.ac.uk/claws7tags.html [Accessed 1 March 2013]

[31] A question is whether word combinations (*open the door*) that are found in the learner corpus and also included in either the *BBI* or *OCDSE* should be listed in my database. Combinations like *open the door* were not included, since they are viewed as free combinations based on our definition of collocations. The two dictionaries were only used to attest well-formedness.

grammatical forms. Therefore, errors in phrasal verbs, determiners, prepositions and the number of nouns were not counted (instances of such errors are **make one's mind*, **hunt a job*, and **give some advices)*.

(2) Errors involving free verb + noun combinations were also eliminated (e.g. **abandon prisons, *build heroes*).

(3) Errors involving the wrong choices of the nouns were eliminated (e.g. **earn his life, ?*solve questions*).

The following chart presents a summary of the procedure adopted for identifying well-formed and erroneous VN collocations:

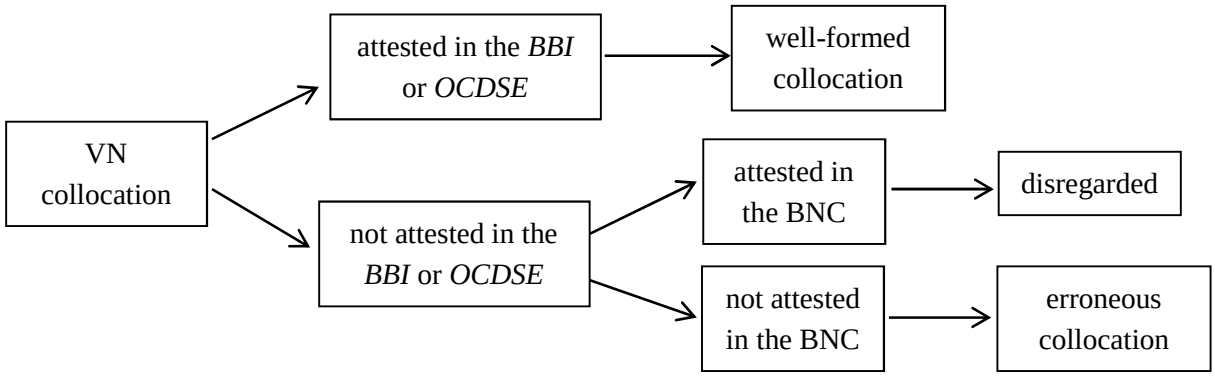


Figure 4-1 Procedures for identifying well-formed and erroneous VN collocations

b. Creating databases for all the collocations produced by the three levels of learners

Excel software was used as a tool for storing all the collocations as well-formed ones (further divided into delexical verb + noun collocations and lexical verb + noun collocations) and erroneous ones (further divided into erroneous delexical verb + noun collocations and erroneous lexical verb + noun collocations), with information on collocation tokens and types recorded. For the convenience of study, verbs and nouns in collocations were lemmatised and articles, determiners and adjectives in between them were not recorded. For example, the following collocations were viewed as instantiations of the same collocation (*make + plan*): *make a plan*, *makes a plan*, *make plans*, *made plans*, etc. One advantage of getting the verbs and nouns in VN collocations lemmatised and recorded into separate and parallel columns was to facilitate subsequent automatic analyses.

The procedure described above was required for the investigation of the developmental patterns

of VN collocations in terms of delexical verb + noun and lexical verb + noun collocations. Answering the next question about the relationship between vocabulary increase and collocation acquisition required the implementation of the following steps:

c. Classifying the lexical verbs in VN collocations into synonym sets

As a first step in examining the relationship between verb increase in a semantic domain and collocation uses, lexical verbs in well-formed and erroneous collocations were classified into synonym sets or synsets, in WordNet parlance. Words are synonyms "if they have a significant similar semantic content" (Saint-Dizier and Viegas, 1995: 18). It is widely acknowledged that no exact synonyms exist in a language and many words are in a loose relationship of synonymy with varied degrees of meaning overlap (Palmer, 1981). For the classification of verbs in VN collocations, one major criterion is the semantic content a word carries. In this sense, Levin's (1993) classification of semantically coherent verbs was adopted. For instance, verbs like *compose* and *create* in the ST2 database and words in the ST6 database like *build*, *form*, and *draw* were categorised into *verbs of creation*. However, it was not sufficient to base synonym classifications exclusively on Levin's classifications, since verbs grouped into one class by Levin were primarily 'syntactic' synonyms with semantically coherent bond. Verbs of one semantic category were not exhaustively listed by Levin (1993), e.g. the verb *establish*, which semantically belongs to 'verbs of creation' but was listed in the category of 'verbs with predicative complement'. Therefore, other referencing sources were consulted, namely the *Oxford Dictionary of Synonyms and Antonyms* (henceforth *ODSA*) (Spooner, 2005) and WordNet (the web interface),[32] a large on-line word reference system in which verbs are organised into synonym sets (cf. Fellbaum, 2010; Miller, 1995; Miller et al., 1990). In the Wordnet system, synonymy is defined as the many-to-one mappings of word forms and concepts. Words representing the same concept are called synonyms (e.g. *boot* and *trunk*). Synonyms are grouped into unordered sets, viz. synsets (Fellbaum, 2010: 232). One advantage of using the WordNet is that like Levin's classification of verbs, it displays the larger semantic domain to which a verb belongs. In sum, three referencing sources were applied in classifying verbs in VN collocations into synsets, i.e. *English Verb Classes and Alternations* (henceforth *EVCA*) (Levin, 1993), the *ODSA* and WordNet. If verbs are co-listed in at least one of the three dictionaries, they were placed in the same set.

[32] http://wordnetweb.princeton.edu/perl/webwn [Accessed 10 May 2013]

Given that verbs are polysemous, their synonyms were located specifically within the sense of the verb in a VN collocation. For example, the verb *discharge* has 11 senses (accordingly, 11 synsets) as listed in WordNet and 5 synsets in the *ODSA*. In the VN collocation – *discharge + duty* produced by ST2 learners, only synonyms of the sense of *discharge* were recorded (e.g. *complete*). Similarly, *acquire* in *acquire + knowledge* was grouped under 'learn' verbs instead of being placed into the synset of 'obtaining'.

Classifying verbs into synsets was not confined to verbs under the same entry covered in the referencing sources. Instead, if two verbs were not listed as synonyms, but they had a shared synonym, the three verbs were grouped in a synset. For example, *fix* and *place* were not listed as synonyms in either the *ODSA* or WordNet, but they were respectively in a synonymous relationship with *attach*, so they were placed in the same synset. Similarly, all three words were synonyms of *put* and they were gathered in one synset. Accordingly, new verbs were gradually accumulated in the synset of 'verbs of putting'.

Besides the above criterion of synonym classification, two more loose criteria were adopted, i.e. context and foreign-language equivalents. Words are defined as synonyms if they both fit in a particular context (Palmer, 1981; Saint-Dizier and Viegas, 1995). These synonym pairs are context-dependent synonyms. Based on this criterion, the verbs *lead* and *live* were synonyms in the given context of *lead/live + life*. Another standard was a cross-linguistic one. According to Benson et al. (1986: 204), one of the definitions of synonymy is foreign-language equivalent. Thus in the data sets of the verbs, *wear* and *dress* were synonyms in the sense that they both share one Chinese translation equivalent (*chuan*), though they behave quite differently in English.

In all, the criteria of classifying verbs in VN collocations in our study include semantic similarity (synonyms), context-dependent synonyms and foreign-language equivalents. Three referencing sources were applied: the *EVCA*, the *ODSA* and WordNet. There was no hierarchy in applying these criteria and resources and instead there were alternatives. For the convenience of study, each synset was given a name according to the classification given by the *EVCA*, e.g. verbs of creation (*compose, create, build*, etc.), and verbs of obtaining (*achieve, earn, receive*, etc.) In cases where there was no umbrella term for the semantic set, a representative verb was used to cover the synset, e.g. *fulfil* verbs incorporating verbs like *fulfil, accomplish, apply*, etc.

Finally, verb + noun collocations with the verbs falling into the synsets classified were

investigated. The purpose was to find out whether there are more VN collocation errors in higher levels within synsets where there is an increase of verbs, and whether these errors are more associated with new verbs than old verbs.

4.7.3 Investigation of adjective + noun and noun + noun collocations

In our study, adjectives in adjective + noun collocations only refer to attributive adjectives, e.g. *bright colour*. Predicative adjectives, although they can form a collocational relationship with the noun subject (e.g. *the colour is bright*) were not considered for the convenience of automatic retrieval of attributive adjective + noun combinations. As is classified in the POS tagset of CLAWS 7, nouns subdivided into 21 categories, e.g. common nouns (singular, plural and neutral for number), nouns of titles, locative nouns, temporal nouns, proper nouns, etc. Only the most frequent and important common nouns were targeted in the extraction of AN and NN combinations, since nouns of titles, locative nouns, temporal nouns and proper nouns do not fall into collocational relationships with adjectives.

The analysis of AN and NN collocations followed the same methodology as VN collocations, except for the retrieval process. Due to the constant position of nouns in the AN and NN combinations, where the noun is directly adjacent to the preceding adjectives or nouns, data were automatically collected by PowerGREP. However, unlike the extraction of verb + noun collocations, identifying AN and NN collocations from the retrieved combinations meant distinguishing them from compounds first. A compound noun is "a fixed expression which is made up of more than one word and which functions in the clause as a noun (Sinclair and Fox, 1990: 25). To demarcate a AN or NN collocation from an adjective + noun or noun + noun compound (e.g. *high school, news bulletin*) is not easy, since they both involve frequent co-occurrence of word constituents, though compounds are much more fixed and mutually predicted and function as one unit of meaning. Distinguishing AN and NN collocations from compounds was not within the scope of our investigation. Instead, recognition of AN and NN collocations were carried out with reference to the two dictionaries – the *BBI* and the *OCDSE*. If an AN or NN combination was listed in either of the two collocation dictionaries, it was taken as a collocation (rather than a compound).

4.8 Summary

This chapter has described the design of this cross-sectional investigation of Chinese learners' collocation written performance. Two major points were covered: the relevant materials utilised and the procedure involved. A summary of the design is presented in the following table.

Table 4-1 A brief summary of the design of the study

Materials	The learner corpus	CLEC
	Groups of learners	ST2, ST5 and ST6 learners
	Types of collocations	verb + noun, adjective + noun and noun + noun collocations
	Software	AntConc 3.2.4w, Wordsmith 5.0, EditPad Pro 7, PowerGREP 4, Microsoft Office Excel 2010 and GraphPad Prism (Version 6.04)
	Referencing dictionaries	The *BBI* (3rd Edition) and the *OCDSE* (2nd Edition)
	Reference corpus	The BNC
Procedure	Processing of the sub-corpora	a. Part-of-speech tagging of the three sub-corpora b. Reliability check (a high tagging reliability)
	Investigation into VN collocations	a. Extracting well-formed and erroneous collocations b. Creating collocation databases c. Classifying lexical verbs in VN collocations into synsets
	Investigation into AN and NN collocations	Automatic extraction of AN and NN combinations; the same methods as analyses of VN collocations were adopted

The next chapter reports the findings of this large scale study of Chinese learners' production of English verb + noun, adjective + noun and noun + noun collocations, which used the methodology presented above.

Chapter 5: Verb increase and the production of verb + noun collocations (1)

5.0 Introduction

This chapter enquires into the relationship between vocabulary increase and collocation uses by L2 learners seen from the overall perspective of the growth from delexical verbs to lexical verbs. As part of the investigation, it presents the overall analyses of VN collocations produced by all the three levels of learners. In addition to the comparison of our findings with previous ones, the learning of verb + noun collocations by L2 learners is analysed from along following lines: the overall results and general patterns of verb + noun collocations produced by the three proficiency levels (Section 5.1), the developmental patterns of delexical verb + noun (abbreviated: DeLeVN) and lexical verb + noun (LeVN) collocations (Section 5.2), the comparison between overall verb growth and VN collocation errors (Section 5.3), followed by a summary of these three overall analyses (Section 5.4).

5.1 Overall analyses (1): general patterns of VN collocations produced by L2 learners

5.1.1 Overall tokens of collocations

Altogether 5,068 instances of collocations (including both well-formed and erroneous ones) were extracted (ST2: 1,579; ST5: 1,660 and ST6: 1,829). In terms of absolute frequencies, the number of collocations produced by each level of learners shows a gradual increase. However, there was no proportionate increase in collocation production if the likewise gradually increasing sizes of each file were taken into consideration (ST2: 208,088; ST5: 214,510; ST6: 226,106). This means that there was no quantitative development in VN collocational knowledge as learners become more proficient. Direct comparison of these results with previous studies, as regards the numbers of VN collocations relative to the size of learner corpora is difficult, given the varied methods they employed in counting

verb + noun collocations and the foci on learners at different proficiency levels. Instead of an exhaustive retrieval of all the possible verb + noun collocations, Laufer and Waldman (2011) calculated VN collocations starting from a predetermined set of nouns with high frequencies in a native speaker corpus and reported fewer collocations (852) in their corpus of advanced learners (approximating the size of our ST6 file). Howarth's (1996; 1998a; b) studies, instead, presented a manual and exhaustive analysis of all the VN collocations and reported around 1,000 collocations in about 25,000 words of L2 academic writing. Compared with the proportion of collocations extracted from the three files in our study, there was a much higher percentage of collocations in Howarth's study, a discrepancy probably due to two factors: the definition of collocations and the focus on learners at different proficiency levels. In Howarth's studies, free and restricted collocations and idioms were all included, which naturally resulted in a larger number. Furthermore, Howarth's subjects were full-time postgraduates of Linguistics and English Language Teaching studying in an English-speaking country, and thus the proportions of collocations produced by these highly proficient learners were unsurprisingly higher.

In light of the homogeneity of learners' proficiency level and collocation retrieval methods, Nesselhauf's (2005) data provides a level of comparison in terms of the production of VN collocations by advanced learners. Compared with the 2,082 VN collocations found by Nesselhauf (2005) in her investigation of the around 150,000-word-writing by advanced German learners of English, the number of collocations produced by our advanced learners (ST6) – 1,829, approximates to the number of VN collocations in her findings. Yet there was a difference, as the proportion of collocations out of the total words in her study was 1.7 times as high as the proportion of collocations in the ST6 database. The main reason for this gap in collocation percentages was that the types of VN collocations were rather restrictedly defined in our study. As is illustrated in Section 4.7.2, only verbs with nouns as objects were included, whilst eight other syntactic patterns which verb-noun collocations fell into were counted by Nesselhauf (e.g. *go to prison, fall in love with somebody*) (cf. Nesselhauf, 2005: 68). So there was no major discrepancy between the proportion of VN collocations obtained in our study and that reported in previous L2 VN collocation research.

Though in general it was difficult to compare the number of VN collocations relative to learner corpus size due to the heterogeneity of L2 VN collocation studies, it can be seen from the quantities of collocation production that L2 learners produced far fewer VN collocations as compared to the percentage of VN collocations produced by native speakers of English as reported in other studies.

When the proportion of verb + object-noun collocations produced by NSs was calculated out of the total verb-noun combinations, the proportion was found to be as high as over 40% (Cowie, 1991; 1992; Howarth, 1996). If VN collocation proportion is counted out of the total corpus size, Howarth (1996; 1998a; b) extracted over 5,000 target collocations out of a native-speaker corpus of only about 240,000 words, a proportion 2.5 times as high as that obtained through our L2 learner data.[33] This quantitative discrepancy in terms of collocation uses has been widely acknowledged and empirically tested (cf. Section 3.2.1.1). That learners used fewer collocations compared with NSs on the one hand, shows a poorer sense of collocations; on the other, it demonstrates the greater use of an 'open choice principle' than of an 'idiom principle' (Sinclair, 1991) by L2 learners. This ''open choice principle'' is further manifested through a non-diversified production of collocation types, discussed below.

5.1.2 Overall types of collocations and collocation frequency distribution

The numbers of collocation types produced by the three groups of Chinese EFL learners were: 285 (ST2), 344 (ST5) and 441 (ST6).[34] The overall number of types (1,070) was found to be rather low compared with tokens (5,068). That means on average one collocation was produced 5 times. But the frequency was not so evenly distributed if we take a closer look at the distribution of collocation frequencies over the overall types in each learner group. Figure 5-1 presents the distribution of collocation frequencies over the 285 types of collocations in the ST2 database.

[33] The proportion of VN collocations in native speaker data is in fact much higher, since Howarth (1996; 1998a; b) started from the most frequent verb lemmas (with a frequency of 10 or more) and then proceeded to extract their noun collocates, which means that there still exist a large number of collocations with less frequent verbs.

[34] Collocations like *made plans, made a plan, make a plan* were regarded in this study as instantiations of one collocation type "*make + plan*".

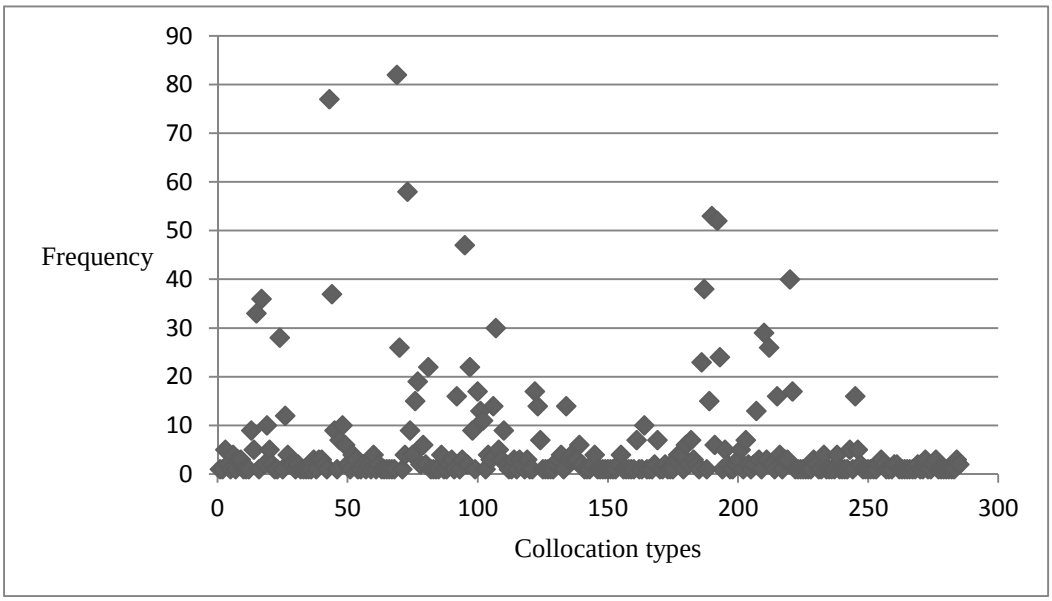


Figure 5-1 VN collocation frequencies distributed over collocation types in the ST2

As is shown in Figure 5-1, a predominant number of collocations had a frequency less than 10, with 5 collocations having a frequency over 50. It can be seen that an overwhelming number of collocations occurred fewer than 5 times, and in fact most of these occurred only once. The same pattern of the frequency distribution across types was also found in the ST5 and ST6 databases. The fact that a majority of the types of collocations were produced less than 5 times demonstrates a varied use of collocations, which further acts as some sign of phraseological competence. However, taking the total tokens into account, this varied use is at the same time accompanied by an overuse of a small number of collocation types.

A further look into the distribution of collocation tokens over their types revealed a huge overuse of a limited number of collocations in all three groups of learners. Table 5-1 shows the types of collocations that were divided into three groups according to frequencies: those with a frequency of 5 or lower (≤ 5), those between 5 and 10 (5-10) and those with a frequency of 10 or more (≥ 10).

Table 5-1 VN collocations divided into three frequency groups

Learners	≤ 5	5-10	≥ 10
ST2	230	16	39
ST5	277	27	40
ST6	381	20	40

As is shown in the above table, a majority of collocations in the three levels occurred less than 5 times (cf. Figure 5-1) and only a small proportion of them had a frequency more than 10. However, this small proportion of collocation produced more than 10 times made up a majority of the overall collocation tokens. Taking the ST2 data for example, 39 types of collocations were used for 1,052 times, making up 67% of the total 1,579 collocations. The same trend for heavy reliance on a limited number of collocation types was revealed in the ST5 and ST6 databases as well (see Figure 5-2 below). Though there was a slight decrease in the proportion of collocations occurring more than 10 times among the overall tokens from ST2 to ST5 and ST6, similar distribution patterns were observed in the ST5 and ST6 collocation databases, i.e. less than 14% collocation types made up more than half of the collocations produced by Chinese EFL learners at over proficiency levels.

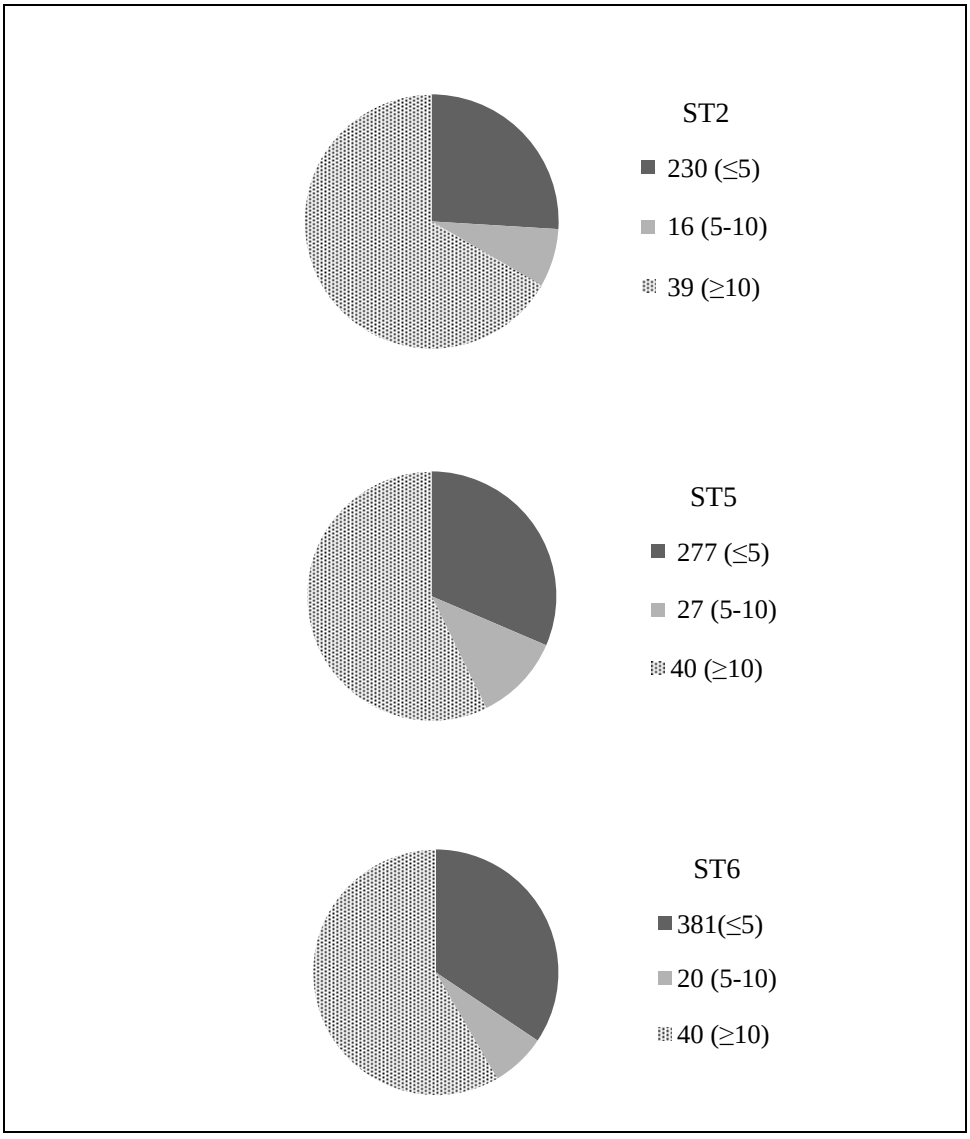


Figure 5-2 The frequency distribution of VN collocation types in ST2, 5&6 databases

On the one hand, the finding obtained in this study mirrors those of previous studies which have identified the phenomenon of phraseological overuse on the part of L2 learners (Ädel and Erman, 2012; Granger, 1998a; Lorenz, 1999; Kaszubski, 2000; etc. cf. Section 3.2.1.1). On the other, the finding of an overuse of a restricted number of collocations by L2 learners was reported in a different way here. A majority of studies find an overuse in L2 collocation data when comparing the uses of a fixed type of formulaic sequences by NNSs with those produced by NSs (e.g. Ädel and Erman, 2012; Altenberg and Granger, 2001; Cobb, 2003; De Cock et al., 1998; Foster, 2001; Granger, 1998a; Lorenz, 1999; Kaszubski, 2000). As discussed in Section 3.2.1.1, it is not surprising that NNSs, still in the process of interlanguage development, are characterised by an overuse of a limited number of collocations, and the attainment of diversified collocation uses, as with the attainment of diversified use of grammatical structures or lexis, is incremental. The finding of collocation overuse revealed in this study is, however, not based on comparisons of non-native and native data. So a rather limited range of collocations occurring more than 10 times in our data were "overused" in the sense that they occurred more frequently than the other types of collocations. Nesselhauf (2005) also identified a set of VN collocations that were most frequently produced but unfortunately her data were not quantified, and the proportion of such an overuse is not known. One striking advantage of examining L2 learners' overuse without comparison with NSs data is that characteristics of collocation uses specific to L2 learners can be authentically investigated.

VN Collocations that were used more than 10 times by at least two learner levels include: *make + progress, make + use, make + friend, make + mistake, take + part, take + care, do + homework, do + exercise, do + good, do + harm, have + dinner, answer + question, lead + life, live + life, pay + attention, play + role, sing + song, solve + problem, spend + time, try + best, wear + clothes* and **learn + knowledge*. The heavy use of a restricted number of collocations by Chinese L2 learners shows the phenomenon of 'collocational teddy bears' (Nesselhauf, 2005: 69) as discussed in Section 3.2.1.1. A general feature of the above expressions is that they are very frequently used in everyday native speaker English. The necessity to use these collocations for communicative purposes may thus facilitate fluent and repeated uses. So a heavy use of certain collocations is not a non-native phenomenon, rather it is a natural characteristic of everyday use. However, though these overused collocations are required in everyday English use, heavy reliance on these collocations, as pointed out in Chapter 1, can add a rigid

flavour to learners' writings (Channell, 1994: 21).

Whether this heavy use is determined by communicative purposes or it is a result of clinging on to 'collocational teddy bears' is not crucial here. What is more important is why L2 learners have such a generally good command of the above VN collocations.[35] Previous studies show that what is common with overused collocations is that they bear a great resemblance with learners' L1 combinations (cf. Granger, 1998a; Kaszubski, 2000). That may suggest an easier acquisition of L2 collocations which are similar to L1 combinations than those differing from L1 patterns. Yet a closer look at the above collocations from a Chinese perspective shows that not all of them are L1-equivalent word combinations, such as delexical verb combinations: *make + progress, make + use, take + part, take + care*, etc.[36] So a necessity for communicative purposes and the frequent input for L2 learners of these overused collocations may well account for such a heavy use. The role of L1 in collocation acquisition will be extensively discussed in Chapter 9.

Though there exists a necessity for the frequent use of collocations in everyday English, as some sign of fluent and idiomatic control of a language, a varied use of collocation types is also needed. As stated in Section 3.2.1.1, the difference of phraseological uses between NSs and NNSs rests in the diversified types produced. Turning to the types of VN collocations produced by all three levels, a between-group comparison of diversification in collocation production was conducted (see Figure 5-3 below).

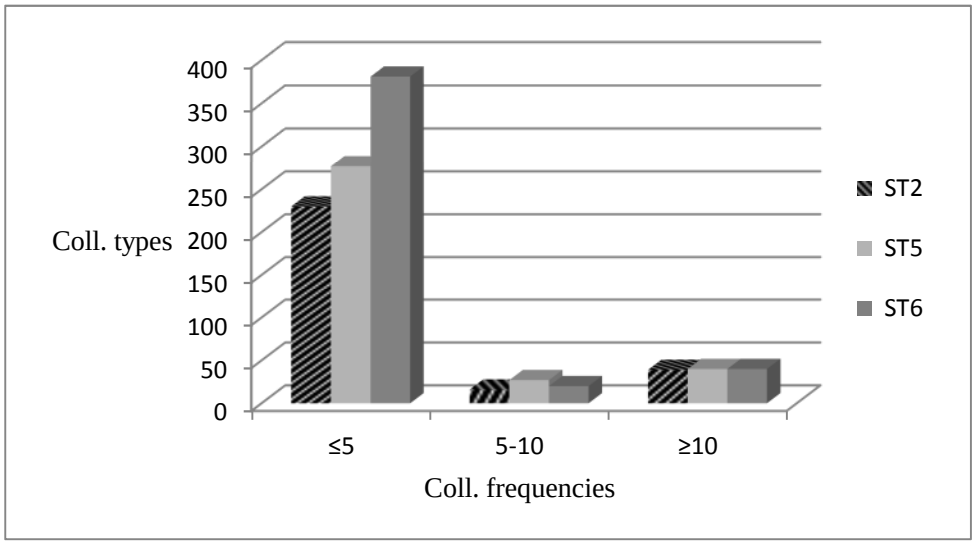


Figure 5-3 Between-group comparison of VN collocations within three frequency groups

[35] Good performance is observed with the exception of **learn + knowledge.*

[36] The Chinese equivalent expressions for *make + progress, make + use, take + part, take + care* are *qude jinbu* (literally: *gain progress*), *shiyong (use), canjia(participate), zhaogu(care)*, among which the last three are Chinese verb lexemes.

As is shown in Figure 5-3 (cf. the corresponding numerical data in Table 5-1), for collocations that were used 5 times or less, there was a clear increase in collocation types in the ST6 level compared with the two lower levels. For collocations with a frequency more than 5, there was not such a growing trend. It shows that despite a heavy use of a rather small number of most frequent collocations at all levels, there was no increase in the types of these frequent collocations across the three proficiency levels, but an increase in far less frequent collocations. That there was an increase in collocation types with the rise of proficiency is consistent with the findings by Gitsaki (1999) and Zhang (1993), who reported that more proficient L2 learners produced more varied collocations than the less proficient L2 learners. It is unsurprising in the sense that more collocations are learned as learners receive more English instruction. So it becomes more important to investigate the quality of their collocation production in terms of misuses.

5.1.3 Collocation misuses

Qualitative deviation is another feature manifested in L2 learners' collocation production. As with findings from previous studies, numerous collocation misuses were also uncovered in our study. The following are some of non-native-like (erroneous) VN collocations: **lay + role, *learn + knowledge, *conduct + crime, *attend + military service, *do + problem,* and **have + progress,* etc. Compared with the uneven distribution of well-formed collocations where a restricted number were overused, the type/token ratio of erroneous collocations was very high (0.57 for ST2 and ST6 learners and 0.4 for the ST5 level), with only three collocations wrongly used on more than 10 occasions (**learn + knowledge, *release + pain, *release + burden*). Thus, the percentage of collocation errors was calculated in terms of collocation types in order to avoid statistical bias caused by uneven distribution of tokens. The well-formed and erroneous collocation types produced by the three levels of learners are presented numerically in Table 5-2. To facilitate comparison, the proportion of erroneous collocations is given out of the total number of collocations.

Table 5-2 Well-formed and erroneous VN collocations in the three levels of learners (types)

Learners	Well-formed collocations	Erroneous collocations	Total
ST2	221	64 (22%)	285
ST5	300	44 (13%)	344
ST6	348	93 (21%)	441

(Notes: ST2 and ST5: $p = 0.0020$ **
ST5 and ST6: $p = 0.0024$ **
ST2 and ST6: $p = 0.7121$ ns
Double asterisks indicate "very significant" and "ns" suggests "not significant".[37])

As shown in the above table, nearly one fourth of the collocations in the ST2 and ST6 databases are erroneous and a rather smaller proportion of collocations (13%) are erroneous in the ST5 database. It is interesting to note that ST5 learners, in the middle level, produced the fewest collocation errors among the three levels. In addition, in order to see whether there is a statistical difference between the proportions of well-formed and erroneous collocations and learner levels, Fisher's test was performed on the above data.[38] Results show that ST5 learners –first and second year English majors produced very significantly fewer VN collocation errors than pre-university middle school students and third and fourth year English majors (ST2 and ST5: $p = 0.0020$; ST5 and ST6: $p = 0.0024$). That collocation errors made by ST5 learners are the fewest was also found by Zhang and Gao (2006) in their analysis of the original error-tagged CLEC. Though they did not investigate correct collocations, they gave the numbers of problematic verb-noun collocations produced by Chinese EFL learners and figures showed that ST5 learners produced the smallest number of errors compared with ST2 and ST6. Taking into other types of erroneous word combinations, viz. noun/noun, noun/verb, adjective/noun, verb/adverb and adverb/adjective combinations, ST5 was also found to produce the smallest numbers of errors in the three levels under discussion (cf. Zhang and Gao, 2006: 32). Therefore, in terms of the frequency of erroneous collocations, the ST5 level is not consistent from the levels of ST2 to ST6. ST5 learners received either one to two more years' English instruction than ST2 learners, and they receive one to two fewer years' instruction than ST6 learners. At the intermediate level, ST5 learners exhibit a higher

[37] The threshold significance level is set as 0.05 by the Graphpad Prism. In the meantime, symbols used by Prism suggesting the level of significance were also adopted, and these symbols together with the p values are: **** ($p < 0.0001$): extremely significant; *** ($0.0001 < p < 0.001$): extremely significant; ** ($0.001 < p < 0.01$): very significant; * ($0.01 < p < 0.05$): significant; ns ($p > 0.05$): not significant (cited from Graphpad statistics guide: http://www.graphpad.com/guides/prism/6/statistics/index.htm?extremely_significant_results.htm) [Accessed 10 June 2013]

[38] Fisher's test can give an exact P value and works fine with small sample sizes. Considering the small number of collocation types, Fisher's test was adopted.

competence than both the lower and higher levels of learners. The question that arises here is that why the middle level outperforms the other two levels. This phenomenon has also been noted by Zareva and Wolter (2012) in L2 word association studies where the intermediate group produced the highest percentage of collocational responses, higher than the advanced group. They further noticed that "the same class (paradigmatic) connections become more prominent as the proficiency of L2 learners of English increases to an advanced level" (Zareva and Wolter, 2012: 59-60). Therefore, in a broad sense, a reverse relationship is suggested between vocabulary growth in the paradigmatic relations and the collocation performance in the syntagmatic relations. To put the matter in simple terms, the more words L2 learners learn, the more they make errors (as seen from the comparison of error ratios between ST5 and ST6). Moreover, a sufficient vocabulary size is undoubtedly important for correct collocation production (as seen from the comparison of error ratios between ST2 and ST5). This relationship between vocabulary growth and collocation is at the heart of our study and will be elaborated in Chapter 6.

Returning for the present to the percentages of erroneous VN collocations produced by L2 learners, the result is similar to other studies of L2 VN collocation production, given the heterogeneity in studies in this field. In Nesselhauf's (2005) investigation into the verb – noun collocations in a corpus of writings by advanced German-speaking learners of English, approximately one third of the collocations were found to be unacceptable or questionable. This proportion was endorsed by Laufer and Waldman (2011), who reported about a third erroneous VN collocations among all the collocations L2 learners produced. It should be noted that there were higher proportions of VN collocation errors in the above studies because more types of errors were included. For Nesselhauf, verb-noun errors include errors of all elements, e.g. verbs, phrasal verbs, nouns, determiners, etc. An example given is the collocation – *come to the conclusion that* and errors in any of these elements were considered (Nesselhauf, 2005: 71). Similarly, errors involving nouns were counted in the study conducted by Laufer and Waldman (2011). The closest point to turn to is the study carried out by Howarth (1996), who found a fourth of the verb – noun collocations his subjects produced were erroneous. Therefore, our finding in terms of the proportion of erroneous collocations is similar to previous L2 VN collocation studies.

Turning to collocation errors as related to L2 proficiency, statistical analysis shows that there was a significant difference between the numbers of erroneous collocations and learner types, viz. ST2 vs. ST5, ST5 vs. ST6, though no significant relationship was found between the ST2 and ST6 learners.

However, the data revealed a persistent proportion of collocation misuses in the two levels (22% in the ST2 level and 21% in ST6 level). This suggests an overall lag in collocation ability, with no sign of decrease in errors with rising proficiency. That there is no decrease in errors accords with the finding of the cross-sectional study conducted by Laufer and Waldman (2011), who found a third erroneous collocations produced by learners at three proficiency levels. Thus, now we have again uncovered a deficiency in the L2 acquisition of collocations, and it becomes important to identify the factor(s) contributing to this lag.

5.1.4 Synopsis of Overall analyses (1)

This section presents the overall analyses of all the verb + noun collocations produced by Chinese EFL learners at three proficiency levels. Overall results were discussed in connection with prior findings in L2 VN collocation studies. In short, this study identified both a quantitative and qualitative deficiency long acknowledged in the collocation performance by L2 learners. Among the approximately 600,000 words of text analysed, only about 5,000 collocations were retrieved, following the criteria set out in the last chapter, thus revealing a quantitative discrepancy in terms of collocation uses and again manifesting a preference for the 'open-choice principle' on the part of L2 learners. This quantitative discrepancy has also been shown through the small number of collocation types (1,070) compared with 5,000 collocation tokens. These figures indicate weak collocational links in the mental lexicon of L2 learners, which corroborates findings from word association tests that L2 learners produced significantly fewer collocational responses than native speakers did (cf. Fitzpatrick, 2006).

In addition, in terms of collocation misuses, it was found that collocation poses problems at all levels, shown by nearly a quarter of all the collocations produced being erroneous. Furthermore, data obtained from this cross-sectional study yields some interesting points: firstly, collocation overuse as reported in previous studies through comparisons of NS and NNS corpora (Ädel and Erman, 2012; Cobb, 2003; De Cock et al., 1998; Durrant and Schmitt, 2009; Foster, 2001; Granger, 1998a) is uncovered from a non-comparison perspective in our study. Through analysing the distribution of collocation tokens over types, this study showed that a small proportion of common collocations make up a majority of the overall collocation tokens. This distribution is quite like word frequency distribution in a corpus of natural language, viz. "a small number of words tend to make up a very large portion of any normal text"

(Milton, 2009: 46). Secondly, between-group comparisons of collocation data revealed some general developmental patterns, viz. the overall number of collocations does not increase with the rise of proficiency (both in terms of tokens and types) but there is a more diversified collocation uses as learners advance to higher levels. Despite the good signs of a development in collocational competence as proficiency rises, there was no decrease in collocation misuses, as collocation errors were persistent even in the ST6 level, depicting a general lag in collocation knowledge. This indicates that collocational knowledge does not improve with the advances of L2 proficiency and the stagnant of collocational knowledge has long been endorsed (e.g. Bahns and Eldaw, 1993; Laufer and Waldman, 2011). The next sections, therefore, moves on to continue the discussion of this lag through examining collocations classified into delexical verb and lexical verb + noun collocations produced by the three levels.

5.2 Overall analyses (2): between-group comparisons of delexical and lexical VN collocations

Vocabulary increase was first broadly measured in terms of the development from delexical verbs to lexical verbs, viz. from very general to more specific verbs in meanings, and then measured locally with reference to particular synsets. One of the main hypotheses is with regard to the increase in verbs from delexical to lexical verbs in collocation production: it is hypothesised that in VN collocation production L2 learners at lower levels make more errors using delexical verbs, whilst those at higher levels make more errors with lexical verbs. If this hypothesis is upheld, it means that the learning of verbs, progressing from delexical to lexical verbs, does not ensure better collocation competence, even though the growth of lexical verbs provides more opportunities for L2 learners to be specific in choosing the right verb to collocate with a noun in specific VN collocations.

As was pointed out in Section 2.2.3, the six commonest delexical verbs targeted are *do*, *give*, *have*, *make*, *take* and *get*. Examples of well-formed and erroneous delexical verb + noun collocations are: *give + comment*, *make + money*, *take + nap*, **give + meeting*, **take + joke*, and **do + game*. Examples of well-formed and erroneous and lexical verb + noun collocations are *achieve + aim*, *claim + right*, *impose + burden*, **ensure + law*, **implement + act*, and **teach + knowledge*. All the well-formed and erroneous DeLexVN and LexVN collocations in the three databases were numerically tabulated in Table 5-3 and graphically presented in Figure 5-4.

Table 5-3 Well-formed and erroneous VN collocations in the three levels of learners (tokens)

Learners	Well-formed collocations		Erroneous collocations	
	DeLexVN	LexVN	DeLexVN	LexVN
ST2	871	596	36	76
ST5	790	768	30	72
ST6	568	1097	34	130

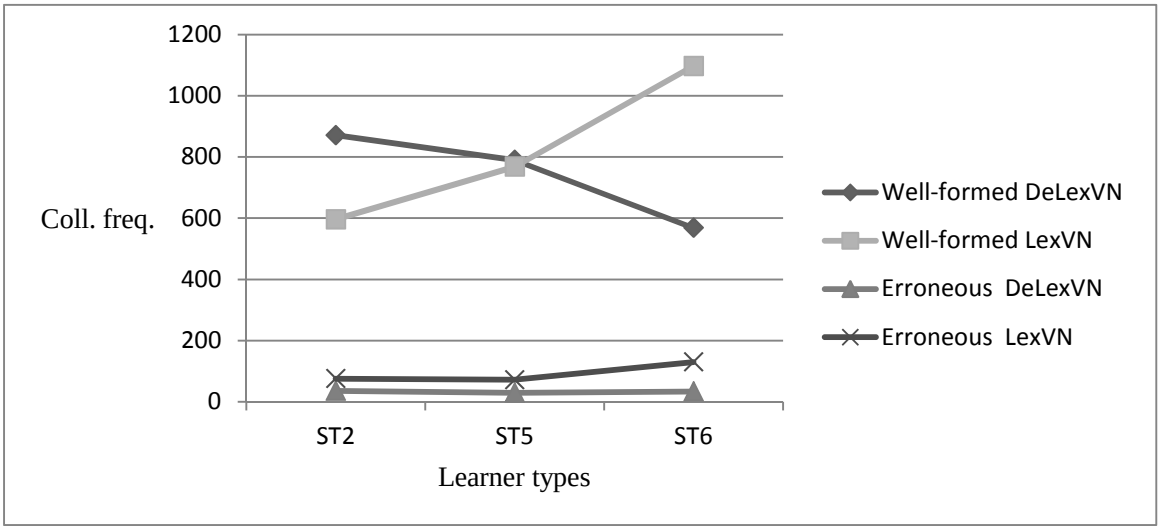


Figure 5-4 Well-formed and erroneous VN collocations in the three levels of learners (tokens)

Firstly, overall developmental patterns were revealed in the well-formed and erroneous DeLexVN and LexVN collocations. As is clearly shown in Figure 5-4, for collocations that are correctly produced, there is with rising proficiency a clear increase in lexical verb +noun collocations, and a decrease in delexical verb + noun collocations. In general, it can be interpreted to the effect that the learning of more lexical verbs leads to L2 learners' better production of VN collocations. This is a good sign of general vocabulary development as well as a development in collocational performance. However, it is interesting to note that at the same time L2 learners make more errors with LexVN collocations as proficiency rises, with DeLexVN collocation errors remaining the same in quantity. Investigation of collocation types rather than tokens in the three levels displayed the same developmental pattern (see Figure 5-5), and revealed a clear trend for the increase in erroneous lexical verb + noun collocations from the ST5 to the ST6 level.

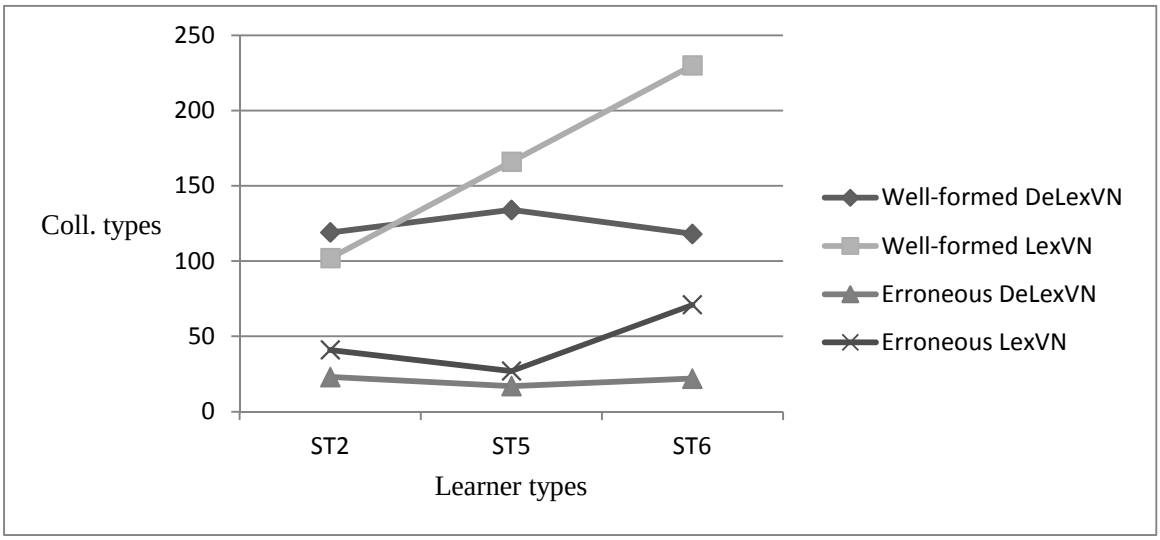


Figure 5-5 Well-formed and erroneous VN collocations in the three levels of learners (types)

The above figures show the overall trend on both the correct and erroneous uses of LexVN and DeLexVN collocations among different levels of learners. In addition to a comparison of frequencies, further statistical analyses were performed to see if these differences reached statistical significance. Analyses were subsequently carried out from two perspectives: comparisons of well-formed LexVN and DeLexVN collocations in the three databases and comparisons of erroneous collocations produced by these three levels of learners.

5.2.1 Between-group comparisons of well-formed DeLexVN and LexVN collocations

Table 5-4 below shows the overall tokens of well-formed delexical and lexical verb + noun collocations produced by the three levels of learners, and also gives the information regarding the ratio of DeLexVN collocations divided by LexVN collocations.

Table 5-4 Well-formed VN collocations produced by the three levels of learners (tokens)

Learners	DeLexVN	LexVN	DeLexVN / LexVN
ST2	871	596	1.5:1
ST5	790	768	1:1
ST6	568	1097	0.5:1

(Notes: ST2 and ST5: $\chi^2 = 22.57$, $p < 0.0001$ ****
ST5 and ST6: $\chi^2 = 90.20$, $p < 0.0001$ ****
ST2 and ST6: $\chi^2 = 199.3$, $p < 0.0001$ ****)

From this table we can see that for ST2 learners, DeLexVN collocations were used 1.5 times as often as LexVN collocations. For ST6 learners, this ratio dropped sharply to 0.5, meaning that the DeLexVN collocations produced by the ST6 level were only 0.5 times the number of LexVN collocations. Yet the ratio for the ST5 level was 1:1, indicating that they produced roughly equal numbers of DeLexVN and LexVN collocations. These ratios demonstrate a clear growth in the production of lexical verbs by ST6 learners.

Next, pairwise comparisons between the three groups were made using chi-square test with Yate's correction.[39] A significant relationship was found between the numbers of delexical verb + noun/lexical verb + noun collocations and learners at different proficiency levels. More specifically, ST5 learners produced very significantly more LexVN collocations than ST2 counterparts ($\chi^2 = 22.57$, $p < 0.0001$); Similarly, ST6 learners produced very significantly more LexVN collocations than the ST5 level ($\chi^2 = 90.20$, $p < 0.0001$); when the comparison was made between ST2 and ST6 learners, the ST6 level produced very significantly more LexVN collocations ($\chi^2 = 199.3$, $p < 0.0001$). These statistical analyses, together with the trend analyses presented in Figure 5-4, indicate that Chinese EFL learners' production of lexical verb + noun collocations increases with rising proficiency.

The above analyses were based on the overall collocation tokens. As has been noted earlier in Section 5.1.2, there was an uneven distribution of collocation tokens among types, so it is also necessary to consider the frequency of collocation types.

Table 5-5 presents the types of well-formed DeLexVN and LexVN collocations in the three databases and the ratios of DeLexVN collocations divided by LexVN collocations.

[39] This was chosen since the "Yates' continuity correction is designed to make the chi-square approximation better". (http://graphpad.com/guides/prism/6/statistics/index.htm?stat_chi-square_or_fishers_test.htm) [Accessed 10 June 2013]

Table 5-5 Well-formed VN collocations produced by the three levels of learners (types)

Learners	DeLexVN	LexVN	DeLexVN / LexVN
ST2	119	102	1.2:1
ST5	134	166	0.8:1
ST6	118	230	0.5:1

(Notes: ST2 and ST5: $p = 0.0416$ *
ST5 and ST6: $p = 0.0060$ **
ST2 and ST6: $p < 0.0001$ ****)

A similar analysis procedure was followed as in the above analyses of tokens. Firstly, in terms of ratios, the numbers of delexical verb + noun collocations decreased gradually as compared with those of lexical verb + noun collocations from the levels of ST2 to ST6, with the ST5 in the middle level. This confirmed to the trend as revealed in the analysis of collocation tokens. Secondly, statistical analyses were performed. Considering that the number of types produced by learners in our study was small in size, Fisher's test was used instead of chi-square test. Statistical significance was revealed between the production of LexVN collocations and learner types (ST2 and ST5: $p = 0.0416$; ST5 and ST6: $p = 0.0060$; ST2 and ST6: $p < 0.0001$). Like the comparison of tokens between ST2 and ST6, a highly significant difference was found in terms of collocation types, which means that Chinese EFL learners' production of lexical verb + noun collocations reliably increases with rising proficiency. The next section goes on to examine the growth trends for erroneous LexVN and DeLexVN collocations.

5.2.2 Between-group comparisons of erroneous DeLexVN and LexVN collocations

The analyses of well-formed delexical verb + noun and lexical verb + noun collocations showed a clear trend towards an increase in the production of lexical verb + noun collocations and decrease in delexical verb + noun collocations as L2 learners' proficiency rises. That indicates a growing collocational competence with the learning of more lexical verbs or nouns. The production of more LexVN collocations by more proficient learners seems unsurprising. It is natural that less proficient learners tend to resort to general words rather than words of specific meanings as constrained by limited vocabulary. However, as the learning of more lexical verbs facilitates better verb choices (e.g. *take* +

attitude, *have* + *attention*, *solve* + *problem* in the ST2, but *adopt* + *attitude*, *attract/catch* + *attention* and *solve/resolve/tackle* + *problem* in the ST6), it is not always facilitative since at the same time more lexical verb + noun collocations were found to be incorrectly used as learners become more proficient (see Table 5-6 and Figure 5-6 below).

Table 5-6 Erroneous VN collocations produced by the three levels of learners (tokens)

Learners	DeLexVN	LexVN
ST2	36	76
ST5	30	72
ST6	34	130

(Notes: ST2 and ST5: $p = 0.7671$ ns
ST5 and ST6: $p = 0.1398$ ns
ST2 and ST6: $p = 0.0354$ *)

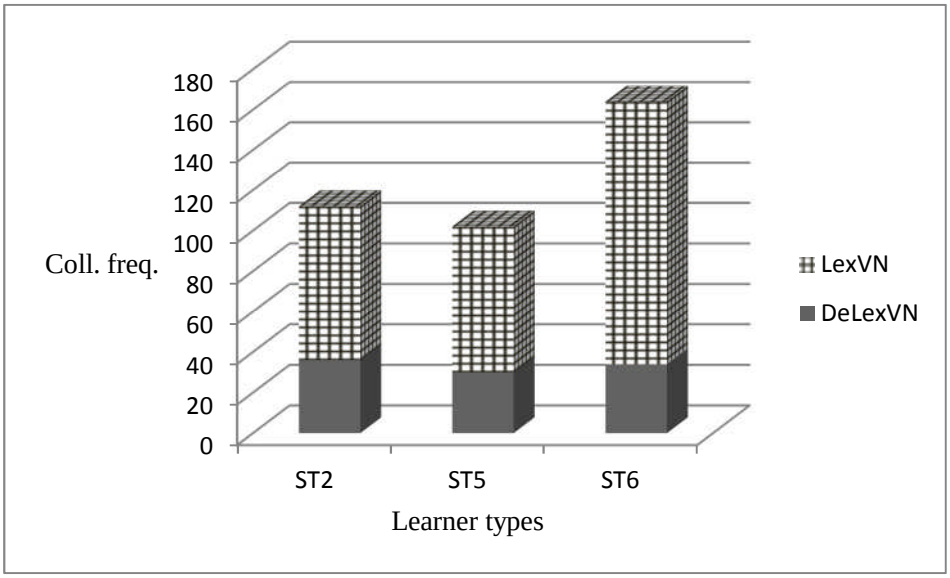


Figure 5-6 Erroneous VN collocations produced by the three levels of learners (tokens)

As is shown in the above graph, there was no increase in erroneous delexical verb + noun collocations but there was an increase in erroneous lexical verb + noun collocations from the ST2 to the ST6 level. Fisher's test further showed a strong trend for increasing lexical verb + noun errors in the ST6 level as compared with the ST2 level ($p = 0.0354$), though no significant difference was found between ST2 and ST5 levels, and ST5 and ST6 levels. As was discussed earlier in Section 5.1.3, the ST5 level stands out in terms of lower number of errors, than with ST2 and ST6 levels. Analyses of erroneous

collocation types were also carried out and results showed no significant difference between groups, although there was an increase in lexical verb + noun errors from the ST5 to the ST6 level (see Appendix Ⅰ).

5.2.3 Synopsis of Overall analyses (2)

This section is concerned with quantitative analyses of the production of well-formed and erroneous delexical verb + noun and lexical verb + collocations by the three groups of learners, with the aim to test whether in VN collocation learning lower levels of L2 learners make more errors with delexical verbs and higher levels make more errors with lexical verbs. This prediction was upheld. It was found that there was an increase in the ratios of erroneous lexical verb + noun collocations from the ST2 level to the ST6 level and ST6 learners produced significantly more errors with lexical verb + noun collocations than ST2 learners. Although for all groups of learners, there were more LexVN collocation errors than DeLex VN collocation errors, ST2 learners produced significantly more DeLexVN errors compared with the ST6 level and the latter produced significantly more LexVN errors.

In addition to the between-group comparisons of erroneous VN collocations, well-formed ones were also analysed. Results showed that there was a gradual increase in lexical verb + noun collocations and a decrease in delexical verb + noun collocations among the three levels of learners. A significant relationship was found between the numbers of LexVN collocations and learner groups, indicating a significantly higher production of lexical verb + noun collocations with the rise of proficiency. Combining this with the results obtained through analyses of erroneous VN collocations, we can see on the one hand, that the increase in lexical verbs facilitates diversified verb + noun collocation production, and on the other, that more errors are involved with lexical verb + noun collocations, even though the increase in lexical verbs means more and better choices for L2 learners in choosing the “right” verbs for the noun collocates. Therefore, in addition to the finding of a poorer performance on the part of learners in lexical verb + noun collocations as measured through the increase of verbs from delexical to lexical verbs in collocation production, there is a need for a detailed analysis of the relationship between lexical verb increase and collocation misuses. This detailed analysis will be performed in Chapter 6.

Another interesting finding in this section was the position of ST5 learners. In terms of overall collocation errors out of the total number of VN collocations, they produced the fewest erroneous

collocations among all levels (proportion of errors in the ST5: 13%; ST2: 22%; ST6: 21%). Meanwhile, if the errors are divided into erroneous DeLex VN and LexVN collocations, the ST5 level is still out of line with the other two levels in terms of the ratio of DeLexVN to LexVN errors. However, in terms of the well-formed VN collocations, the number of collocations produced by ST5 learners is around the average of those produced by ST2 and ST6 learners (ST5: 1,558; ST2: 1,467; ST6: 1,665). The same is true for the ratios of well-formed DeLexVN to LexVN collocations in the ST5 level (in terms of tokens: ST5: 1:1; ST2: 1.5:1; ST6: 0.5:1; in terms of types: ST5: 0.8:1; ST2: 1.2:1; ST6: 0.5:1). We can see that ST5 is in the transitional stage. Even though the ST5 level does not show a significant difference between the two levels in terms of DeLexVN and LexVN collocation errors, it is consistent with the trend for increasing use in LexVN and decreasing use in DeLexVN collocations. As shown in Figure 5-4, there is a downward trend in delexical verb + noun collocations and upward trend in lexical verb + noun collocations. So ST5 is just at the place for the trend to be monotonic. There is a slight progression in English proficiency between ST5 and ST6 learners since ST5 is just lower than ST6, who have not spent much more time on exposure in English than the ST5. However, there are significant differences between the ST2 and ST6 level both in the number of well-formed and erroneous DeLexVN collocations and LexVN collocations. The sharp difference between the lowest level (ST2) and the highest level (ST6) makes the comparison between these two levels more noteworthy. Based on this observation, comparisons in the following sections were mainly carried out between these two levels.

5.3 Overall analyses (3): verb growth and collocation errors

Up to this point, it can be seen that there is a significant increase in lexical verb collocations, either correctly or incorrectly produced as learners proceed to the advanced level. Before turning to the detailed analyses in Chapter 6 of verb increase in a particular semantic set and VN collocations associated with these verbs, an overall analysis was performed on the growth rate of all the verbs and nouns used by the three groups. The aim of this quantitative analysis is to get a panoramic view of vocabulary increase and its relationship with error growth. The aim is to compare the growth rate of lexical verbs and the rates of collocation errors, in order to see globally the interconnection of these two rates. Seen through the above finding that there was a gradual increase in lexical verb + noun collocations both in terms of tokens and types, it is predicted that there is a considerable increase in

lexical verbs and/or nouns.

Verbs and nouns were automatically retrieved through regular expressions performed by PowerGREP and then lemmatised by Wordsmith (cf. Section 4.6). Examples of lemmatised verbs are *go* (lemma), including *goes, going, gone, went*; *legalise* (lemma), including *legalised/legalized, legalises/legalizes, legalising/legalizing*.[40]

Altogether the frequencies of lemmatised verbs/nouns used by the three groups, their growth rates and growth rates of lexical verb + noun collocations are presented in Table 5-7.

Table 5-7 Growth rates of lemmatised verbs, nouns and LexVN collocations

	Verbs	Nouns	Well-formed (tokens)	Well-formed (types)	Erroneous (tokens)	Erroneous (types)
ST2	1473	3345	596	102	76	41
ST5	1771	3857	768	166	72	27
ST6	2049	4199	1097	230	130	71
Growth rates (ST2 to ST6)	39%	26%	84%	125%	71%	73%

The above table illuminates two interesting aspects. From the perspective of the quantities of lemmatised verbs and nouns, the transitional stage of ST5 was further confirmed (cf. Section 5.2.3). To be more specific, there is a gradual increase in verbs and nouns from the ST2 to ST6, and the ST5 is again in mid-position allowing the upward trend to be monotonic. It is not surprising that L2 learners learn more and more verbs and nouns with increasing exposure to English. Another interesting point we can observe here is regarding the growth rates of verbs and nouns in comparison with the growth rates of well-formed and erroneous LexVN collocations. With the learning of more verbs and nouns, the possibilities of combining them into well-formed collocations increase as well (growth rates of well-formed collocations: 84% and 125%). This can be interpreted to the effect that with the increase in lexical verbs in learners' overall vocabulary, there are more chances for them to locate the right lexical verbs to produce correct verb + noun collocations. At the same time, the learning of more nouns means more diversified combinations of lexical verbs into well-formed VN collocations. However, the chances that this vocabulary growth (for verbs: 39% and for nouns: 26%) may lead to errors increase as well,

[40] The reason why words were lemmatised before their growth rate was calculated is that the various forms of one word should be viewed as one word to avoid repetitive calculation. If all verb forms were included, the four forms of the verb *legalise* were instead counted as four words in the ST6. But as a matter of fact, for learners they've learnt only one verb – *legalise*.

seen through the high growth rates of erroneous lexical verb + noun collocations (71% in terms of tokens and 73% in terms of types). On the whole, these data suggest that the more learners learn, the more chance there is they will make mistakes.

In addition, Table 5-7 also shows a higher growth rate of verbs (39%) than nouns (26%). So the worsening collocation performance in lexical verb + noun collocations can be inferred as more linked to verb increments than noun increments. In what follows, detailed analysis of verb increase was conducted, in order to see collocation errors that are linked to verbs with an increase in a given synonym set. Furthermore, the fact that nouns increase 26% from the ST2 to ST6 suggests that learning nouns also plays a role in production of VN collocations. Considering this, a case study is conducted to consider the ratios of collocation errors associated with the learning of new nouns (see Section 6.3).

5.4 Synopsis of the overall analyses of verb + noun collocations

Sections 5.1, 5.2 and 5.3 set out to quantitatively examine the production of verb + noun collocations by three levels of Chinese EFL learners. Unlike most previous L2 collocation studies, this study started with an exhaustive extraction of both the well-formed and erroneous VN collocations, and these collocations were further divided into delexical verb and lexical verb collocations with a view to looking into verb vocabulary growth, from delexical verbs to lexical verbs. An apparent-time design was adopted, assuming that the performance of different age groups of learners at different proficiency level is indicative of a continuous developmental process. The above three sections presented the overall results obtained through general quantitative analyses and findings were as follows:

a. Results of the overall collocations (only around 5,000) out of over 600, 000 words of text support the findings of the ‘open choice principle’ employed by L2 learners in language production by comparison with NSs. Though it is difficult to compare the number of VN collocations relative to the size of writings by L2 learners due to the heterogeneity of L2 VN collocation studies, compared with previous findings with regard to NS performance, L2 learners produced a far smaller number of collocations. In addition, the overall number of collocation types (1,070) is rather low compared with tokens. Findings showed a heavy use of a rather small number of collocation types, i.e. less than 14% types of collocations taking up more than half of the collocations produced by Chinese EFL learners at all proficiency levels. These figures together indicate poor L2 phraseological competence, as

collocations are sparsely and repetitiously used by L2 learners.

b. Collocation overuse, as has been widely recognised, was discovered in this study and yet in a different way. Unlike previous studies, the finding of overuse was not based on comparisons of native and non-native data. Instead, comparisons of collocation production were performed within learner data in our study. A rather limited type of collocations were "overused" in the sense that they occurred more frequently than other collocation types, making up more than half of all collocation tokens. Therefore, learners' collocation interlanguage is characterised by a small number of frequent collocations making up large portions of all collocation uses.

c. Collocation misuses were found at all levels, with varying percentages. In general, nearly a quarter of the collocations produced by L2 learners were erroneous. The quantity of errors was analysed with regard to L2 proficiency, and statistical analysis showed that there was no significant difference between the numbers of erroneous collocations and learners of the ST2 and ST6 level. However, there were a persistent proportion of collocation misuses with the rise of proficiency (22% in the ST2 level and 21% in ST6 level). Rather than a decrease in errors as L2 learners' proficiency rises, collocation misuses remain at the same level, which indicates a lag in collocation acquisition.

d. In terms of the production of delexical verb + noun and lexical verb + collocations by the three groups of learners, an upward trend for well-formed lexical verb and downward trend for delexical verb collocations were found from the ST2 to the ST6 level. Further statistical analyses confirmed this trend by showing significant increase in well-formed lexical verb + noun collocations with the rise of proficiency.

e. Although there was a significant increase in well-formed lexical verb collocations, which indicates a rising collocational competence at the ST6 level, there were also significantly more lexical verb collocation errors at the ST6 level compared with the ST2 learners. It was observed that there was an increase in the ratios of erroneous lexical verb + noun collocations from the ST2 level to the ST6 level and lower levels of learners made more errors with delexical verbs and higher levels made more errors with lexical verbs. ST6 learners produced significantly more errors with lexical verb + noun collocations than ST2 learners in terms of collocation tokens.

f. The growth rates of verbs and nouns were measured and then compared with the growth rates of lexical verb + noun collocations. Results indicated that with the learning of more verbs and nouns, the possibilities of combining them into well-formed collocations increase, but errors increase sharply as

well. The huge increase of lexical verb collocation errors in the ST6 level as compared with the ST2 can be considered as an association more with the general growth in verbs than the nouns, given the more rapid growth of the verbs than the nouns.

In sum, from a developmental perspective, there emerges a complex developmental pattern of VN collocation knowledge, a quantitative progression but qualitative degradation in the development of Chinese L2 learners' knowledge of collocations. From the perspective of well-formed collocations, there is an increasing and more diversified production of collocations with the rise of proficiency. In spite of this positive sign of development, the percentage of collocation errors are still high at the ST6 level. They produced significantly more lexical verb errors than the ST2 level, suggesting a poorer performance on the part of learners in lexical verb + noun collocations as measured through the increase in verbs from delexical to lexical verbs in collocation production. General verb increase was found from the lowest to the highest level. It seems the more verbs/nouns learners acquire, the more they make errors. So there is a need for a detailed analysis of the relationship between lexical verb increase and collocation misuses. The next chapter will therefore be concerned with analysing the collocations of verbs as classified into synonym sets.

Chapter 6: Verb increase and the production of verb + noun collocations (2)

6.0 Introduction

The main research goal in the present study is to answer the question whether, within specific semantic domains of the verbs occurring in verb + noun collocations produced by all levels of learners, there are more chances of these verbs in higher levels to lead to collocational errors than they form the correct ones. Thus Section 6.1 presents detailed analyses of VN collocations where verbs were classified into synsets, aiming to investigate the relationship between verb increase and collocation errors. Section 6.2 gives a summary of the detailed analyses of verbs in synsets and learners' collocation performance with these verbs; Section 6.3 looks into whether there are other factors accounting for a lag in collocation, i.e. the acquisition of new nouns.

6.1 Detailed analyses – verb increase and collocation uses

The groups of learners first targeted were the ST2 and ST6, i.e. the lowest and the highest levels. The reasons why the ST5 level was not included as the first step in analysis are as follows: firstly, the ST5 level, as the transitional stage, produced the fewest erroneous collocations as compared with the other two levels. Results from statistical tests showed that they produced very significantly fewer VN collocation errors than both the ST2 and ST6 groups of learners (cf. Section 5.1.3). Secondly, analysis of erroneous lexical verb collocations in the ST5 file yielded the same result: they produced the fewest erroneous lexical verb collocations and no significant relationship was found between the ST5 level and the other two levels with regard to the number of LexVN errors produced. However, there was a strong trend for increasing LexVN errors at the ST6 level as compared with the ST2 level (cf. 5.2.2). Therefore, the increase in lexical verbs was sharper in the ST6 level as compared with the ST2 level than as compared with the ST5 level. So we started by classifying verbs in VN collocations in the ST6 level into synsets, then classified the lexical verbs in VN collocations in the ST2 level into synsets, and

compared VN collocations within these synsets between the two levels. Finally, verbs in the VN collocations in the ST5 level were added for general comparison with the other two levels, in order to see whether there is a consistent trend in learners' collocation performance within these synsets.

As was presented in Section 4.7.2, the criteria for classifying verbs in VN collocations were semantic similarity (synonyms), context-dependent synonyms and foreign-language equivalents. Three sources for determining semantic similarity were referenced: the *EVCA*, the *ODSA* and WordNet. Verbs in the VN collocations in the ST2 and ST6 databases were classified into synsets such as, verbs of creation (e.g. *compose, create, build*), and verbs of obtaining (e.g. *achieve, earn, receive*), etc. Verb + noun collocations produced by Chinese L2 learners were limited in quantity (cf. section 5.1.2), so the verbs in collocations were found to be infrequent. Due to the rather infrequent uses of verbs in collocations produced by EFL learners with proficiency ranging from the basic level to the advanced level, only a limited number of synsets that occurred in both databases were obtained (see Table 6-1 for the 16 synsets classified)

Table 6-1 Synsets occurring both in ST2 and ST6 VN collocation databases

Synsets		Verbs		
		ST2	ST6	No.
1	verbs of creation	compose, create, draw, hold, launch, raise, set	arouse, chart, build, draft, draw, enact, establish, form, hold, launch, publish, raise, set, stir	7
2	“fulfil” verbs	discharge, fulfil	accomplish, apply, carry out, commit, conduct, enforce, exercise, exert, fulfil, implement, perform, realise	10
3	verbs of obtaining	achieve, earn, gain, gather, grasp, receive	achieve, catch, earn, gain, grasp, reach, receive, seize	2
4	verbs of putting	lay	attach, fix, impose, lay, place, put	5
5	“settle” verbs	settle, solve	charge, settle, solve, resolve, tackle, undertake	4
6	“learn” verbs	know, learn, study,	acquire, learn, master, study	1
7	verbs of transfer of a message	teach, tell	impart, instruct, teach, tell	2
8	“keep” verbs	hold, keep	hold, keep, maintain	1
9	“follow” verbs	follow, obey	adopt, follow, obey	1
10	“play” verbs	play	act, play	1
11	“change” verbs	change	change, shift	1
12	“break” verbs	break	break, violate	1
13	“live” verbs	lead, live	lead, live	0
14	“wear” verbs	dress, wear	dress, wear	0
15	“drive” verbs	drive, ride	drive	-1
16	“pay” verbs	devote, pay	pay	-1

(Note: The ‘No.’ column represents the number of verbs at the ST6 level that are more than verbs at the ST2 level.)

As shown in the above table, there was an increase in verbs in the first 12 synsets, but the increase varied in different synsets. More dramatic verb increase at the ST6 level were found in the semantic sets of verbs of creation, *fulfil* verbs, verbs of putting and *settle* verbs than other synsets like verbs of obtaining, *learn* verbs, verbs of transfer of a message, *keep* verbs, *follow* verbs, *play* verbs, *change* verbs and *break* verbs. But for the last 4 synsets, there was no such increase in the quantity of verbs in the ST6 level. The proliferation of verbs in the higher level is a natural process as learners learn more words with more instruction they receive. The more verbs in a semantic field learners learn, the more specific they can be in expressing meanings. However, as is pointed out by Wolter (2006), L2 learning is not merely restricted to expanding vocabulary size: the depth of vocabulary knowledge is of equal

importance and one measure of vocabulary depth is the learning of syntagmatic connections between words. But as will be shown below, greater specificity was acquired at a loss for L2 learners. In the following detailed discussion of both the well-formed and erroneous VN collocations of these verbs in the 12 synsets, we set out to examine whether these verbs lead to more errors than they form correct collocations.

6.1.1 Analysis of VN collocations within synsets identified at the ST2 and ST6 levels

Both well-formed and erroneous VN collocations involving verbs in the16 synsets in Table 6-1 were respectively recorded at the ST2 and ST6 levels of learners. For classifying erroneous collocations, errors involving both the wrongly used verbs and the target verbs falling in the synsets were included. In other words, errors included not only the verbs within the synsets that were inappropriately produced, but also verbs that should be produced but not. For example, **create (compose) + song* was classified as a collocation error in the synset of verbs of creation, since the wrongly used verb (*create*) and the target verb (*compose*) have the semantics of *creation*. In addition, **make (compose) + poem* was also counted as a collocation error falling in the synset of verbs of creation, given that the target verb *create* was in the semantic field of *creation*. Detailed classification of well-formed and erroneous collocations of the verbs within these synsets in ST2 and ST6 is provided in Appendices Ⅱ and Ⅲ respectively. Analyses were performed on L2 learners' collocation performance in synsets with a verb increase (i.e. the first 12 synsets in Table 6-1) and synsets with no increase in verbs in the ST6 level (i.e. the last 4 synsets). Table 6-2 presents the total number of collocation types within two different kinds of synsets (for detailed information about the frequencies in each synset see Appendix Ⅳ). The frequency of tokens was not considered in the following analyses so as to avoid skewing the overall results, since there was an unbalanced distribution of tokens within a limited range of collocation types (cf. Section 5.1.2).

Table 6-2 Frequency of well-formed and erroneous VN collocations in the 16 verb synsets (ST2 and ST6)

	ST2		ST6	
	WFC	EC	WFC	EC
Synsets with a verb increase	39	22	126	65
Synsets with no verb increase	11	2	7	2

(Notes: 'WFC' stands for well-formed verb + noun collocations; 'EC' for erroneous verb + noun collocations.)

Within the 12 synsets where there was an increase in verbs in the ST6 level (e.g. synsets of verbs of creation, *fulfil* verbs, etc.), well-formed VN collocations increased dramatically from 39 to 126 in frequencies. However, there was also an increase in collocation errors from 22 in the lowest level to 65 in the highest level. In contrast, among the 4 synsets where no increases in verbs were found in the ST6 level (e.g. synsets of *live* verbs, *wear* verbs, *drive* verbs and *pay* verbs), collocation errors remained constant from the ST2 to the ST6 (2 types of errors in total in each level). In terms of proportions, the percentage of erroneous collocations involving verbs in synsets with a verb increase out of the total number of collocations produced by ST2 learners was 36% (22/(39+22)), and for ST6 learners, the percentage was 34% (65/(126+65)). The percentage of collocation errors that ST6 learners made in synsets with a verb increase was roughly the same as that in the ST2 level. This finding indicates a lag in collocational knowledge for more proficient learners. More precisely, even though ST6 learners were more advanced and acquired more lexical verbs, they were as likely to make verb + noun collocation errors as much less proficient learners (ST2 learners). There was no sign of an improving competence on VN collocations with the rise of proficiency.

Not only was a lag found in learners' collocation performance in synsets with an increase in verbs, the occurrence of collocation errors involving these synsets in the ST6 was found to be more limited to elaborated synsets than it was in the ST2 level. The total number of erroneous collocations produced by the two groups of learners respectively were 64 (ST2) and 93 (ST6) (cf. Table 5-2). So the proportion of collocation errors associated with verbs in the 12 synsets in ST2 was 34% (22/64). For ST6 learners, the ratio was twice as high as that of ST2 learners – 70% (65/93). An increase in erroneous collocations in these synsets was found. Again, this gross analysis of collocation errors out of the total number of errors indicated that the more verbs that were learned by higher levels, the more collocation errors were produced.

What has been found up to now supports the general prediction that verb increase is a factor responsible for the stagnant development of collocational knowledge. However, caution is needed here, since collocation errors in the above analyses include both verbs that are old, i.e. verbs produced by the ST2 level in VN collocations, and verbs that are new in the ST6 level, i.e. newly learned verbs that were not found in ST2 VN collocation databases. The 65 collocation errors produced by ST6 learners involve both errors with old verbs and new verbs. For example, given that the verb *draw* in verbs of creation has been used by the lower level (e.g. *draw + conclusion*), it was considered as an already-acquired verb for learners at higher levels. Similarly, *conduct* was not used by ST2 learners in VN collocations but was present in the ST6 level, so it was considered as a new verb. In the calculation of erroneous verb + noun collocations in Table 6-2, errors involving both the old verb (**make (draw) + conclusion*) and the new verb (**conduct (commit) + crime*) were included. Therefore, ST6 learners' collocation performance on old verbs and new verbs should be distinguished, in order to look at whether new verbs are associated with more errors than they form correct collocations.

In the process of distinguishing errors associated with old verbs and new verbs in the ST6 level, the following criteria were adopted: if errors involved new verbs (e.g. **publish (enact) + law, publish* was a new verb in the ST6), they were put in the category of errors with new verbs; if the error involved old verbs, but the target verb was a new verb (e.g. *make (conduct) + exam, conduct* was a new verb in the ST6), it was classified as errors with new verbs; if the error involved old verbs (e.g. **draw (formulate) + theory, draw* is an old verb for the ST6 level), but the target verb was not a new verb in the synsets identified, it was considered as an error with old verbs. Following these criterion, VN collocations associated with the verbs in the sysnsets identified were divided into those with old verbs and new verbs. Examples of well-formed VN collocations associated with old verbs are: *launch + war, set + fire*; examples of well-formed VN collocations associated with new verbs are: *chart + course, draft + law*; examples of collocation errors associated with old verbs are **make (draw) + conclusion, *take (launch) + career*; examples of errors with new verbs are **arouse (cause) + trouble, *take (conduct) + survey*. The frequency information of collocation errors, divided into errors with old and new verbs in the 12 synonym sets, is tabulated in Table 6-3 below.

Table 6-3 VN collocation production involving old and new verbs at the ST6 level

	Synsets	Old verbs			New verbs		
		Verbs	WFC	EC	Verbs	WFC	EC
1	verbs of creation	draw, hold, launch, raise, set	13	4	arouse, chart, build, draft, enact, establish, form, publish, stir	6	11
2	“fulfil” verbs	fulfil	3	1	accomplish, apply, carry out, conduct, enforce, exercise, exert, implement, perform, realise	23	16
3	verbs of obtaining	achieve, earn, gain, grasp, receive	18	2	catch, reach, seize	6	6
4	verbs of putting	lay	2	2	attach, fix, impose, place, put	12	6
5	“settle” verbs	settle, solve	2	0	charge, resolve, tackle, undertake	4	1
6	“learn” verbs	learn, study	0	4	acquire, master	1	1
7	verbs of transfer of a message	teach, tell	3	1	impart, instruct	1	2
8	“keep” verbs	hold, keep	12	2	maintain	3	0
9	“follow” verbs	obey, follow	4	1	adopt	4	0
10	“play” verbs	play	2	2	act	0	1
11	“change” verbs	change	1	1	shift	1	0
12	“break” verbs	break	3	1	violate	2	0
Total			63	21		63	44
			84			107	
ER			25%			41%	

(Notes: ‘WFC’ stands for well-formed verb + noun collocations; ‘EC’ for erroneous verb + noun collocations; ‘ER’ represents the ratio of the errors out of all the collocations examined in the column.)

As is shown in Table 6-3, the overall number of well-formed collocations with old verbs and new

verbs did not show an increase, but errors involving new verbs increased sharply. The error percentage associated with new verbs out of the number of their collocation uses is 41%, while that of old verbs is only 25%. Apart from a comparison in percentages, further statistical analysis was performed. Fisher's test revealed a significant difference between old and new verbs in terms of the number of erroneous collocations (p = 0.0216; see Table 6-4 below). In other words, collocation errors involving new verbs are significantly more likely than errors with old verbs.

Table 6-4 Collocation uses involving old verbs and new verbs at the ST6 level

	Well-formed coll.	Erroneous coll.	Total
Old verbs	63	21	84
New verbs	63	44	107

(Note: p = 0.0216 *)

Turning now to the synsets where L2 learners had more problems with new verbs than with old verbs, it becomes clear from Figure 6-1 below that errors with new verbs falling into the semantic domains of verbs of creation, *fulfil* verbs, verbs of obtaining, verbs of putting, *settle* verbs and verbs of transfer of a message occurred more often than with old verbs. This result is strongly linked with synset classification in which there is a proliferation of verbs just within the six synsets (cf. Table 6-1). Therefore, verb increase as an inhibiting factor on target-like L2 collocation performance is again supported.

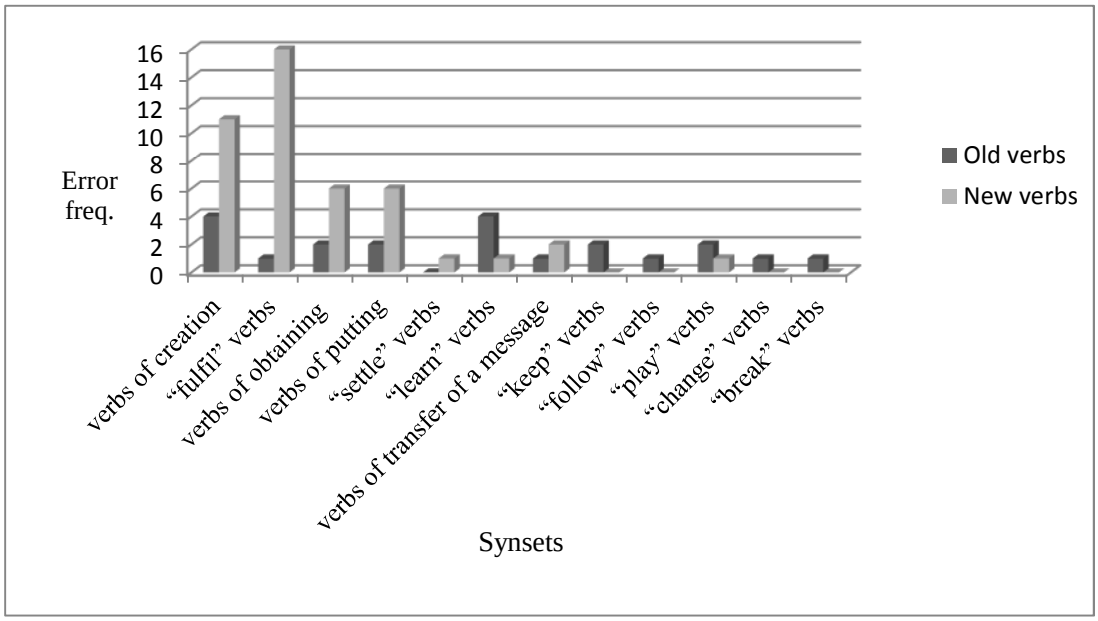


Figure 6-1 Collocation errors involving old and new verbs in the ST6 synsets

6.1.2 Analysis of VN collocations within synsets identified at the ST2, ST5 and ST6 levels

The previous section has addressed the relationship between verb increase and collocation errors in the lowest (ST2) and highest (ST6) levels and quantitative analysis shows: 1) errors in the synsets with more verbs learnt in the ST6 level than the ST2 level increased with rising proficiency; 2) errors with new verbs were significantly more likely than with old verbs. In this section, the focus shifts to the middle level (ST5), so as to see whether there is a consistent trend in the ST5 level in terms of both verb increase in the 12 synsets and collocation uses linked with these verb synsets. It is predicted that the performance of ST5 learners is consistent, i.e., compared with the ST2 level there is an increase in verbs in the synsets and at the same time an increase in collocation errors associated with the verbs in the synsets, and there are fewer verbs and collocation errors, compared with the ST6 level.

All the lexical verbs in the verb + noun collocations in the ST5 database were classified into synsets following the same procedure and criteria applied in the analyses of the other two levels. Table 6-5 lists the classification of verbs in the ST5 level together with the verbs in synsets identified in the ST2 and ST6 levels.

Table 6-5 Verb synsets classified from ST2, ST5 and ST6 VN collocation databases

Types	Synsets	Verbs		
		ST2	ST5	ST6
1	verbs of creation	compose, create, draw, hold, launch, raise, set	arouse, build, conduct, draw, establish, form, hold, launch, produce, publish, raise, set	arouse, chart, build, draft, draw, enact, establish, form, hold, launch, publish, raise, set, stir
2	“fulfil” verbs	discharge, fulfil	apply, enforce, fulfil, perform, practice	accomplish, apply, carry out, commit, conduct, enforce, exercise, exert, fulfil, implement, perform, realise
3	verbs of obtaining	achieve, earn, gain, gather, grasp, receive	achieve, catch, earn, gain, grasp, reach, receive, seize	achieve, catch, earn, gain, grasp, reach, receive, seize
4	verbs of putting	lay	attach, lay, place, put, set	attach, fix, impose, lay, place, put
5	“settle” verbs	settle, solve	resolve, solve	charge, settle, solve, resolve, tackle, undertake
6	“learn” verbs	know, learn, study	learn, master, study	acquire, learn, master, study
7	verbs of transfer of a message	teach, tell	teach, tell	impart, instruct, teach, tell,
8	“keep” verbs	hold, keep	hold, keep	hold, keep, maintain
9	“follow” verbs	follow, obey	adopt, follow, obey	adopt, follow, obey
10	“play” verbs	play	play	act, play
11	“change” verbs	change	change	change, shift
12	“break” verbs	break	break, violate	break, violate
13	“live” verbs	lead, live	lead, live	lead, live
14	“wear” verbs	wear, dress	wear	wear, dress
15	“drive” verbs	drive, ride	ride	drive
16	“pay” verbs	devote, pay	pay	pay

In terms of the variety of verbs falling in the first 12 synsets, the ST5 level falls in the middle between the lower level (ST2) and the higher level (ST6). In each of the 12 synsets, they produced

verbs no more than the higher level and no less than the lower level. A clear and consistent trend for verb increase is shown from the above table. More and more verbs were learned in the semantic domains of verbs of creation, *fulfil* verbs, verbs of obtaining and verbs of putting. On the whole, the quantity of verbs in the ST5 level is more like the verbs produced by ST2 learners, since in the 6 synsets of s*ettle* verbs, *learn* verbs, verbs of transfer of a message, *keep* verbs, *play* verbs, *change* verbs and *live* verbs, the ST5 level shows no increase in verbs compared with the lowest level.

Collocations with the verbs in the 16 synsets in the ST5 level were divided into correct and erroneous uses, as shown in Appendix Ⅴ. Then similar analyses of well-formed and erroneous collocation uses associated with verbs in the 12 synsets were performed on the ST5 data (see the numerical presentation of collocation uses associated with the 12 synsets in the three levels in Table 6-6, and for the detailed frequency information, see Appendix Ⅵ).

Table 6-6 Well-formed and erroneous VN collocations in the 16 verb synsets (ST2, ST5 and ST6)

	ST2			ST5			ST6		
	WFC	EC	Total	WFC	EC	Total	WFC	EC	Total
The first 12 synsets	39	22	61	61	22	83	126	65	191
The last 4 synsets	11	2	13	6	1	7	7	2	9

(Note: ‘WFC’ stands for well-formed VN collocations, and ‘EC’ for erroneous VN collocations.)

For the last four synsets in which higher levels did not produce more verbs than the lower levels, no increase in well-formed and erroneous VN collocations was found from the lowest to the highest level (well-formed collocations: from 11 to 6 to 7; errors: from 2 to 1 to 2). In contrast, the overall figures for well-formed and erroneous VN collocations in each proficiency group showed that within the synsets where there is a verb increase (the first 12 synsets), more and more well-formed collocations (from 39, to 61 and then to 126 types) were produced; at the same time errors increased as well from the ST2 to the ST6 level (from 22 to 65 types). There was not an error increase from the ST2 to the ST5 level, which may be due to the slight verb increase in sysnets in the ST5 level compared with the ST2 level. In addition, in terms of proportions, the percentage of collocation errors involving the 12 synsets in the ST5 was the lowest: 27% (for ST2: 36%; ST6: 34%) (see Table 6-7 below). When the numbers of errors are placed into the bigger context of the total collocation errors in

each level, there is a clear increase in the proportions of errors involving verbs in the 12 synsets.

Table 6-7 Proportions of VN collocation errors associated with the 12 synsets with a verb increase

Levels	Errors in synsets	Total errors	Total colls. in synsets	ER^1	ER^2
ST2	22	64	61	34%	36%
ST5	22	44	83	50%	27%
ST6	65	93	191	70%	34%

(Notes: 'colls.' stands for collocations; 'ER^1' represents the ratio of errors in synsets out of the total number of collocation errors in each level; 'ER^2' represents the ratio of errors in synsets out of the total number of both well-formed and erroneous collocations involving the 12 synsets.)

As Table 6-7 reveals, the ratios of errors out of VN collocations associated with the synsets identified did not show much decrease with rising proficiency, indicating a general lag in collocational knowledge. In addition, out of the total number of VN collocation errors, collocation errors within the 12 synsets increased markedly in proportions. Again this trend conforms to the trend that has been discussed earlier, i.e., verb + noun collocation errors produced by the learners became more and more limited to the synsets with an increase in verbs, as learners' proficiency rises. If these errors are localised in each synset (see Figure 6-2 for the depiction), a clear error increase can be seen at the ST6 level, which encompasses the most verbs in the 12 synsets. This increase is most manifested in the first four synsets, which see the greatest increase in verbs from the ST2 to the ST6 levels (cf. Table 6-1).

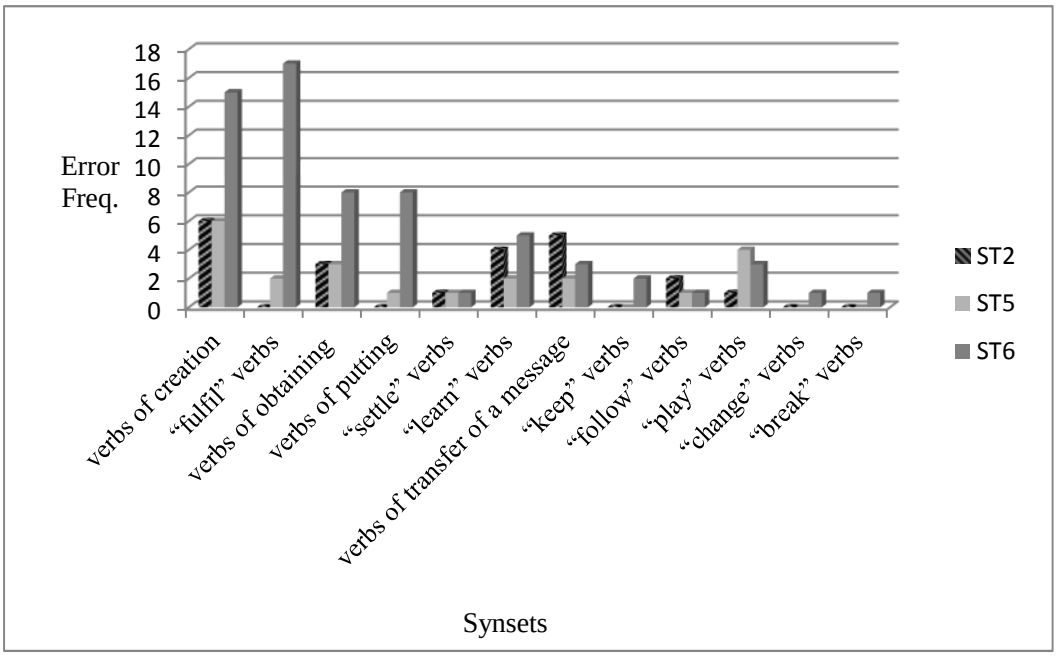


Figure 6-2 VN collocation errors with the verbs in the twelve synsets across the three levels

Up to this point, synsets that were particularly problematic for L2 learners emerge. As can be seen from the above figure, verbs of creation, *fulfil* verbs, verbs of obtaining and verbs of putting pose particular problems for advanced Chinese learners of English. Previous research has shown that they are also susceptible to errors for advanced German-speaking learners of English. Examining the erroneous VN collocations produced by German advanced learners of English, Nesselhauf (2005) reported similar semantic verb groups where L2 learners had particular problems. As in our study, the verbs wrongly used included those inappropriately produced and those that were intended but not produced. One group of semantically related verbs susceptible to error is what we classified here as verbs of obtaining, though Nesselhauf did not provide an umbrella term. However, she gave a list of the verbs: *achieve, reach, acquire, obtain* and *gain*. Both learners in her and our studies had considerable difficulty with the right collocational choice. For example, **reach* + *recognition* was wrongly used instead of *receive* + *recognition*; and **get* + *conclusion* instead of *reach* + *conclusion.* Another verb group with which Nesselhauf reported learners' difficulty was verbs meaning 'carrying something out', which in our study were the *fulfil* verbs encompassing the largest number of errors for the ST6 learners. Both the German and Chinese learners were confused in choosing the correct verbs: e.g. *accomplish* and *conduct* were produced instead of *commit* in collocation with *crime*; delexical verbs like *do* and *make* were also wrongly chosen for the noun *crime*; *fulfil* + *plan* was produced rather than *implement* + *plan*; and **implement* + *act* was wrongly used instead of *perform* + *act*, etc.

In addition, Nesselhauf reported that the verb group comprising verbs such as *create, establish, set* and *set up* also posed particular problems for learners and this group of verbs was here classified as verbs of creation. Examples of errors within this sysnset are **publish + law* instead of *enact + law, *stir + consciousness* instead of *raise + consciousness, *raise + discussion* instead of *arouse + discussion.*

In general ST6 learners struggled in choosing the right lexical verbs from sets of semantically related verbs, and in a few cases delexical verbs were used instead of lexical verbs, e.g., **make + exam (conduct), *take + survey (conduct), *build + regulation (enact), *build + tie (establish), *make + conclusion (draw), *take + military service* (*perform*). However, ST2 learners were more likely to use delexical verbs instead of lexical verbs within the synsets classified in this study. For example, among the 22 VN collocation errors produced by ST2 learners in the synsets, a third were found to be linked with delexical verbs, e.g. **give + meeting, *make + result, *do + problem, *get + lesson*, etc.

Meanwhile, in the following cases of collocations, the erroneous verbs were observed to share certain phonological resemblance to the target verbs, e.g. the wrongly used verb – *arise* – and the target – *arouse* – in collocation with *discussion, draw* and *draft* followed by *law*. These instances can be seen as a result of phonological interference, which is consistent with word association studies showing that learners' responses tended to have a phonological similarity to stimulus words whilst native speakers produce associations based on syntagmatic and paradigmatic relations (Meara, 1978; 2009).

Even in synsets where no errors were found in the ST2 level, e.g. *fulfil* verbs, verbs of putting, *keep* verbs, *change* verbs and *break* verbs, this does not necessarily mean a full acquisition of verbs in these sets (e.g. verbs like *discharge, fulfill, lay*). Instead, it indicates the limited vocabulary that ST2 learners have. Although a limited vocabulary may lead to success in choosing the right verbs for a noun in certain collocations, it may cause imprecision in collocation production when the need for using other collocations arises. However, with the expansion of vocabulary size, more chances of incorrect production arise as learners are faced with more difficulty choosing the right verb from a set of semantically related verbs.

6.2 Synopsis of detailed analyses of verb increase and collocation uses

Through the above detailed analyses of verb increase within a certain semantic set and their

relationship with collocation performance, the development of learners' VN knowledge was observed to stagnate. On the one hand, the percentages of errors involving the 12 synsets identified from learners' production of VN collocations remained roughly constant (36% at the ST2 level and 34% at the ST6 level) with rising proficiency. In addition, errors with new verbs in the ST6 level were significantly more likely to be made than old verbs. On the other, the ratios of collocation errors with the verbs in the synsets out of the total number of errors increased successively from the lowest to the highest level (ST2: 34%; ST5: 50% and ST6: 70%). As learners proceeded to more advanced levels, the occurrence of collocation errors became more and more limited to synsets with a verb increase. Verb classes most susceptible to errors were verbs of creation, *fulfil* verbs, verbs of obtaining and verbs of putting, where there was a considerable increase in the number of verbs. Thus we consider that the increase of verbs in a particular semantic domain is an inhibiting factor for the learning of VN collocations.

There is a view that collocations defeat even the most proficient non-native speaker because of the arbitrary restrictedness in collocations (Nesselhauf, 2003; 2005). Collocations are not just semantically motivated, but also involve arbitrarily restricted selection. For example, *blonde hair* in English is felicitous, but **blonde paint* is not and *auburn hair* is used to describe women, but not men (Schmitt and Carter, 2004: 14). The restrictedness nature of collocation has been considered as the most important factor correlating with learners' difficulties with collocation production (Nesselhauf, 2003; 2005).

In light of our findings, what poses great difficulties for L2 learners is to distinguish among a group of semantically related verbs (e.g. *perform* vs. *implement*, *conduct* vs. *commit*, etc.). In the erroneous collocation **conduct + crime* instead of *commit + crime*, it can be seen that they both share the semantic features of 'carry out something', but differ from each other in the sense that "commit" denotes 'doing something illegal or bad', but "conduct" means 'organising something and carry it out'.[41] Likewise, in the following collocation error: **implement + act*, the learner who has made such an error may know a partial meaning of the verb *implement* (i.e. the semantic component as 'carrying out something') but not its complete semantics. This partial acquisition led to the learner's incorrect belief that the erroneous verb *implement* can be combined with the noun *act*. However, *implement* means more than that. The misused verb (*implement*) and the target verb (*perform*) both belong to the

[41] The meanings of the two verbs were quoted from *Collins COBUILD Advanced Learner's English Dictionary* (2006).

semantic field of *fulfil* verbs and verbs in this set have a small number of semantic features in common, but they are distinguished by specific meanings. Both verbs have the semantic component "carry out something", but *implement* is distinguished from *perform* in that it implies: 'to ensure what has been planned is done' (e.g. *implement a plan*). For the verb *perform*, it simply suggests 'doing a (usually) complicated task or action'. So it is inferred that when a learner has an incomplete command of *implement*, i.e. only the semantic component as 'carrying out something', but not its distinctive feature of 'ensuring something that has been planned is completed', collocation errors like *implement + act* are likely to be made. Therefore, seen from the erroneous VN collocations produced by Chinese L2 learners, only a fraction of verb semantics was acquired by learners, but not its distinguishing features from a set of semantically related verbs. In this sense, acquisition of verb semantics is important for successful learning of collocations.

6.3 An alternative explanation: new nouns and collocation uses

The research hypothesis tested above was that verb increase is the main factor responsible for the stagnant L2 collocation development. Accordingly, it is predicted that other factors, e.g. the acquisition of new nouns is not the main inhibiting factor in collocation performance. Our prediction is thus that in the majority of new nouns produced by higher levels of learners, learners produce correct VN collocations. The prediction that noun increase is not the inhibiting factor would be further confirmed if the percentage of new nouns in erroneous collocations remained constant within the levels of ST5 and ST6.[42]

In order to examine whether collocation lag is a result of new noun acquisition, specifically, whether collocation errors are made because learners learn a large proportion of new nouns, the empirical requirement was to identify newly acquired nouns by L2 learners. It was not feasible to see if a noun was new or old through asking the learner him/herself at the time of their writing. The identification of new nouns in a higher level was therefore implemented by examining nouns produced by lower levels of learners. New nouns were those which were not used by lower levels (i.e., with no occurrences in the file at lower levels) whilst old nouns referred to those that were both used by the two

[42] Given the constraint of locating new nouns in the lowest ST2 level, the proportion of collocation errors where new nouns occur cannot be obtained.

groups of learners (i.e. with occurrences in both the files of the two groups). Taking the ST2 and ST6 learners for the purposes of illustration, the nouns only occurring in the ST6 file were assumed to be newly acquired nouns by ST6 learners and old ones were those that occurred in both ST6 and ST2 sub-corpora.[43]

The search for new nouns was performed automatically. With the ST2 and ST6 as examples, procedures involving the identification of new nouns in the collocations produced by ST6 learners were as follows:

a. Store all the nouns in the VN collocations produced by ST6 learners in a text file;
b. Generate a list of all the nouns in the ST2;
c. Use Wordsmith (the *Match* function) to delete all the matched nouns between the wordlist of nouns in ST6 collocations and the nouns in ST2, and get a list of new nouns in the ST6.

Once the above procedure had been carried out, analyses of new nouns were first performed at the ST6 level (new nouns as new compared with the ST2 level), then new nouns (new nouns as new compared with the ST5 level) in the ST6 were analysed and finally new nouns in the ST5 level (new as compared with the ST2 level) were analysed.

(1) New nouns in the ST6 VN collocations (new as compared with the ST2 level)

Altogether there were 264 nouns in the ST6 VN collocations, out of which 72 were new nouns that did not appear in the ST2 file, and 192 old nouns that were used both by ST2 and ST6 learners. A further categorisation was carried out among the new nouns that occurred in the erroneous verb + noun collocations and correct collocations. The results are presented in Table 6-8.

[43] This can only be assumptions, since it might be as well that some new nouns in ST6 were actually acquired by the ST2 but not used (e.g. *button, decision*) or among old nouns in the ST6 and ST2, some were idiosyncratic uses by one learner (e.g. *criminal, principle*) and not acquired by general ST2 learners. These cases did exist but were rare, so assumptions were on the whole justified. In addition, since groups of learners were targeted, the individual differences between learners could not be spotted.

Table 6-8 New nouns and old nouns in ST6 VN collocations (new nouns as compared with ST2)

	New nouns	Old nouns	Total
Erroneous coll.	18 (25%)	42	60
Correct coll.	54 (75%)	150	204
Total	72(100%)	192	264

As is shown in the above table, 25% of the overall number of new nouns in ST6 VN collocations were in erroneous collocations, which means that three quarters of these newly acquired nouns were in correct collocations. It then becomes interesting to see whether collocation errors occur because of a lack of new verbs for newly acquired nouns or whether it is just a matter of learners' inability to associate the new nouns with already acquired verbs. If the first case, the inhibiting role of the new nouns will be manifested, as learners encounter two difficulties: newly acquired nouns and a lack of appropriate verbs. If the collocating verbs for the new nouns are already acquired but not correctly used with the new nouns, we can infer learners' split learning of collocations into individual words, rather than association of a newly acquired noun with an old verb. The 18 new nouns and their collocating verbs were analysed in the erroneous verb + noun collocations produced by ST6 learners, shown in Table 6-9 below.[44] Table 6-9 also provides information concerning whether or not the erroneous and the target verb collocates of the new nouns are present in the ST2 file. If the target verb (e.g. *impose*) was not present in the ST2 file (signaled by a minus symbol "-"), this verb was considered as a new verb for the ST6 learner. Otherwise it was an old verb (e.g. *play*).

[44] Among the 18 nouns, there is a noun phrase – *military service*, which was regarded as one noun for the convenience of analysis.

Table 6-9 18 new nouns in ST6 erroneous VN collocations and their verb collocates (new nouns as compared with ST2)

Nouns	Erroneous verbs	Target verbs	Erroneous verbs in ST2	Target verbs in ST2
burden	give	impose	+	-
burden	lay	impose	+	-
burden	release	relieve	-	-
role	lead	play	+	+
role	lay	assign	+	+
role	act	play	+	+
role	serve	play	+	+
consciousness	stir	raise	-	+
disadvantage	surpass	outweigh	+	-
regulation	break	violate	+	-
regulation	build	enact	+	-
survey	take	conduct	+	-
threat	do	pose	+	-
threat	impose	pose	-	-
murder	conduct	commit	-	-
prejudice	reflect	hold	-	+
prejudice	cast	hold	-	+
treaty	draw	sign	+	+
chat	make	have	+	+
competence	exert	demonstrate	-	+
imagination	cause	excite	+	-
load	pull	carry	+	+
mercy	cast	have	-	+
recognition	reach	receive	+	+
military service	attend	perform	+	+
military service	take	perform	+	+
measure	make	take	+	+

(Note: '+' means that the verb appears in the ST2 sub-corpus, and '-' represents an absence of the verb in the ST2 sub-corpus.)

To see whether ST6 learners wrongly used a new verb instead of an old verb to collocate with a new noun, or they used an old verb instead of another old/new verb for the new noun, the 27 collocation pairs of the 18 new nouns in ST6 erroneous collocations displayed in the above table were further classified into three categories, according to whether the erroneous and target verbs were already learnt by ST2 learners:

a. New nouns in the erroneous VN collocations where the target verb was absent in lower levels of learners (11 VN collocations)

Instances of this category are **give + burden, *lay + burden, *release + burden, *surpass + advantage, *break + regulation, *build + regulation, *take + survey, *do + threat, *impose + threat, *conduct + murder*, and **cause + imagination*. That the target verbs for these erroneous collocations (e.g. *impose, relieve, outweigh*, etc.) were not used by ST2 learners suggests that they were new to ST6 learners or had not been fully acquired yet. With *give + burden as an example, the target verb *impose* was not used by ST2 learners, suggesting *impose* may be a new verb to the ST6 learner who had acquired a new noun – *burden*. Thus learners were found to make collocation errors as such by using a verb they have already acquired (in the case of *burden*, they used *give* and *lay*). Among the 11 collocations where collocation errors may take place as a result of a lack of new verbs, 8 of the erroneous verbs in the VN collocations are old verbs with appearance in the ST2 file, meaning they may have been already acquired by ST6 learners. This is natural given that L2 learners haven't acquired the target new verbs and have to use old verbs instead. Learners acquire a new noun, but they may not acquire the collocating verbs, and collocation errors may thus occur. In the other three instances where new verbs have not been acquired, newly acquired verbs were incorrectly used with the new nouns, i.e. **release + burden, *impose + threat, and *conduct + murder.*[45]

In all, the 11 types of erroneous collocations in ST6 learner group can be viewed as a lack of new collocating verbs with newly acquired nouns.

b. New nouns in the erroneous VN collocations where the target verb was present in lower levels of learners but the erroneous verbs was absent (5 VN collocations)

As shown in Table 6-9, this category includes **stir + consciousness* instead of *raise + consciousness, *reflect/cast + prejudice* instead of *hold + prejudice, *exert + competence* instead of *demonstrate + competence* and **cast + mercy* instead of *have + mercy*. That the target verbs were used in lower levels of learners suggests that the verbs might be already known to ST6 learners, and yet they used newly acquired verbs to collocate with the newly acquired nouns.

c. New nouns in the erroneous VN collocations where both the erroneous and target verbs were present in lower levels of learners (11 VN collocations)

[45] The target verbs for the three new nouns (*relieve, pose and commit*) and new erroneous verbs (*release, impose, conduct*) all share partial phonological resemblance, which could be seen as phonological interferences and the target verbs may have been known by learners but not fully acquired yet.

This type arises when learners acquire a new noun, and misuse a known verb with another already acquired verb. Erroneous collocations involving misuses of old verbs include **lead + role* (correct verb: *play*), **lay + role (assign), *act + role (play), *serve + role (play), *draw + treaty (sign), *make + chat (have), *pull + load (carry), *reach + recognition (receive), *attend + military service (perform), *take + military service (perform), *make + measure (take)*. Errors of this type can be inferred as a split learning of VN collocations, i.e. new nouns were learnt in an isolated way instead of being learnt in collocational relationships with already learnt verbs.

In all, the percentage of new nouns that are linked to collocation errors in the ST6 is 25%, which means that among 100 nouns that are newly acquired by L2 learners, learners correctly find a collocating verb in 75 of the cases. Even in errors involving the wrong choices of verbs for the newly acquired nouns, less than a half of the errors involving new nouns (41%: 11/27) arise when learners do not know the new target verb collocates (as is illustrated in the Category a.). However, in more than half of the cases (59%: 16/27), errors arise when the target verb may have already been acquired by ST6 learners for the newly acquired nouns (as is illustrated in Categories b and c). These figures can be interpreted to the effect that collocation errors with new nouns occur even though learners in most cases do not lack the verbs for newly acquired nouns. For example, either new verbs are misused instead of an old verb (e.g. *stir + consciousness instead of *raise, *reflect + prejudice* instead of *hold*) or another old verb is misused instead of another old target verb (e.g. **lead + role* instead of *play, *make + measure* instead of *take*). On the one hand, that learners use newly acquired verbs to combine with newly acquired nouns can be viewed as boldness in collocation learning, i.e. they are experimenting with verbs that have been newly learnt. On the other hand, the fact that L2 learners fail to associate nouns with known verbs in collocations suggests the learning of new nouns in isolation, instead of being learnt as prefabricated chunks with the already acquired verbs. This finding supports Wray's (2002) claim of a split learning of collocations into individual items, or the inability to pay attention to collocational relationships between words on the part of L2 learners.

Therefore, the role played by new nouns in the collocation lag in ST6 learners can be viewed as no more than a minor one, influencing a limited percentage of new nouns in erroneous collocations (25%). In most cases where collocation errors with new nouns occur, it is because of an inability to associate the new nouns with already acquired verbs. On this basis, collocation lag is not a result of newly acquired nouns.

(2) New nouns in the ST6 VN collocations (new as compared with the ST5 level) and new nouns in the ST5 VN collocations (new as compared with the ST2 level)

It should be recalled that ST2 learners are senior middle school students, representing the lowest level in the CLEC corpus, and ST6 learners are third and fourth year university English majors, representing the highest proficiency level within the corpus. Between these two levels of proficiency, there is the ST5 level of first and second year English majors. Therefore, in order to ensure the continuity of between-group comparisons, VN collocation performance of ST5 learners was taken into consideration as well. Following the same procedure as in the analysis of ST6 as compared with the ST2 level, learners' collocation performance in terms of new nouns in the ST6 (new as compared with the ST5 level) and ST5 (new as compared with the ST2 level) was analysed. Frequencies are presented in Tables 6-10 and 6-11.

Table 6-10 New and old nouns in ST6 VN collocations (new nouns as compared with ST5)

	New nouns	Old nouns	Total
Erroneous coll.	5 (14%)	55	60
Correct coll.	31 (86%)	173	204
Total	36 (100%)	228	264

Table 6-11 New and old nouns in ST5 VN collocations (new nouns as compared with ST2)

	New nouns	Old nouns	Total
Erroneous coll.	6 (11%)	28	34
Correct coll.	48 (89%)	165	213
Total	54 (100%)	193	247

As a generalisation from Tables 6-8, 6-10 and 6-11, it becomes clear that the percentages of new nouns in erroneous collocations are low and remain roughly constant at the two higher levels. The percentage of new nouns in erroneous collocations in ST5 is 11% and the ratio is 14% for ST6 learners. So again the prediction that the acquisition of new nouns is not an inhibiting factor in collocation performance is upheld. It is interesting that if the new nouns at the ST6 level are identified as new with reference to the nouns produced by the ST2 learners, the percentage is 25%, exactly the sum of 11% and 14%. It follows that when the nouns in the highest level (ST6) are compared with those produced

by the lowest level (ST2), there are more new nouns obtained than are compared with a lower level (ST5). The three percentages (11%, 14% and 25%) suggest two significant points: the proficiency of the three levels of learners is continuously developing (as manifested through a gradual increase in newly acquired nouns); the percentages of new nouns in erroneous collocations appear to describe a coherent trend, supporting the validity of the analysis method in our study.

Turning now to the detailed analyses of new nouns in erroneous VN collocations, new nouns in erroneous collocations in the ST6 (new as compared with the ST5 level) and those in the ST5 (new as compared with the ST2 level) were analysed (see Tables 6-12 and 6-13 below).

Table 6-12 New nouns in ST6 VN erroneous collocations and their verb collocates (new nouns as compared with ST5)

Nouns	Erroneous verbs	Target verbs	Erroneous verbs in ST5	Target verbs in ST5
treaty	draw	sign	+	+
competence	exert	demonstrate	-	+
load	pull	carry	+	+
recognition	reach	receive	+	+
military service	attend	perform	+	+
military service	take	perform	+	+

(Note: '+' means that the verb appears in the ST5, and '-' represents an absence of the verb in the ST5.)

Table 6-13 New nouns in ST5 VN erroneous collocations and their verb collocates (new nouns as compared with ST2)

Nouns	Erroneous verbs	Target verbs	Erroneous verbs in ST2	Target verbs in ST2
role	do	play	+	+
role	act	play	+	+
role	occupy	play	+	+
role	lay	play	+	+
drum	hit	beat	+	+
eyebrow	frown	raise	-	+
offence	make	commit	+	-
regulation	do	enact	+	-
utmost	make	do	+	+

(Note: '+' means that the verb appears in the ST2, and '-' represents an absence of the verb in the ST2.)

From the above tables, it can be inferred that in nearly all the cases of new nouns in erroneous collocations, ST5 and ST6 learners may know the verb but fail in using them with the new nouns (except **make + offence*, **do + regulation*). Even when they do not lack the collocating verb for a new noun, errors still arise, which again reveal a split learning of collocations.

In conclusion, the research hypothesis that it is the increase in verbs that is mainly responsible for the stagnant collocation performance was upheld. In this section we attempted an alternative explanation to see if the learning of newly acquired nouns is also a factor responsible for the collocation lag. It was found that the occurrence of new nouns is not the main factor responsible for stagnant development of collocation performance of L2 learners. The percentages of new nouns in erroneous collocations are rather low (11% and 13%) and these figures remain roughly constant at both higher levels. In addition, even though new nouns are used in erroneous collocations, it was found that it is mainly not due to a lack of new collocating verbs, since the target verbs may have been acquired. Even though the target verbs have already been acquired, L2 learners fail to associate them with the new nouns in collocation, which suggests the learning of new nouns in isolation, rather than in chunks.

Chapter 7: L2 learners' performance on adjective + noun and noun + noun collocations

7.0 Introduction

In the learning of verb + noun collocations by Chinese learners of English, verb increase in a set of semantically related verbs has been shown above to pose great difficulties for learners. There is a lag in collocational knowledge as observed in learners' performance on verb + noun collocations. In light of this finding, this chapter will now explore two other important and frequent types of collocations: adjective + noun and noun + noun collocations, in order to investigate the inhibiting factor of vocabulary growth in collocation learning.

For the convenience of comparison, analyses of AN and NN collocations were performed only on the ST2 and ST6 data, given that the ST5 data show a certain inconsistency in the quantity of collocation errors in VN collocations (cf. Section 5.1.3). Learners' performance on AN collocations was firstly presented in Section 7.1, followed by analyses of their production of NN collocations (Section 7.2).

7.1 Analyses of adjective + noun collocations

Examples of frequent adjective + noun collocations produced by ST2 and ST6 learners are (collocations with a frequency over 10): *active part, best wishes, civil war, rapid progress, associate professor, developed countries, developing countries, economic development, environmental protection, living condition, natural resources, etc.* Erroneous AN collocations in the two databases are (with the target adjectives given in brackets): **beautiful supper (delicious), *deep language (rich), *heavy sufferings (great), *large laughter (loud), *large voice (loud), *light river (clean), *sharp match (fierce), *classical song (classic), *clear welcome (warm), *feminine movement (feminist),* etc.

The overall number of well-formed and erroneous AN collocations is presented in Table 7-1 (for tokens) and Table 7-2 (for types).

Table 7-1 AN collocations produced by ST2 and ST6 learners (tokens)

	Well-formed coll.	Erroneous coll.	Total
ST2	337 (95%)	17 (5%)	354 (100%)
ST6	986 (99%)	15 (1%)	1001 (100%)

(Note: $\chi^2 = 10.99$, $p = 0.0009$ ***)

Table 7-2 AN collocations produced by ST2 and ST6 learners (types)

	Well-formed coll.	Erroneous coll.	Total
ST2	100 (88%)	14 (12%)	114 (100%)
ST6	217 (96%)	9 (4%)	226 (100%)

(Note: $\chi^2 = 7.01$, $p = 0.0081$ **)

It is noteworthy that there are not so many instances of erroneous AN collocations in the corpus. Only 17 cases were identified in the collocations produced by the lower level learners and 15 instances by the higher level. Apart from this, the two tables reveal two important aspects. Firstly, it is evident that the ratios of AN collocation errors are not high for both group of learners in terms of either tokens or types. These ratios are 12% for ST2 learners and 4% for ST6 leaners. In contrast to the proportion of verb + noun collocation errors (for ST2, 22%, and for ST6, 21%, cf. Section 5.1.3), they are much lower. Secondly, learners at the higher level produced significantly more well-formed collocations than the lower level, and there was a significant error decrease. The decrease in AN collocation errors stands in sharp contrast to the production of VN collocations, for which the proportion of errors does not show a clear decrease.

Therefore, based on quantitative analyses, the data show that Chinese learners do not seem to have great difficulties with adjective + noun collocations. In addition, their knowledge of AN collocations improves with rising proficiency. These results corroborate the findings from previous studies of L2 learners' learning of AN collocations (e.g. Gitsaki, 1999; Siyanova and Schmitt, 2008; Zhang and Chen, 2006; cf. Section 3.2.1.2). Adjective – noun collocations have been identified as "easy" and "early acquired" type of collocations (Gitsaki, 1999) and more proficient learners had better command of AN collocations than lower levels (Zhang and Chen, 2006).

7.2 Analyses of noun + noun collocations

Examples of frequent noun + noun collocations produced by ST2 and ST6 learners are: *art festival, basketball match, book shop, fire fighter, swimming pool, crime rate*, etc.[46] Erroneous NN collocations include **artist festival (art festival), *homehold duties (household duties),* and **scientist book (science book),* etc. Altogether there are 82 and 46 instances of erroneous noun + noun combinations in each database. A detailed look into the errors involving noun + noun combinations shows that not all of these errors are collocation errors, i.e. errors associated with the wrong choices in words of the same word class. Instead, a large proportion of them are colligation errors, i.e. errors linked with wrong choices in grammatical categories. Unlike collocations referring to co-occurrence of word combinations, colligation refers to the co-occurrence of grammatical choices. So a collocation error is an error with wrong lexical selections (but the word class is correct), e.g. **learn knowledge* rather than *acquire knowledge*. A colligation error is an error with word classes of the words in a word combination, e.g. **industry city* rather than *industrial city*. Table 7-3 presents the NN colligation errors in the ST2 and ST6 databases).

[46] In the creation of the ST6 NN collocation database, two collocations – *mercy killing* (which has been used for 255 times) and *prison system* (107 times) – were deleted so as to avoid statistical skewedness. The high occurrence of these two collocations is because they are topic-related: two of the topics given to the English majors are "the legalisation of euthanasia in China" and "the abolition of prisons".

Table 7-3 Noun + noun colligation errors produced by ST2 and ST6 learners

ST2		ST6	
Errors	Target words	Errors	Target words
flowers exhibition	flower	electricity lamps	electric
foreigner teacher	foreign	environment consciousness	environmental
happiness family	happy	families members	family
industry city	industrial	feminism movement	feminist
interest book	interesting	feudalism society	feudalist
socialism country	socialist	globe economy	global
socialism reformation	socialist	heat debate	heated
sport ground	sports	heat topic	heated
sport meeting	sports	importance step	important
sports trousers	sport	industry revolution	industrial
summer's holiday	summer	limit recourses	limited
history's test	history	medicine fee	medical
math's test	math	nationality defence	national
freedom life	free	nature process	natural
people computer	personal	scenery spots	scenic
		socialism construction	socialist
		socialism countries	socialist
		society evolution	social
		society factor	social
		society problem	social
		society wealth	social
		examples sentences	example
		capitalism countries	capitalist
		science way	scientific
		economy development	economic
		economy growth	economic
		stars hotel	star

In most of the colligation errors, learners wrongly use a noun instead of its adjectival form, e.g. *foreigner teacher* rather than *foreign teacher, electricity lamps* rather than *electric lamps, environment consciousness* instead of *environmental consciousness*, etc. In English, both a noun and adjective can function as the modifiers of a following noun, and this grammatical feature seems to be baffling learners in choosing which to collocate with the nouns that follow. Another factor of the misuses of nouns for adjective modifiers may be a cross-linguistic one; a noun modifier before another noun is very common in the Chinese syntax. What is more interesting among these errors is that sometimes learners are

conscious of the typical grammatical feature of English and try to use adjectival modifiers such as possessives before the nouns. As the examples in Table 7-3 show, *summer's holiday, history's test* and *math's test* are indications of an awareness of adjectival modifiers. Yet the overgeneralisation leads to colligation errors.

In terms of the quantities of these errors, learners do not seem to get better with NN colligations as their proficiency level rises. On the contrary, there is a worsening performance in noun + noun colligations. Tables 7-4 and 7-5 present the frequencies of colligation errors and non-colligation errors in the two proficiency groups.

Table 7-4 Colligation and non-colligation NN errors in the ST2 and ST6 levels (tokens)

	NN colligation errors	NN non-colligation errors	Total
ST2	34 (43%)	46 (57%)	80 (100%)
ST6	38 (83%)	8 (17%)	46 (100%)

(Note: $p < 0.0001$ ****)

Table 7-5 Colligation and non-colligation NN errors in the ST2 and ST6 levels (types)

	NN colligation errors	NN non-colligation errors	Total
ST2	15 (41%)	22 (59%)	37 (100%)
ST6	27 (77%)	8 (23%)	35 (100%)

(Note: $p = 0.0020$ **)

From the two tables, it can be seen that in percentage terms, 83% of the NN combination errors ST6 learners made are colligation errors, whilst the percentage for ST2 learners is 43%. Fisher's test on the frequencies of tokens and types showed a significant difference between learner types in terms of the production of colligation errors. ST6 learners made significantly more errors with noun + noun colligations than ST2 learners. This could be that the influence of the noun + noun structure in the L1 Chinese is very persistent even at higher levels. We shall turn to the L1 influence on the learning of collocations by L2 learners in Chapter 9.

Besides the NN colligation errors, there are also a few cases where the entire expression does not make sense in English and should be a noun or a totally new expression, i.e. **smile sound* for *laughter*,

**hill-medicine* for *yam*, **football door* for *goal*, **book table* for *desk*, **mother school* for *alma mater*, **warning clock* for *alarm bell*, **psychology doctor* for *psychiatrist* and **song words* for *lyric*. Such unacceptable expressions are the direct word-by-word rendering of the Chinese characters into English. This may be a strategy adopted by L2 learners as they turn to the direct translation of the Chinese expression (*shanyao*) when they haven't acquired the English word (*yam*).

With the colligation errors and errors of the entire expression excluded, the remaining 16 NN combinations in the ST2 database are erroneous noun + noun collocations. Collocation errors produced by ST2 learners are: **artist festival (art festival), *basketball ground (basketball court), *football court (ground), *football movement (football games), *hand master (head master), *hand teacher (head teacher), *heart illness (heart disease), *homehold duties (household duties), *saw materials (raw materials), *scientist book (science book), *speech match (speech contest), *pity girl (poor girl), *middle night (mid-night), *singers match (singing match), *end game (final game), *beauty match (beauty contest)*. The 7 NN collocation errors in the ST6 database are: **feminine movement (feminist movement), *feminism women (feminist women) *graduation certification (graduation certificate), *life standard (living standard), *mountain slides (land slides),* and **song star (music star)*. The overall number of well-formed and erroneous NN collocations is presented in Tables 7-6 (for tokens) and 7-7 (for types).

Table 7-6 Noun + noun collocations in the ST2 and ST6 levels (tokens)

Learners	Well-formed coll.	Erroneous coll.	Total
ST2	618 (95%)	31(5%)	649 (100%)
ST6	792 (99.2%)	6 (0.8%)	798 (100%)

(Note: $\chi^2 = 21.68$, $p < 0.0001$ ****)

Table 7-7 Noun + noun collocations in the ST2 and ST6 levels (types)

Learners	Well-formed coll.	Erroneous coll.	Total
ST2	202 (93%)	16 (7%)	218 (100%)
ST6	307 (98%)	6 (2%)	313 (100%)

(Note: $\chi^2 = 8.20$, $p = 0.00425$ **)

As is revealed from the above tables, the proportions of erroneous noun + noun collocations are

very low: 7% for ST2 and 2% for ST6 learners. Statistical analyses showed that higher level learners made very significantly fewer errors than the lower group in the use of NN collocations. This indicates a better command of NN collocations as learners' proficiency rises. Comparing this finding with the production of AN collocations, it shows a similar pattern, i.e. better performance on NN collocations was observed with the rise of L2 proficiency.

7.3 Synopsis of the analyses of adjective + noun and noun + noun collocations

This chapter expands the analyses of Chinese learners' collocation performance to two other frequent types of collocations: adjective + noun and noun + noun collocations. A good performance was observed in the production of both AN and NN collocations by Chinese EFL learners at two proficiency levels. There was a very low proportion of AN collocation errors (ST2: 12%; ST6: 4%) and NN colloation errors (ST2: 7%; ST6: 2%), although the ratios of AN collocations errors were more than those of NN collocation errors. Meanwhile, learners' knowledge of AN and NN collocations improves with rising proficiency, as there were significantly more well-formed AN and NN collocations at the higher level. The picture emerging from adjective + noun and noun + noun collocations is different from learners' performance on verb + noun collocations. The next chapter will compare in detail Chinese learners' production of the three types of collocations and make possible interpretations accounting for such differences.

Chapter 8: Comparison and interpretation of learners' performance on the three types of collocations

8.0 Introduction

This chapter presents and compares the results obtained from the analyses of Chinese learners' verb + noun, adjective + noun and noun + noun collocations. Section 8.1 performs a comparison of erroneous collocations among the three types of collocations; Section 8.2 analyses and compares the overall growth of verbs, adjectives and nouns and relates them to the production of collocations; Section 8.3 analyses in detail the synset density of the three word classes and offers interpretations of the differing performances on the VN, AN and NN collocations.

8.1 Collocation errors in the three types of collocations

Table 8-1 presents the overall percentages of collocation errors among the production of verb + noun, adjective + noun and noun + noun collocations by Chinese EFL learners at the basic and advanced levels.

Table 8-1 Error ratios of VN, AN, and NN collocations produced by ST2 and ST6 learners

Learners	Verb + noun coll.	Adjective + noun coll.	Noun + noun coll.
ST2	22%	12%	7%
ST6	21%	4%	2%

A comparison of the ratios of erroneous collocations among the three types of collocations at each proficiency level (e.g. in the ST2 level, 22% for VN, 12% for AN and 7% for NN) shows that Chinese L2 learners, irrespective of their proficiencies, performed best in NN collocations, followed by AN collocations, and performed worst in VN collocations. A cross-group comparison of the ratios demonstrates a varied developmental pattern, i.e. no clear decrease in collocation errors with the rise of proficiency in the production of VN collocations, but a decrease in errors in the production of AN and

NN collocations. Combining this result with the findings from statistical tests, no significant relationship was found between learner levels and erroneous VN collocations (cf. Section 5.1.3), suggesting that there is no significant decrease in VN collocation errors. However, there were very significantly fewer AN and NN collocation errors at the ST6 level than the ST2 level, suggesting an improvement in acquisition of these two collocation types.

Therefore, a varied collocation performance by L2 learners was observed. Simply put, there is a stagnant development of VN collocation knowledge but improved AN and NN collocation performance. VN collocations have long been acknowledged as difficult for L2 learners. As for the acquisition of AN and NN collocations, Philip (2007: 1) claims that they are easier to learn because

> many noun + noun and adjective + noun collocates reflect linguistically what can be observed in the real world, and so their lexical co-occurrence seems natural and relatively unproblematic. However, as language moves away from the concrete and observable towards the abstract and, apparently, arbitrary, the likelihood of a collocate to be deemed 'logical' by a learner diminishes, as does its accurate recycling in free production.

This claim holds water in that many noun + noun and adjective + noun collocations in our databases are concrete lexical co-occurrences that can be observed, i.e. *air conditioner, bus stop, football ground, blue sky, full moon, heavy rain*, etc. However, the observable property of certain collocations does not guarantee an easy acquisition because of the arbitrariness in lexical co-occurrences. For example, *heavy rain* is concrete and observable, but it does not rule out the possible combinations of **dense rain, *strong rain,* or **powerful rain*. Learners were found to make errors with concrete NN collocations as well, such as **basketball ground, *mountain slides, *light river*, etc. At the same time, many abstract AN and NN collocations were correctly produced, i.e. *environment protection, labour force, sales volume, cheap trick, deep sorrow,* etc. So the observable feature of adjective + noun and noun + noun collocations as proposed by Philip (2007) cannot account for learners' relative ease with these two types of collocations.

Another possible explanation of the relative difficulty in acquiring VN collocations and relative ease in acquiring AN and NN collocations is based on the observation that words of different parts of speech differ in their tendency to cluster, i.e. singular nouns and base forms of verbs are highly collocational while adjectives and adverbs are not (Kjellmer, 1990). This could be interpreted to mean that verbs and nouns are more collocational and therefore bring greater learning burden for foreign language learners. In other words, the collocational density of verb + noun collocations poses more problems for learners than AN and NN collocations. This could account for the overall learning burden

of VN collocations, but to account for the different performance on the three types of collocations, we argue for the vocabulary growth factor and predict that the synonym densities of adjectives and nouns are lower than those of verbs, thus resulting in better performance in the AN and NN collocations and worse performance in VN collocations. We shall discuss this in Sections 8.2 and 8.3.

8.2 Vocabulary growth and collocation errors

Through adopting a similar approach in examining the relationship between the growth of lexical verbs and the growth of verb + noun collocation errors (cf. Section 5.3), we calculated the growth rates of all the adjectives and nouns used by the three levels of learners (including adjectives and nouns in sentences like “a healthy diet is very important for people”), aiming to build a panorama of vocabulary growth and erroneous collocations.

Table 8-2 Growth rates of adjectives, nouns and collocation errors

	Lemmatised adjectives	Lemmatised nouns	Tokens of AN coll. errors	Types of AN coll. errors	Tokens of NN coll. errors	Types of NN coll. errors
ST2	1175	3345	17	14	31	16
ST5	1850	3857	-	-	-	-
ST6	2287	4199	15	9	6	6
Growth rate (ST2 to ST6)	95%	26%	-12%	-36%	-81%	-63%

(Note: The numbers of lemmatised adjectives and nouns are type frequencies.)

As shown from Table 8-2, the numbers of adjectives and nouns produced by ST5 learners occupy a middle position, confirming again a continuous rise in proficiency from the ST2 to the ST6 levels. At the same time, there is a dramatic increase in adjective uses from the ST2 to ST6 level (95%), but a clear decrease in AN collocation errors. A less dramatic increase in nouns and a more marked decrease in NN collocation errors are also displayed in the above table. Therefore, in comparison with the VN data, different trends emerge with regard to performance on verb + noun, adjective + noun and noun + noun collocations, i.e. there is an increase in VN collocation errors and a decrease in AN and NN collocation errors (see Figure 8-1 for graphic depiction).

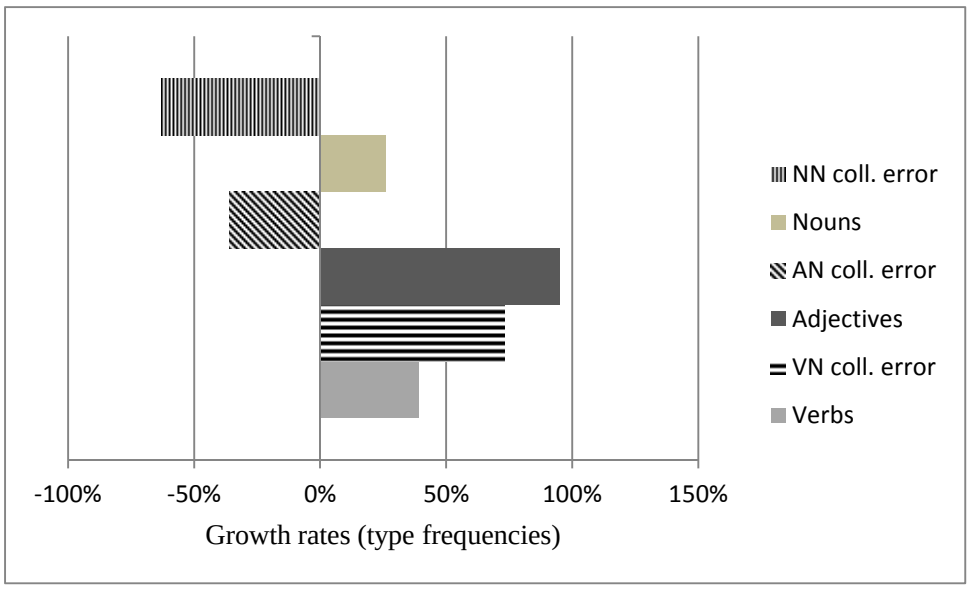


Figure 8-1 Overall growth rates of the verbs, adjectives and nouns and collocation errors

If the error increase is viewed in association with the increase of vocabulary, Figure 8-1 shows that VN collocation errors increase as learners acquire more and more verbs and nouns, but AN and NN collocation errors decrease even though learners acquire more adjectives and nouns. Thus, we see that vocabulary growth plays a different role depending on word class-specific collocations. More specifically, the increase in learners' overall vocabulary in verbs has a negative correlation with VN collocation learning, whereas the expansion in adjective and noun vocabulary contributes to the learning of AN and NN collocation. The detailed analysis of verb increase within certain synonym sets confirmed a lag in VN collocational knowledge, as collocation errors involving new verbs were significantly more likely than errors with old verbs. The proportion of errors in synsets where new verbs have been acquired increased markedly. So verb increase is an inhibiting factor in VN collocation acquisition. However, with the increase in adjectives and nouns, AN/NN collocation errors in contrast decreased. Therefore, following the procedures adopted in detailed analysis of verb synsets, grouping adjectives and nouns into synsets and examining collocation uses within these sets was required. Analyses of adjective and noun synsets will be performed in the following section in order to investigate in detail the vocabulary factor.

8.3 Synsets and collocation production

The classification of the adjectives and nouns in the databases into synonym sets was performed with reference to the WordNet and the *ODSA*. These classifications were much more difficult than the classification of verbs into synsets since adjectives and nouns were more diversified than verbs in the collocation databases. Starting from the highest level where more adjectives are produced in AN collocations, adjectives listed in either of the reference sources as synonyms were recorded as one synset. Accordingly only seven synsets with a rather limited number of adjectives were categorised by using the same reference works as with verbs. They are adjectives describing broadness (*broad, full, wide*), adjectives denoting keenness *(keen, sharp*), "deadly" adjectives (*deadly, fatal, lethal*), "clean" adjectives (*clean, clear, light*), "distant" adjectives (*distant, remote*), "dense" adjectives (*dense, heavy*) and "daily" adjectives (*daily, everyday*). Adjectives in the ST2 collocation databases are more diversified and only "daily" adjectives (*daily, everyday*) were identified.

The difficulty in grouping adjectives in synsets may be attributed to the types of adjectives that were frequently used by learners, e.g. classifying adjectives (e.g. *academic, annual*).[47] A large number of adjectives used by ST2 and ST6 learners in AN collocations are of a classifying nature (e.g. among the 142 types of adjectives in the ST6 AN collocation database, 70 are classifying ones, cf. Appendix Ⅶ). Classifying adjectives, as their name suggests, function to group nouns into different categories, so these adjectives themselves are too broad in scope to be categorised in synsets. For example, synonyms of *capitalist, chemical, domestic,* and *solar* have only a few synonyms as referenced in WordNet.

As it is difficult to classify adjectives into different synsets, the nouns produced in the NN collocations, such as *air, alarm, art* are also difficult to group. So in general the synonym density of adjectives and nouns in the databases is lower than that of verbs, which may be the reason why AN and NN collocations were more accurate than VN collocations. Studies of the synsets of verbs, adjectives and nouns have confirmed such a decrease in synonym density. According to the statistics published online for WordNet 3.0 database, the ratios of synsets as compared to the total number of verbs, adjectives and nouns respectively are 1.19 for verbs, 0.85 for adjectives and 0.70 for nouns (see Table

[47] Adjectives are classified into 4 categories (Sinclair and Fox, 1990: 63f): qualitative adjectives (which identify qualities someone or something has, e.g. *happy,* and *intelligent*), classifying adjectives (which identify someone or something as a member of class, e.g. *financial* and *intellectual*), colour adjectives (identifying the colour of something, e.g. *blue* and *gree*n) and emphasising adjectives (which are used to emphasise feelings, e.g. *complete* and *absolute*).

8-3 for the raw statistics cited from WordNet).

Table 8-3 Numbers of words, synsets and senses in WordNet[48]

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Therefore, based on the statistics from the WordNet, in general there are more synsets for verbs than adjectives, and more synsets for adjectives than nouns. That decreasing synonym density of the three word classes was also verified through computational analysis of WordNet (cf. Kamps et al., 2004; Tufis and Stefanescu, 2011). In the graphs drawn by Kamps et al. (2004: 1116) through collecting all words in the WordNet, and relating words that can be synonymous, they observed a giant component: in the verb-subgraph there is a component of size 6,365 (or 57% of all verbs); in the adjective-subgraph there is a component of size 5,427 (or 25% of all adjectives) and in the noun-subgraph there is a connected component of size 10,922 (or 10% of all nouns). These figures show that more verbs than adjectives and more adjectives than nouns are related in synsets.

However, synset analysis of words in WordNet is very general. A more directly focused study relevant to learner data in our research was conducted, through narrowing down the investigation of the density of the three word classes into our databases. A straightforward and simple way to measure synonym density is to compute the average number of synonyms for each verb, adjective and noun. Considering the large number of words in learner data, a random sampling of words in the ST2 databases was adopted. The sampling procedure is as follows: altogether 80 lexical verbs, 60 adjectives and 278 nouns in the VN, AN and NN collocations were identified, and then listed alphabetically. Every four verbs, every three adjectives and every fourteen nouns were chosen in the three wordlists to form the sample in this case study. In all, 60 words were selected, i.e. a random selection of 20 verbs in VN collocations, a random selection of 20 adjectives in AN collocations and a random selection of 20 nouns

[48] Quoted from the table found at http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc2 [Accessed 10 May 2014]

in NN collocations. These words are:

a. 20 verbs in VN collocations: *answer, break, catch, comb, create, discharge, earn, follow, grasp, kick, lead, obey, pass, play, remember, see, show, sow, teach, wear*

b. 20 adjectives in AN collocations: *blue, botanical, capitalist, classical, common, crisp, deep, double, fair, firm, founding, glib, happy, historic, living, low, natural, political, public, strong*

c. 20 nouns in NN collocations: *ball, break, center, colour, diamond, fashion, gambling, head, lab, light, name, party, police, program, restaurant, sentence, steel, telephone, trip, world*

Synonyms of those 60 words were located in WordNet. However, not all the synonyms for each word are counted. WordNet provides the synonyms for each word according to their different senses, since different senses of the word lead to different synonyms. As Table 8-3 shows, the average polysemy figure of verbs is the highest – 2.17 (25047/11529), much more than the average polysemy of adjectives: 1.40 (30002/21479) and of nouns: 1.24 (146312/117798).[49] So the inclusion of all synonyms irrespective of word senses will naturally add more synonyms for verbs than adjectives and nouns. For example, 59 senses are listed for the verb *break*, but for the adjective *botanical*, there is only 1 sense. Therefore, only the synonyms of the particular sense of the word in question are taken into consideration. The sense of a word is determined by its following collocates. Take the verb *answer* for example. WordNet software (version 2.1) yielded the following 10 senses for this verb:[50]

Sense 1
answer, reply, respond -- (reply or respond to; “She didn’t want to answer”; “answer the question”; “We answered that we would accept the invitation”)
=> state, say, tell -- (express in words; “He said that he wanted to marry her”; “tell me what is bothering you”; “state your opinion”; “state your name”)
Sense 2
answer -- (give the correct answer or solution to; “answer a question”; “answer the riddle”)
=> solve, work out, figure out, puzzle out, lick, work -- (find the solution to (a problem or question) or understand the meaning of; “did you solve the problem?”; “Work out your problems with the boss”; “this unpleasant situation isn’t going to work itself out”; “did you get it?”; “Did you get my meaning?”; “He could not work the math problem”)
Sense 3
answer -- (respond to a signal; “answer the door”; “answer the telephone”)
=> react, respond -- (show a response or a reaction to something)
Sense 4
answer, resolve -- (understand the meaning of; “The question concerning the meaning of life

[49] See also the statistics on the WordNet website: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc2 (Accessed 8 April 2014)
[50] For the convenience of checking the synonyms of the verb in question, Wordnet (2.1) was referenced instead of the web interface, as in the software different senses of the verb are numbered, and synonyms are neatly listed.

cannot be answered")

=> solve, work out, figure out, puzzle out, lick, work -- (find the solution to (a problem or question) or understand the meaning of; "did you solve the problem?"; "Work out your problems with the boss"; "this unpleasant situation isn't going to work itself out"; "did you get it?"; "Did you get my meaning?"; "He could not work the math problem")

Sense 5

answer -- (give a defence or refutation of (a charge) or in (an argument); "The defendant answered to all the charges of the prosecution")

=> refute, rebut -- (overthrow by argument, evidence, or proof; "The speaker refuted his opponent's arguments")

Sense 6

answer -- (be liable or accountable; "She must answer for her actions")

=> be -- (have the quality of being; (copula, used with an adjective or a predicate noun); "John is rich"; "This is not a good answer")

Sense 7

suffice, do, answer, serve -- (be sufficient; be adequate, either in quality or quantity; "A few words would answer"; "This car suits my purpose well"; "Will $100 do?"; "A 'B' grade doesn't suffice to get me into medical school"; "Nothing else will serve")

=> satisfy, fulfil, fulfil, live up to -- (fulfil the requirements or expectations of)

Sense 8

answer -- (match or correspond; "The drawing of the suspect answers to the description the victim gave")

=> match, fit, correspond, check, jibe, gibe, tally, agree -- (be compatible, similar or consistent; coincide in their characteristics; "The two stories don't agree in many details"; "The handwriting checks with the signature on the check"; "The suspect's fingerprints don't match those on the gun")

Sense 9

answer -- (be satisfactory for; meet the requirements of or serve the purpose of; "This may answer her needs")

=> meet, satisfy, fill, fulfil, fulfil -- (fill or meet a want or need)

Sense 10

answer -- (react to a stimulus or command; "The steering of my new car answers to the slightest touch")

=> react, respond -- (show a response or a reaction to something)

In the particular collocation *answer* + *question* produced by ST2 learners, the senses of *answer* are 'reply/respond to, or give the correct answer or solution to, or resolve', corresponding to Senses 1, 2 and 4. Therefore, the number of synonyms for *answer* is 9, which include *reply to, respond to, solve, work out, figure out, puzzle out, lick, work* and *resolve*. Troponyms (e.g. *state, say, tell* in Sense 1) which denotes the manner of answering a question are not counted as synsets for *answer*, whilst hypernyms or superordinates (e.g. solve, work out, figure out, puzzle out, lick, work in Sense 2) are included.

The same procedure was performed on the verb *break*. In the collocations *break* + *rule* and *break* +

record, Senses 6 and 14 correspond to the senses of *break* in the two word combinations (see the following senses extracted from WordNet software (version 2.1)) and accordingly 16 synonyms were located.

Sense 6

transgress, offend, infract, violate, go against, breach, break -- (act in disregard of laws, rules, contracts, or promises; “offend all laws of humanity”; “violate the basic laws or human civilization”; “break a law”; “break a promise”)

=> disrespect -- (show a lack of respect for)

Sense 14

better, break -- (surpass in excellence; “She bettered her own record”; “break a record”)

=> surpass, outstrip, outmatch, outgo, exceed, outdo, surmount, outperform -- (be or do something to a greater degree; “her performance surpasses that of any other student I know”; “She outdoes all other athletes”; “This exceeds all my expectations”; “This car outperforms all others in its class”)

Altogether the 60 words and their number of synonyms are presented in the following table.

Table 8-4 Selected words in the learner databases and the number of synonyms

verbs	No. of syns.	adjectives	No. of syns.	nouns	No. of syns.
answer	9	blue	2	ball	0
break	16	botanical	1	break	4
catch	5	capitalist	1	center	0
comb	6	classical	1	colour	0
create	2	common	4	diamond	0
discharge	8	crisp	6	fashion	3
earn	9	deep	6	gambling	0
follow	3	double	2	head	2
grasp	7	fair	7	lab	0
kick	1	firm	2	light	0
lead	2	founding	0	name	0
obey	3	glib	1	party	0
pass	4	happy	9	police	0
play	6	historic	2	program	1
remember	6	living	0	restaurant	0
see	7	low	7	sentence	0
show	4	natural	2	steel	0
sow	9	political	0	telephone	1
teach	3	public	1	trip	2
wear	3	strong	3	world	0
Total	113	Total	57	Total	13

(Note: ‘No. of syns.’ is short for the number of synonyms.)

As the above table reveals, the synonyms for verbs (113) far outnumber those for adjectives (57), and the synonyms for adjectives outnumber those for nouns (13). The result confirms the findings from computational studies of synsets in WordNet, viz. the synonym density for verbs, adjectives and nouns is on a decreasing scale. In the light of the fact that verbs have more synsets than adjectives and adjectives have more synsets than nouns, we get a better understanding of why L2 learners perform worse on verb + noun collocations and better on adjective + noun and noun + noun collocations. Since there are more synonyms for verbs, the more verbs in a synset learners acquire, the more likely they are to make collocation errors. In contrast, the fewer synonyms for adjectives and nouns may explain why learners seldom made errors in choosing the right collocates, although they were confused with the grammatical forms of words and made colligation errors in NN collocations.

8.4 Synopsis of the findings in this chapter

This chapter has been concerned with a comparison of learners' performance on verb + noun, adjective + noun and noun + noun collocations. Two overall findings have emerged with regard to the three collocation types: a relatively poorer performance on VN collocations was observed in Chinese EFL learners but better performance was found on AN and NN collocations. Moreover, learners' knowledge of VN collocations did not increase with rising proficiency but their knowledge of AN and NN collocations improved as they proceeded to higher levels.

A possible reason for such a difference was explored with regard to the overall density of the synonyms for the three syntactic categories of words as collocators: verbs, adjectives and nouns. Computational analyses both of the synsets of the three word classes in WordNet and in a case study revealed that verbs generally have more synonyms than adjectives, while adjectives have more synonyms than nouns. Therefore, the better performance on AN and NN collocations can be accounted for through the lower density in synonyms. In this regard, the analyses of AN and NN collocations in this chapter consolidate the prediction that vocabulary growth is an inhibiting factor in collocation acquisition. To be more specific, for word classes where there is little increase in a synonym set, collocation errors are seldom made (as for adjective and nouns in AN and NN collocations); where there are increases in words in synsets, chances of errors subsequently increase (as for verbs in VN collocations).

Chapter 9: The role of L1 in collocation learning[51]

9.0 Introduction

As has been reviewed in Section 3.2.2, the L1 of the learner plays a considerable role in the learning and production of L2 collocations. On the one hand, previous empirical collocation studies into L2 learners' collocation performance show that L1-influenced errors make up a large proportion of errors even at advanced levels (Laufer and Waldman, 2011; Nesselhauf, 2005). For example, Laufer and Waldman (2011) reported that over 60% of the verb + noun collocation errors produced by intermediate and advanced learners were L1-induced. This proportion is not found to decrease over time. A heavy reliance on their mother tongue in collocation production is also manifested through the overuse of certain collocations that are linked to lexical combinations in the L1 and the underuse of collocations that are mismatched between the two languages (e.g. Granger, 1998; Kaszubski, 2000). On the other hand, research exploring the psychological reality of L2 collocation learning in terms of L1 and L2 congruence and non-congruence found that L2 learners perform better on congruent collocations than non-congruent ones in lexical decision tasks. It is also suggested that non-congruent collocations stored in memory are processed autonomously without word-by-word mediation of the L1. However, there is still a paucity of investigation into whether L2 learners produce congruent collocations with more accuracy than non-congruent ones, and whether non-congruent collocations once learnt, are less susceptible to errors. Therefore, we set out in this chapter to investigate the role of L1 in L2 collocation learning in terms of the production of congruent and non-congruent collocations. It is worthwhile for this study to provide a point of comparison with existing research in terms of the influence of L1 lexical network on L2 collocation learning. In addition, little research evidence has been provided with regard to the role of the learners' mother tongue in different types of collocations, e.g. verb + noun and adjective + noun collocations. So the second goal of this chapter is to compare learner performance with regard to congruent and non-congruent VN and AN collocations in order to see if there are differences in terms of cross-linguistic influence. It is hypothesised that:

[51] This chapter has been written into a paper titled *Cross-linguistic Influence on the Production of L2 Collocations: A Corpus-based Study of Chinese EFL Learners' Collocation Learning.* The paper was given at the 9th Newcastle upon Tyne Postgraduate Conference in Linguistics, 4 April, 2014.

1. L2 learners perform better with congruent collocations than non-congruent collocations in collocation production.
2. Non-congruent collocations that are correctly used by learners at lower levels are not wrongly used by learners at higher levels.
3. The L1 plays a different role in verb + noun and adjective + noun collocations.

To test the above hypotheses, the notion of congruence in classifying L2 collocations was firstly clarified (Section 9.1). Then within-group comparison of well-formed and erroneous congruent and non-congruent VN collocations was carried out to see whether congruent collocations were produced with a higher accuracy than non-congruent ones (Section 8.2); Next, between-group comparison on the well-formed and erroneous congruent and non-congruent VN collocations was performed in order to see whether, from a developmental perspective, non-congruent collocations that were used correctly at lower levels were not wrongly used by higher levels (Section 9.3); Finally, within-group comparison of positive and negative L1 influence with verb + noun and adjective + noun collocations was made to examine the role of the L1 in the learning of different types of collocations (Section 9.4). The last section (Section 9.5) presents a summary of this chapter.

9.1 The notion of congruence

Congruence between L1 and L2 is approached from the perspective of cross-linguistic translation equivalence (e.g. Bahns, 1993; Marton, 1977; Nesselhauf, 2003; 2005; Philip, 2007; Wolter and Gyllstad, 2011; Yamashita and Jiang, 2010). Yet as Nesselhauf (2005: 221) acknowledges, the notion of congruence is difficult to grasp. In contrastive or translation studies, translation equivalence refers to the correspondence of forms or constructions between the source language and the target language (Johansson, 2007: 23). Here, translation equivalence used in determining collocation congruence in two languages is measured at a word-for-word level. In other words, if word elements in a collocation in one language have direct word-for-word translational equivalence in another language, then it is considered as a congruent collocation; otherwise, it is a non-congruent collocation (Nesselhauf, 2005; Wolter and Gyllstad, 2011). An example of congruent collocation given by Wolter and Gyllstad (2011) is *give an*

answer, which can be rendered on a word-for-word basis in Swedish as *ge ett svar*. For non-congruent collocations, they give the example of *pay a visit*, for which the literal Swedish translation would be **betala ett besÖk,* which is infelicitous.

A difficulty in capturing the notion of congruence is deciding how to measure whether two words are "direct translation equivalents" in two languages. For example, the English *strong tea* is idiomatically translated into Chinese as *nong cha*. Yet *nong cha* is literally **dense tea* when directly translated into English. In this case it is not certain whether the English collocation *strong tea* shares "direct" translation equivalence with the Chinese expression or not. Therefore, in order to measure direct translational equivalence, the strategy of "back translation" is adopted, where the translated word in question is translated back into English out of context, and we see if it can be translated to the English word in question (Altenberg and Granger, 2002: 17; Nesselhauf, 2005: 221). At the same time, to reduce human judgment and make the classification feasible, the classification of congruent and non-congruent collocations was approached in our study with the assistance of two bilingual dictionaries. The procedure was as follows:

First, congruence is only considered at the level of content words;[52]

Next, the noun in the VN collocation is translated into Chinese;

Thirdly, the meaning of the verb is looked up in a bilingual dictionary (*Oxford Advanced Learners' English-Chinese Dictionary (7th Edition)*) (*OALECD*) with reference to the meaning of the following noun, and then the Chinese meaning is located;

Fourthly, we check if the Chinese verb translation together with the Chinese noun make a felicitous sequence. If not, then it is non-congruent collocation; if so, we take the fifth step:

Finally, if the sequence obtained is a felicitous Chinese sequence, a check is made as to whether, in the back translation process, the Chinese verb in question would out of context be readily translated into the English word in question, with reference to a Chinese-English bilingual dictionary-*New Century Chinese-English Dictionary (NCCED)*. If so, it is a congruent collocation; if not, it is considered as a non-congruent one.

The following three collocations exemplify the detailed procedures:

[52] Grammatical words usually behave differently across languages. That is especially true of the Chinese language, which has fewer words functioning as prepositions and they are used less frequently than in English. The Chinese does not have articles (Cross and Papp, 2008: 68; LÜ, 2002). So grammatical words are disregarded and only content words of verbs and nouns in our database were considered.

Example 1: MAKE + CONTRIBUTION

Analysis:

1. The Chinese meaning of CONTRIBUTION is *gongxian*.
2. The sense of MAKE in the *OALECD* in the context of MAKE + CONTRIBUTION is 'create' and in Chinese *zuo*.
3. The Chinese sequence *zuo gongxian* is felicitous.
4. The Chinese verb *zuo* in the *NCCED* can be readily translated into *make* in English.

Conclusion: MAKE + CONTRIBUTION is a congruent collocation.

Example 2: TAKE + NOTE

1. The Chinese meaning of NOTE is *biji*.
2. The sense of TAKE in the *OALECD* in the context of TAKE + NOTE is 'write down' and in Chinese *ji*.
3. The Chinese sequence *ji biji* is felicitous.
4. The Chinese verb *ji* in the *NCCED* cannot be readily translated as *take* in English, or rather it is literally translated as *write down*.

Conclusion: TAKE + NOTE is a non-congruent collocation.

Example 3: PAY + VISIT

1. The Chinese meaning of VISIT is *canguan.*
2. The sense of PAY in the bilingual dictionary in the context of PAY + VISIT is 'used with some nouns to show that you are giving or doing the thing mentioned' and there is no equivalence in Chinese.

Conclusion: PAY + VISIT is a non-congruent collocation.

Using the above procedure, all the well-formed collocations were classified into congruent and non-congruent collocations. For the erroneous collocations, the target collocation (e.g. *acquire knowledge* but not **learn knowledge*) was classified as congruent or non-congruent. Next, the erroneous verb was judged together with the noun collocate to see if it is a word-for-word equivalent of a Chinese expression. If so, it was classified as a Chinese transfer error. Otherwise, it was taken as a non-transfer

error.

Altogether, six tags were designed to classify well-formed and erroneous collocations:

(I, W): non-congruent and well-formed collocations;

(C, W): congruent and well-formed collocations;

(I, N): non-congruent and negative transfer errors;

(C, N): congruent and negative transfer errors;

(I, N-): non-congruent and non-negative transfer errors;

(C, N-): congruent and non-negative transfer errors;

In testing the three hypotheses proposed in Section 9.0, both quantitative and qualitative analyses were performed, with statistical comparisons carried out along three dimensions: (a) within-group comparison of well-formed and erroneous congruent and non-congruent VN collocations; (b) between-group comparison of the well-formed and erroneous uses of congruent and non-congruent VN collocations; (c) within-group comparison of positive and negative L1 influence in verb + noun and adjective + noun collocations. The next three sections accordingly present the results of these three groups of comparisons.

9.2 Within-group comparison of well-formed and erroneous congruent and non-congruent VN collocations

Well-formed congruent collocations are the collocations tagged as (C, W); well-formed non-congruent ones are (I, W). Erroneous congruent collocations include collocations tagged as (C, N) and (C, N-), whilst erroneous non-congruent collocations are (I, N) and (I, N-).

First of all, a within-group comparison of the frequencies of well-formed and erroneous congruent and non-congruent collocations was performed on the group of ST2 learners by using chi-square tests. The results are shown numerically in Table 9-1 (for the frequency of collocation types, see Table 9-2). Table 9-1 presents the tokens of well-formed and erroneous congruent collocations and as well well-formed and erroneous non-congruent collocations used by ST2 learners, along with the percentages of erroneous collocations constituting out of the total number of congruent and non-congruent collocations. Though erroneous collocations did not make up a high proportion of total collocations produced, a chi-square test with Yate's correction was performed and a significant

difference was found between the number of erroneous collocations and collocation types as congruent or non-congruent – that is, there were significantly more errors with congruent collocations than non-congruent collocations (χ^2= 45.39, $p < 0.0001$).

Table 9-1 Well-formed and erroneous congruent and non-congruent collocations in the ST2 (tokens)

Types	Congruent coll.	Non-congruent coll.	Total
Well-formed coll.	727 (88.7%)	740 (97.5%)	1467
Erroneous coll.	93 (11.3%)	19 (2.5%)	112
Total	820 (100%)	759 (100%)	1579

(Note: χ^2 = 45.39, $p < 0.0001$ ***)

Table 9-2 Well-formed and erroneous congruent and non-congruent collocations in the ST2 (types)

Types	Congruent coll.	Non-congruent coll.	Total
Well-formed coll.	117 (69.6%)	104 (88.9%)	221
Erroneous coll.	51 (30.4%)	13 (11.1%)	64
Total	168 (100%)	117 (100%)	285

(Note: χ^2 = 13.59, p = 0.0002 ***)

When collocation types are considered, both the percentages of erroneous collocations among congruent collocations (30.4%) and among non-congruent collocations (11.1%) increase compared with collocation tokens presented in Table 9-1. Following a similar statistical analysis using Fisher's test, a similar result was obtained: there are significantly more errors with congruent collocations and more well-formed non-congruent collocations produced by ST2 learners (χ^2= 13.59, p = 0.0002). This result runs counter to what Nesselhauf (2005) observes, where the percentage of collocation errors among congruent collocation tokens was around 17% and among non-congruent collocations was 42%. The underlying reason for such a substantial difference in the results obtained may be the different criteria adopted in categorising the two types of collocations; namely, congruent and non-congruent. In Nesselhauf's data-set, congruence was measured at the levels of phrasal verbs, prepositions and nouns in verb + noun combinations (2005: 222). That may have led to the classification of more non-congruent collocations which otherwise might be congruent ones. According to Salkie (2002: 56), items in closed grammatical classes normally behave differently across languages. Subsequently, the likelihood of collocation errors with grammatical words (prepositions, the number of noun) increases

due to their often greatly differing use between languages. In this study, congruence was only measured at the level of content words, with colligations disregarded. As was pointed in Chapter 4, only the verbs that were correctly or wrongly used were targeted, and combinations where a noun was wrongly used were not included in the present study.

A difference in L2 collocation performance depending on varying L1 backgrounds has been reported in previous L2 collocation studies (e.g. Biskup, 1992; Wang and Shaw, 2008). Thus, another reason accounting for a difference between Nesselhauf's (2005) finding and ours may lie in the language background of the L2 learners targeted. In her study, the L1 of the learners is German, which, together with the English language, belongs to Indo-European languages, whilst Chinese is one category of Sino-Tibetan languages. Thus, more differences would be expected in Chinese than German when compared with English. Based on this observation, Chinese learners of English may encounter more difficulties in producing congruent collocations than German learners of English.

The above result was obtained for the collocations used by the ST2 learner group. As the same procedure was employed in the data of ST6 learners, similar outcomes were produced when the frequencies of well-formed and erroneous collocations among congruent and non-congruent collocations were compared, i.e. the chi-square test shows that there were significantly more errors among congruent collocations than non-congruent collocations, as a statistically significant relationship was found between the number of erroneous collocations and whether collocations were congruent and non-congruent (see Table 9-3 for token frequencies and Appendix Ⅷ for type information).

Table 9-3 Well-formed and erroneous congruent and non-congruent collocations in the ST6 (tokens)

Types	Congruent coll.	Non-congruent coll.	Total
Well-formed coll.	1047 (88.1%)	618 (96.4%)	1665
Erroneous coll.	141 (11.9%)	23 (3.6%)	164
Total	1188 (100%)	641 (100%)	1829

(Note: $\chi^2 = 33.97$, $p < 0.0001$ ***)

The statistical analyses of collocations produced by ST2 and ST6 learners show that congruent collocations pose more difficulties than non-congruent ones. The greater difficulty with congruent collocations seems to contradict findings from psycholinguistic experiments conducted by Yamashita and Jiang (2010), Wolter and Gyllstad (2011), according to whose findings a group of highly proficient

non-native speakers both processed L1–L2 collocations (i.e. congruent collocations) than L2-only collocations (i.e. non-congruent collocations) with faster reaction times and recognised the former with higher receptive scores. So L2 collocational links in the mental lexicon of L2 learners are likely to be mediated by their L1 and thus congruent collocations gain more legitimacy in the mental lexicon. This is perhaps true in the sense of collocational storage and recognition, but a different picture emerges when this link is activated in the production process.

A detailed analysis was further performed on erroneous congruent collocations, aiming at discovering why congruent collocations seem to pose more difficulties for Chinese learners of English. Erroneous collocations in the ST2 learner group were identified. A detailed investigation into the 93 instances of erroneous congruent collocations in ST2 revealed that a large proportion of the errors were a result of 'partial congruence' between the two languages. Congruent collocations are easier for L2 learners in locating the appropriate verb (e.g. *answer*) with the preselected noun (e.g. *question*) through direct rendering from their mother tongue. As in the case of *answer+ question*, the corresponding English verb – *answer* – is a one-to-one match with the Chinese verb – *huida* (although *reply to* and *respond to* are synonyms of *answer* in this sense and are also in collocational relationship with *question*, they are not believed to be an exact match for *huida*, which is instead rendered in Chinese as *huifu*). However, the problem for L2 learners is that one-to-one correspondence in languages is not prevalent. As shown in the errors in Table 9-4 below, there exists 'differentiation' between Chinese and English, meaning that the native language has one form, whereas the target language has two or more forms (Gass and Selinker, 2008: 100). For example, there is a one-to-many correspondence in the following forms and subsequently errors are induced by such mismatches.

Table 9-4 Erroneous congruent collocations attributable to ‘differentiation’

Chinese words	English words	Error tokens
zuo (做)	make, do (exercise), compose (poem), play (game)	5
kan (看)	see, read (book)	3
chuan (穿)	dress, wear (clothing)	1
shuo (说)	say, tell (joke)	2
ting (听)	hear, listen to (music)	3
dakai (打开)	turn on, open (book)	1
canjia (参加)	attend, participate in (activity)	1
biaoyan (表演)	play, perform (play)	3
chengren (承认)	concede, accept, admit (mistake)	2
chuangzuo (创作)	create, compose (poem)	4
zhuazhu (抓住)	catch, seize (time)	1
Total		26

(Note: Nouns in brackets are given as the noun collocates of the target English verbs).

Additionally, another notable type of congruent collocation error can be attributed to ‘coalescing’, referring to the opposite of ‘differentiation’ where the native language has more than one form corresponding to only one form in the target language (Gass and Selinker, 2008: 100f). Unlike errors attributable to differentiation, where learners seem to have difficulties choosing the right form from several possible forms in the L2, in cases of coalescing, what happens is that L2 learners have to know that among several expressions in their native language (e.g. *huode zhishi* (literal translation: *acquire knowledge*), *xuexi zhishi*: literal translation: **learn knowledge*), only one expression (e.g. *huode zhishi*: English translation: *acquire knowledge*) corresponds to the L2 expression (e.g. *acquire knowledge*). Examples of errors attributable to coalescing are shown in Table 9-5.

Table 9-5 Erroneous congruent collocations attributable to ‘coalescing’

Chinese sequences	English sequences	Error tokens
gei jianyi/ yuanyin(给建议/原因)	give advice/reason	3
shuo jianyi/ yuanyin (说建议/原因)		
huode zhishi (获得知识)	acquire knowledge	27
xuexi zhishi (学习知识)		
zhangwo zhishi (掌握知识)		
chuanshou zhishi (传授知识)	impart knowledge	5
jiao zhishi (教 知识)		
fuxi zhishi (复习知识)	review knowledge	1
jiyi zhishi (记忆知识)		
Total		36

In Table 9-5, there are many-to-one correspondences between the Chinese and English language and only the first sequence in the left column corresponds correctly to the sequences in English. For instance, the concept of *acquire knowledge* can be expressed in Chinese in at least three forms: *acquire knowledge/*earn knowledge, *learn knowledge and *grasp knowledge*. **learn knowledge* has been commonly produced by Chinese learners and negative transfer subsequently occurs. **learn knowledge, *grasp knowledge* together with **teach knowledge* are the commonest Chinese expressions when expressing the concepts of acquire knowledge and impart knowledge.

It can be seen that 66.7% ((26+36)/93*100) of the erroneous congruent collocations are caused by a partial congruence (or in Nesselhauf's (2005) terminology, 'partial non-congruence'), where one-to-many or many-to-one correspondences occur in the native and target languages and only one equivalent of several expressions that can be used in one language is acceptable in another. These two factors are found to be responsible for the susceptibility to errors for congruent collocations. Just as Farghal and Obeidat (1995: 323) pointed out, reliance on L1 does not "always result in positive transfer since the one-to-one correspondence hypothesis holds in only few cases". Relying on the L1 by L2 learners commonly leads to negative transfer.

9.3 Between-group comparison of the well-formed and erroneous congruent and non-congruent VN collocations

Tables 9-6 and 9-7 below present the results of well-formed congruent and non-congruent collocation tokens and erroneous ones produced by ST2 and ST6 learners (for types, see Appendices Ⅸ and Ⅹ).

Table 9-6 Well-formed congruent and non-congruent VN collocations in ST2 and ST6 (tokens)

Types	ST2	ST6	Total
Congruent coll.	727 (49.6%)	1047 (62.9%)	1774
Non-congruent coll.	740 (50.4%)	618 (37.1%)	1358
Total	1467 (100%)	1665 (100%)	3132

(Note: $\chi^2 = 55.85$, $p < 0.0001$ ***)

Table 9-7 Erroneous congruent and non-congruent VN collocations in ST2 and ST6 (tokens)

Types	ST2	ST6	Total
Congruent coll.	93 (83.0%)	141 (85.8%)	234
Non-congruent coll.	19 (17.0%)	23 (14.2%)	42
Total	112 (100%)	164 (100%)	276

(Note: $p = 0.50$ ns)

Likewise, statistical analyses were conducted on the data in both tables above. Two different results were obtained: a significant relationship between well-formed non-congruent collocations and learner groups was found, whilst for erroneous non-congruent collocations, no statistical significance was observed. Significantly more non-congruent collocations were correctly used by ST2 learners than by the ST6 group (both in terms of tokens and types). As is shown in Figure 9-1, there is an increase in congruent collocation uses and decrease in non-congruent collocation uses as learners' proficiency grows. The chances of erroneous congruent and non-congruent collocations increase as well but not sharply. Additionally, the increase in erroneous congruent collocations is sharper than of non-congruent ones as is presented in Figure 9-2.

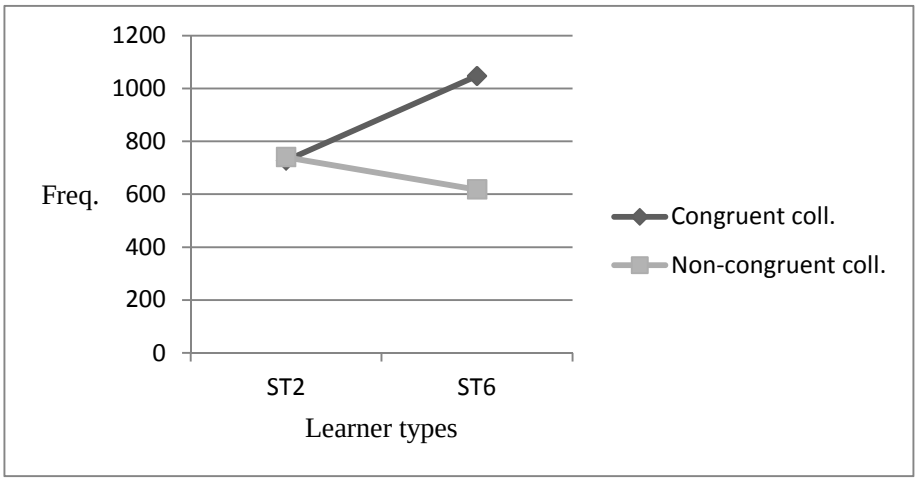


Figure 9-1 Well-formed congruent and non-congruent collocation tokens in the ST2 and ST6

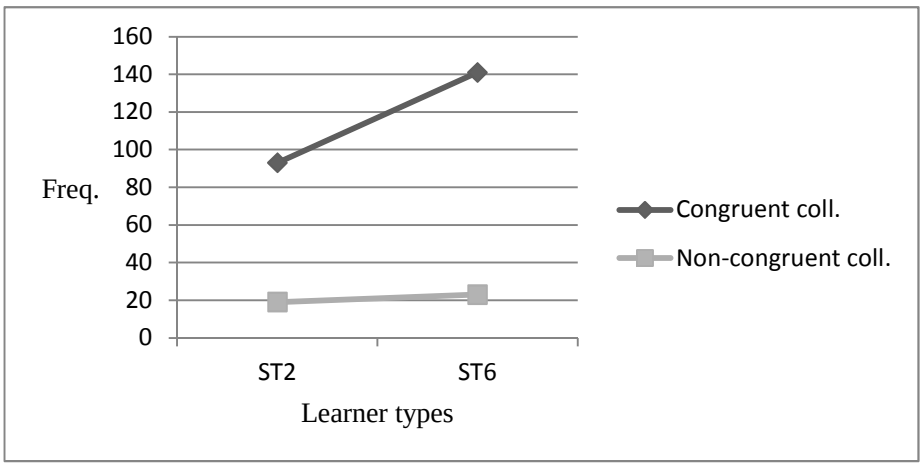


Figure 9-2 Erroneous congruent and non-congruent collocation tokens in the ST2 and ST6

Since there are no direct translation equivalents between non-congruent collocations in learners' L1 and L2, it would be expected that they are more difficult for learners and more susceptible to errors. However, as the data reveals, the number of errors with non-congruent collocations is significantly lower than errors with congruent ones (cf. Tables 9-1; 9-2 and 9-3). It could be that non-congruent collocations are avoided in production, and when there is a need to express the concept, paraphrases are used.[53] For example, in Farghal and Obeidadt's data, subjects tended to paraphrase *light food* as *food little fat, heavy drinker* as *drinks too much*. However, Table 9-1 indicates that there are more non-congruent collocations that are correctly used than congruent ones, so these Chinese learners showed no sign of such an avoidance strategy.

Another reason that non-congruent collocations are less prone to errors may be that they are acquired as wholes and, once acquired, they are less prone to errors. This claim is supported by psycholinguistic evidence that once stored in memory, L2 collocations, like formulaic sequences, are processed independently of the L1 (Conklin and Schmitt, 2008; Jiang and Nekrasova, 2007; Wolter and Gyllstad, 2011; Yamashita and Jiang, 2010). For the present research, evidence was gathered in order to empirically test their claim: to investigate developmentally whether the non-congruent collocations that were correctly used by ST2 leaners were not wrongly used by ST6 learners.

The steps taken in examining the uses of non-congruent collocations were as follows:

a. Check if the well-formed non-congruent collocations in the ST2 were also correctly used by ST6 learners;

b. Check the overlapping non-congruent collocations that were wrongly produced by the ST6 learners

[53] See also the paraphrasing strategy employed by L2 learners in Bahns and Eldaw (1993) and Farghal and Obeidadt (1995).

and the well-formed non-congruent ones in the ST2. Results showed that among the 104 types of well-formed non-congruent collocations in the ST2, 33 types are also correctly used by ST6 learners (the others being absent in the ST6 data). Among the erroneous uses of collocations by ST6 learners, all of them except two cases[54] were absent from the ST2 database of well-formed non-congruent collocations, indicating that erroneous non-congruent collocations at the ST6 level were not produced in large numbers at the ST2 level. This result functions as evidence in support of our claim of holistic learning of non-congruent collocations. To be more specific, if a majority of errors with non-congruent collocations at the ST6 level have been correctly used in the ST2 level, non-congruent collocations may not be learnt holistically and are prone to errors. What we found is that non-congruent collocation errors in the ST6 were not frequently found in the database of well-formed non-congruent ones produced by ST2 learners. In sum, though there are not many overlapping uses of non-congruent collocations, our data showed that non-congruent collocations that are correctly used by lower levels of learners are not susceptible to errors at higher levels.

A detailed look at the non-congruent collocations in our data gives two striking categories of non-congruence: lexical non-congruence and structural non-congruence.[55] The former means that the English collocation and the Chinese expression share the same verb + noun structure, but the verb in question does not have a direct translation equivalent in both languages; structural non-congruence means that the English VN collocation does not correspond to a VN sequence in the Chinese, rather the Chinese uses a single lexeme. *Take + pill* is an example of lexically non-congruent collocation, since the Chinese equivalent *chi + yao* is also a verb + noun structure, yet the Chinese verb *chi* literally corresponds to *eat* in English (**eat + pill*). Other examples of this type include *file + suit* (Chinese: tiqi+ susong), *reach + conclusion* (Chinese: dechu + jielun); *Make + fun*, on the other hand, is a structurally non-congruent collocation, in which the translation equivalent in Chinese is *quxiao*, a single verb lexeme as listed in the *Contemporary Chinese Dictionary* (fifth edition, 2005) (henceforth: *CCD*). *Take + bath* (Chinese: xizao), *set + fire* (Chinese: *shenghuo*) also fall into the category of structurally non-congruent collocations.

Altogether there are 310 lexically non-congruent collocations and 429 structurally non-congruent

[54] The two non-congruent collocations that were correctly used by the ST2 level but wrongly used at the ST6 level are **make (draw) + conclusion* and **put (pay) + attention*, both of which are due to the Chinese transfer.

[55] Nesselhauf (2005: 222) distinguishes two types of non-congruence: lexical and non-lexical non-congruence, the former of which corresponds to the sense in the present study whilst the latter, referring specifically to elements other than lexical words (e.g. prepositions) that are not congruent between languages.

collocations that are correctly used in the ST2. For ST6 learners, the number of tokens of well-formed lexically non-congruent collocations is 161 and of structurally non-congruent collocations, 457. Structurally non-congruent collocations make up a predominant proportion of the total non-congruent collocations that are successfully produced by L2 learners. As Wang (2011) reported in his study of Chinese learners' learning of light verb + noun collocations, learners seldom made errors and seemed to have little difficulty in acquiring collocations where there are structural differences between the forms of expression between Chinese and English.

Past studies of L2 collocation acquisition suggest that learners do not pay attention to collocation relationships and thus collocations are seen as compositional combinations of words rather than as a phenomenon of co-selection (Laufer and Waldman, 2011; Philip, 2007; Wray, 2002). Yet seen through the better performance in non-congruent collocations by L2 learners, special attention might be actually paid when the fully salient non-congruent collocations are encountered in learning. Thus these collocations might be 'restructured' in their mental lexicon and memorised as wholes, bypassing the route of the L1. Wolter (2006: 743-744) argues for this restructuring process by explaining that "in some cases their existing L1 lexical/conceptual network will suffice, and slotting L2 lexical items into the network will be fairly straightforward. In other cases, however, the network itself will need fundamental restructuring in order to accommodate divergent properties". Our finding that non-congruent collocations are less likely to be at fault than congruent collocations and are less prone to errors once acquired by lower levels of learners empirically supports findings from psycholinguistic studies, e.g. once an L2-only collocation (i.e. non-congruent collocation) "is recognised as a legitimate collocation in the L2, it becomes stored as such psychologically and when the first word in the collocation is observed the second word of the collocation is anticipatorily activated" (Wolter and Gyllstad, 2011: 442; also Yamashita and Jiang, 2010). So we argue that non-congruent collocations are learnt holistically, and are not as prone to compositionality as the learning of congruent collocations.

Considering learners' relative ease in acquiring the non-congruent collocations, it might be argued that non-congruent collocations are more 'noticed' than congruent ones. As pointed out by Schmidt (1990: 129), "noticing is the necessary and sufficient condition for converting input to intake". What make them more noticeable are the divergent features between the target language and their mother tongue, e.g. when there is no word-for-word translation equivalent of the L2 collocations in their L1. In this sense, non-congruent collocations receive more perceptual salience (as James (1996) called it)

than congruent ones. In addition, if the non-congruent collocations are processed and stored holistically, the significantly more non-congruent collocations at lower levels of learning and more congruent collocations at higher levels of learning (see Table 9-6) conforms to the model proposed by Krashen and Scarcella (1978) and also Wray (2002): in early stages of L2 learning, formulaic sequences are memorised as wholes but subsequently "language development proceeds analytically, in the 'one word at a time' fashion" (Krashen and Scarcella, 1978: 297). This analytic approach leads to an analysis of a formulaic sequence in terms of individual words, one which does not retain word co-occurring information.

9.4 Within-group comparison of positive and negative L1 influence with VN and AN collocations

Adjective + noun collocations produced by ST2 and ST6 learners were classified into congruent ones and non-congruent ones following the same procedure with verb + noun collocations. Examples of congruent AN collocations are: *active part, absolute truth, bad luck, blue sky*, etc. Non-congruent collocations include *active volcano, narrow escape, promissory note, heavy smoke,* etc. In this section, the role of the L1 in the two types of collocations (VN and AN collocations) was examined in order to see whether its influence is proportionate in different word-class collocations.

It is assumed that congruent collocations are correctly used owing to L1 positive transfer, though this assumption is somewhat arbitrary and speculative, since within the learners' "black box", it is not clear whether congruent collocations are stored and produced wholly without mediation through their L1. However, one may justify this assumption, on the basis that learners take less reaction time and make fewer errors in responding to congruent collocations than non-congruent ones, which suggests that the former are stored in the learners' mental lexicon via L1 mediation (cf. Wolter and Gyllstad, 2011; Yamashita and Jiang, 2010). So the collocations that are a result of positive transfer are collocations tagged as (C, W), and negative transfers are collocations tagged as (I, N) and (C, N).

L1 influence was first measured in the VN collocations produced by ST2 and ST6 learners. The overall number of L1-influenced collocations among the VN collocations produced by ST2 learners was 801 (calculated as the overall collocations tagged as (C, W), (I, N) and (C, N)), which makes up 51% of all the collocations (1,578 tokens of collocations). Similarly, 66% collocations in the ST6 level were

either positively or negatively influenced by the Chinese. In general over 50% of the collocations produced by Chinese learners may be traced to the influence of their L1: this figure was also reported by Wang (2011), in whose investigation of Chinese college students' acquisition of English light verb + noun collocations, 61.84% of the subjects' production of L2 light verb + noun collocations were positively or negatively transferred from Chinese.

Next, an analysis was performed on the negative transfer of VN collocations between ST2 and ST6 learners (see Table 9-8 below for a numeric presentation). It shows a decrease in transfer errors in the ST6 level (from 66% to 29%). In addition, according to statistical analysis, ST6 learners made significantly more non-transfer errors than ST2 learners, which suggests a weakening L1 influence on the production of L2 collocations with learners' rising proficiency. Yet the influence of the L1 is still strong as nearly one third of the errors are L1-induced ones in the advanced level.

Table 9-8 Transfer and non-transfer VN collocation errors produced by ST2 and ST6 learners

Types	ST2	ST6	Total
Transfer errors	74 (66%)	48 (29%)	122
Non-transfer errors	38 (34%)	116 (71%)	154
Total	112 (100%)	164 (100%)	276

(Note: $p < 0.0001$ ***)

For AN collocations in the two groups of learners, the L1 influence was found to be larger than in VN collocations: 95% for ST2 learners and 96% for the ST6 group. Compared with the percentages obtained above in VN collocations, the L1 seems to play a bigger role in AN collocations. In the following analysis, we investigated whether its role is allocated proportionately in positive and negative transfer in the two types of collocation.

Tables 9-9 and 9-10 present respectively the numbers of VN and AN collocations due to positive transfer and negative transfer in the two groups of learners (for types, see Appendices Ⅺ and Ⅻ).

Table 9-9 Positive and negative transfer in VN and AN collocations in the ST2 (tokens)

Types	VN coll.	AN coll.	Total
Positive transfer (C, W)	727 (91%)	263 (97%)	990
Negative transfer ((I, N) and（C, N))	74 (9%)	8 (3%)	82
Total	801 (100%)	271 (100%)	1072

(Note: $p = 0.0005$ ***)

Table 9-10 Positive and negative transfer in VN and AN collocations in the ST6 (tokens)

Types	VN coll.	AN coll.	Total
Positive transfer (C, W)	1047 (96%)	960 (99.8%)	2007
Negative transfer ((I, N) and（C, N))	48 (4%)	2 (0.2%)	50
Total	1095 (100%)	962 (100%)	2057

(Note: $p < 0.0001$ ***)

The two tables reveal that in both groups there is significantly more positive transfer in AN collocations and more negative transfer in VN collocations, irrespective of the tokens and types examined. That there is more negative transfer in VN collocations than AN collocations produced by L2 learners has also been found by Parastuti et al. (2009). They investigated the collocations used by Indonesian English learners of English and reported that among all the negative transfer errors, the percentage of negative transfer for verb (creation + activation) + noun collocations was the largest – 54.24%, and the second largest was adjective + noun collocations – 18.64%.

That AN collocations are less error-prone may be because they are an early acquired type of collocations whereas VN collocations are found to be the most difficult collocations acquired by L2 learners (Gitsaki, 1999). So it may be the relative ease with AN collocations that enables a reduction in negative L1 transfer, while the relative difficulty with VN collocations increases the possibility of L1 interference. Another explanation may relate to the degree of congruence between AN and VN collocations. Of the AN collocations both in ST2 and ST6 databases, 78% and 97% respectively are congruent collocations (e.g. *blue sky*: Chinese: *lan tian*). So one possible explanation is that between Chinese and English adjectives correspondences are more often one-to-one, leading to more successful learning of AN collocations; whilst 'differentiation' or 'coalescing' normally exists between the verbs

in two languages, thus causing more collocation errors.

9.5 Synopsis of findings in this chapter

The findings are summarised with regard to the three hypotheses proposed earlier in this chapter, namely:

1. L2 learners perform better in congruent collocations than non-congruent collocations.
2. Non-congruent collocations that are correctly used by learners at lower levels are not wrongly used by learners at higher levels.
3. The L1 plays a different role in verb + noun and adjective + noun collocations.

For Hypothesis 1, we found that there were more congruent collocations than non-congruent collocations that were correctly used by both groups in either tokens or types (except for the tokens in ST2). However, there were significantly more errors with congruent collocations than non-congruent collocations. So this hypothesis is rejected in light of the data showing that congruent collocations actually posed more difficulties than non-congruent ones. Further, detailed analysis was performed on erroneous congruent collocations so as to locate the factors inhibiting the correct production of congruent collocations. It is found that a large proportion of the errors can be attributed to 'partial congruence' between their mother tongue and English, as in the forms of 'differentiation' and 'coalescing'.

With regard to Hypothesis 2, between-group comparisons on the well-formed and erroneous uses of congruent and non-congruent verb + noun collocations were conducted. This hypothesis is upheld as for non-congruent collocations that were correctly produced by learners of lower levels, they were seldom wrongly used by higher levels.

Concerning hypothesis 3, within-group comparisons of positive and negative L1 influence with verb + noun and adjective + noun collocations were carried out. Statistical analysis revealed that there was significantly more positive transfer in AN collocations and more negative transfer in VN collocations, irrespective of learner types. That indicates that the L1 plays a different role in word-class specific collocations.

In conclusion, our findings regarding the cross-linguistic influence in the learning and production of L2 collocations hold significant implications which can be explored in connection with previous

SLA theory. Discussion of these implications will be presented in Chapter 10.

Chapter 10: Summary and conclusions

The final chapter begins by summarising the key findings reported in Chapters 5, 6, 7, 8 and 9. Then theoretical and pedagogical implications for L2 collocation learning are discussed with a view to findings revealed in this study (Section 10.2). The thesis concludes by mentioning the limitations of the present study and suggesting ways forward in further research into L2 learners' collocation learning (Section 10.3).

10.1 Summary

The mastery of collocations is a key indicator of second language learners' overall proficiency in the field of second language acquisition. The importance of collocational knowledge to the attainment of native-like fluency has been widely acknowledged (e.g. Palmer, 1933; Pawley and Syder, 1983; Wray, 2002). Yet collocation acquisition poses great problems even for fairly proficient L2 learners. Much L2 collocation research has been devoted to an investigation into learners' knowledge and use of collocations, with the finding that there are both quantitative (e.g. overuse and underuse) and qualitative deficiencies (e.g. misuse) in their collocation production. Apart from the deficiencies in collocation uses uncovered by previous studies, collocation is believed to be acquired late and lag behind other aspects of SLA (Henriksen, 2013; Schmitt and Carter, 2004). However, studies identifying the factor(s) associated with the lag of collocational knowledge have been few and our research fills this gap through investigating the relationship between vocabulary growth and the learning of collocations by L2 learners.

A cross-sectional study of the collocation performance of Chinese learners of English at three proficiency levels was conducted. Learners' collocation performance was investigated through their production of three frequent and important lexical collocations: verb + noun, adjective + noun and noun + noun collocations, with the main focus on the most difficult type of collocations – verb + noun collocations. Before pulling together the detailed results, an overall picture of collocation production by Chinese EFL learners at the three proficiency levels is presented.

Corroborating findings from previous L2 collocation studies (cf. Section 3.2.1), both quantitative

and qualitative deficiencies were identified in Chinese L2 learners' performance in verb + noun collocations. Quantitative deficiency was exemplified in the small number of VN collocations extracted from three sub-corpora (only about 5,000 tokens and 1,070 types of collocations retrieved from writings of approximately 600,000 words). That there was an insufficient use of collocations by L2 learners suggests that learners more often combine individual words creatively in language production rather than making use of formulaic language, which further indicates L2 learners' poor sense and command of prefabricated units in language acquisition and production (Foster, 2001; Laufer and Waldman, 2011; Kjellmer, 1991; Wray, 2002). Another aspect of quantitative deficiency lies in the heavy use of a limited range of collocation types.[56] Heavy reliance on a small number of collocations, together with a small quantity of collocations produced, indicated a poor phraseological competence among Chinese learners.

In addition to a quantitative deficiency in collocation production, learners' problems with collocation learning have been identified through the large proportion of collocation misuses. Nearly a quarter of VN collocations were recognised as erroneous collocations. Successful production of VN collocations not only posed problems for all levels of learners, but there was no sign of improving collocation performance with rising proficiency. On the one hand, there was no sign of increasing collocation production from the ST2 to the ST6 level (cf. Section 5.1.1). On the other, in terms of the erroneous collocations produced, though there was a significant decrease in collocation errors from the ST2 to the ST5 level, errors significantly increased again from the ST5 to the ST6 level. This uneven developmental path was attributed to the learning of more verbs at the ST6 level, which inhibited the acquisition of collocations (cf. Chapter 6). As the overall trend predicts, collocational knowledge did not improve since learners at the ST6 level, despite advancing L2 proficiency, made about the same proportion of errors as the lowest ST2 learners.

The thesis went on further to explore the role of the verb increase in this collocation lag. Verb increase was firstly measured broadly in terms of the development of verbs from delexical to lexical verbs. Investigation into delexical verb + noun and lexical verb + noun collocations revealed that there was a gradual increase in the production of lexical verb + noun collocations and decrease in delexical verb + noun collocations among the three levels. A significant relationship was found between the numbers of well-formed lexical verb + noun collocations and learner levels, indicating a significantly higher production of lexical verb + noun collocations with the rise of proficiency. However, in erroneous

[56] For example, 14% collocation types were used more than 10 times by learners at the lowest level, making up 67% of all the collocations retrieved.

lexical verb + noun collocations, there was a dip first from the ST2 to the ST5 level, followed by a sharp increase from the ST5 level to the ST6 level. ST6 learners produced significantly more lexical verb + noun collocation errors than ST2 learners, indicating poorer performance in lexical verb + noun collocations than delexical verb + noun collocations. This means that with more lexical verbs learnt, the chances of these lexical verbs leading to collocation errors increased as well.

When the lexical verbs in both well-formed and erroneous lexical verb + noun collocations produced by all levels of learners were arranged into synonym sets, it was found that collocation errors were seldom made where there was no growth in verb synsets. However, there was an increase in collocation errors in synsets with a verb increase. As learners proceeded to more advanced levels, the occurrence of collocation errors was found to become more and more limited to synsets with verb increases. Verb classes most susceptible to errors were verbs of creation, *fulfil* verbs, verbs of obtaining and verbs of putting, where there was a considerable increase in the number of verbs at the higher level. A marked lag in learners' knowledge of VN collocations was observed, as more proficient learners produced the same proportion of errors as learners of lower levels in terms of the synsets identified. When verbs in these sets were divided into new and old verbs, errors with new verbs at the ST6 level were significantly more likely to be made than errors with old verbs. Therefore, we conclude that the increase in verbs in a particular semantic domain is an inhibiting factor for the learning of collocations: it was suggested that learners may only have an incomplete command of the semantics of the new verb, i.e. the basic meaning of that verb is acquired but not its distinguishing features as distinctive from a set of semantically related verbs. This study suggests that acquisition of verb semantics is important for successful learning of L2 collocations.

In addition to verb growth as an inhibiting factor in collocation learning, further analysis was performed to examine whether newly acquired nouns were also a factor responsible for the lag. Results showed that the percentage of new nouns in erroneous collocations produced by higher levels was rather low – around 13%, a figure which remained roughly constant at both ST5 and ST6 levels. That means in a majority of newly acquired nouns, VN colocations were target-like. Even though new nouns were used in erroneous collocations, it was found that this was not mainly due to a shortfall of new verbs collocating with the newly acquired nouns; in fact the target verbs may have been acquired (e.g. **stir + consciousness* instead of *raise + consciousness, *reflect/cast + prejudice* instead of *hold + prejudice*). So the occurrence of new nouns is not an inhibiting factor for the stagnant development of L2 learners'

collocational knowledge.

This thesis has also investigated learners' performance in two other frequent types of collocations: adjective + noun and noun + noun collocations. Better performance was discovered in the production of adjective + noun and noun + noun collocations than verb + noun collocations. A comparison of the ratios of erroneous collocations among the three types of collocations showed that L2 learners, irrespective of proficiency level, performed best on noun + noun collocations, followed by adjective + noun collocations, and performed worst on verb + noun collocations. Not only was a better performance was observed on AN and NN collocations produced by the three levels of learners, but there was also a clear progression overall in collocational knowledge with regard to these two types of collocations. However, learners' knowledge of VN collocations, lagged, as we saw.

This finding of differing collocation performance depending on category type has contributed to answering the question raised by Siyanova and Schmitt (2008: 453) – whether other types of L2 collocations (e.g., verb–noun, verb–adverb) would be produced at a similar level as adjective + noun collocations. The answer to their question is negative as Chinese L2 learners had much better command of noun + noun collocations than adjective + noun collocations, and better command of adjective + noun collocations than verb + noun collocations.

The present study attempted to account for such differing performance in different types of collocations in terms of vocabulary growth within synonym sets. Classifying verbs into synsets was found to be more natural than adjectives and nouns in the collocation databases. Combining with synonym analyses of the words in WordNet, and a study of the synonym density of randomly-selected verbs, adjectives and nouns used by learners, it was discovered that synonym density of the three types of words is on a decreasing scale. That semantic property may account for L2 learners' better performance in AN and NN collocations and worse performance in VN collocations. In this regard, the prediction that vocabulary growth is an inhibiting factor in collocation acquisition was again upheld.

As an important factor that cannot be ignored in L2 acquisition, the role of L1 in collocation learning was also examined in our study. Contrary to Bahns's (1993) claim (cf. Chapter 3) that only collocations which are non-congruent with learners' L1 collocations need to be taught to learners, we found that congruent collocations were more prone to errors for Chinese learners of English. That was because cases of one-to-one correspondence between the two languages are few and partial congruence is common between the two languages, i.e. differentiation (one-to-many correspondence) and

coalescing (many-to-one correspondence). As for non-congruent collocations, it was found once they were acquired, they were seldom susceptible to errors. L1 was also found to play a different role depending on the types of collocations, as we observed that there was more negative transfer for verbs in verb + noun collocations and more positive transfer for adjectives in adjective + noun collocations.

The main findings of our research are: (a) vocabulary growth was identified as a factor responsible for the stagnant collocation performance in verb + noun collocations; (b) learners performed differently in verb + noun, adjective + noun and noun + noun collocations, with verb + noun the most difficult to acquire, and noun + noun collocations the easiest; (c) learners' L1 played a different role depending on the types of collocations, i.e. more negative transfer in verb + noun collocations and positive transfer in adjective + noun collocations.

These findings contribute to a more comprehensive understanding of collocation learning by second language learners. Most importantly, our research has identified the vocabulary growth factor as an inhibiting force in collocation acquisition, i.e. the learning of new semantically related verbs in a synset leads to more collocation errors. In this regard, it goes beyond previous L2 collocation studies by providing an explanation for the stagnant development of collocation knowledge which has been widely reported (cf. Chapter 3). Put simply, the more words they learn, the more likely learners are to make collocation errors. Furthermore, the misuses of semantically related words in collocations by L2 learners indicate what has been acquired of the semantics of the erroneous verb and what has been not. In the case of **implement* + *act*, the core meaning of *implement* – 'to carry out/do' – was acquired but not its distinguishing semantic property – 'to ensure that what has been planned is done'. In this regard, our finding has contributed to illuminating an aspect of second language acquisition – the linguistic target of learning. Through the misuses of semantically related verbs in synsets in verb + noun collocations, we found that learners failed to produce the correct collocation when they only had acquired the core meanings of a verb but not the distinctive semantic properties of that verb. Therefore, collocation acquisition requires complete acquisition of the semantics of a word.

Additionally, different from previous studies uncovering learners' difficulties with non-congruent collocations either in the production process (cf. Nesselhauf, 2005), or in the collocation recognition process (cf. Yamashita and Jiang, 2010; Wolter and Gyllstad, 2011), our finding in terms of L1 influence shows that congruent collocations deserve more attention than non-congruent collocations for Chinese L2 learners. It also finds that non-congruent collocations once acquired, are less prone to

errors, which, from the production perspective, verifies previous findings obtained from psycholinguistic experiments. In light of these, this thesis contributes to a more comprehensive understanding of the L1 influence on the learning of L2 collocations. In the next section, theoretical and pedagogical implications for L2 collocation learning will be discussed based on our findings.

10.2 Implications

10.2.1 Theoretical implications

Although our data are based on L2 learners' production, indicating that inferences drawn from this study about psycholinguistic aspects in learners' mental lexicon are tentative, the findings in terms of the vocabulary growth factor and the role of L1 in L2 learners' collocation learning hold theoretical implications for the lexical organisation in the mental lexicon of L2 learners of English and the cross-linguistic influence in L2 collocation learning.

10.2.1.1 Implications for lexical organisation in the L2 mental lexicon

The finding that learners misuse semantically similar words in collocation production indicates that words are primarily semantically linked in the L2 lexicon. Most words are stored in the mental lexicon via the establishment of semantic associations and semantically similar words are stored nearby (e.g. Channell, 1988; Howarth, 1998a; Wolter, 2001; Zareva and Wolter, 2012). Language production involves the selection of appropriate words according to the meaning to be conveyed. Psycholinguistic evidence has been gathered "in favor of a psycholinguistic model in which words with like meanings are 'close together' in accessing terms" (Channell, 1988: 90; cf. Albert and Obler, 1978). So in the production of word combinations, a choice among the alternatives of a group of semantically related words has to be made. The clustering of words with similar meanings thus produces an interference effect in selecting the right words. Even native speakers encounter semantic interference in producing the target collocations. This interference effect is observed in the mis-collocations produced by native speakers (either intentionally or unintentionally). Evidence concerning native speakers' collocation

misuse is sparse in the literature. Howarth (1998a) is among the few who investigated both the NSs' and NNSs' phraseological errors. He proposes two types of plausible explanations to account for the occurrences of lexical mis-collocations produced by NSs: collocational overlaps and blends. In **draw a contrast*, the error can be seen as the result of filling in a collocational gap within a partially overlapping cluster: *draw a distinction, make a distinction, make a contrast* but not **draw a contrast*. In **place weight*, the error arises out of a blending of two pairs of collocations: *place emphasis* and *attach weight*. Approaching these errors from the perspective of the semantics of the erroneous verb and the target verb, we get the generalisation that they are semantically related verbs, belonging to the synsets identified in the present study. The verbs of *draw* and *make* are in the same set denoting verbs of creation. *Place* and *attach* are in the same set of verbs of putting. For other L1 collocation errors given by Howarth, such as **reach a justice*, the verb *reach* and the target verb *achieve* fall in the same set of verbs of obtaining (cf. Section 6.1). For another set of verbs listed by Howarth, e.g. *compile, draw up, make, produc*e and *write*, they are semantically related as verbs of creation and it follows that collocation errors are made by native speakers through wrongly selecting one of them (e.g. *compile*) to collocate with a noun (e.g. *memorandum*).

Similar to the semantic interference for native speakers in selecting the right word from a set of semantically similar words, L2 learners encounter the same interference as the expansion of their vocabulary. However, there is a fundamental difference in the semantic interference effect between NSs and NNSs. Native speakers may deviate from standard collocational forms either deliberately or unintentionally and in fact there are only a very small number of erroneous collocations produced by NSs (Howarth, 1998a). However, for L2 learners, the semantic interference effect is stronger owing to an incomplete acquisition of the semantics of the semantically related words. As the empirical evidence obtained in the data shows, learners' collocation production gets worse when new semantically related words are learnt (e.g. **concede + mistake* rather than *admit + mistake*). As Chapter 6 reports, verbs in synsets increased dramatically with the rise of proficiency, so did the increase in collocation errors. Misuses of verbs such as *conduct, commit, accomplish, enforce, implement* and *perform* are believed to be caused by the partial acquisition of the core meanings (i.e. 'carry out' or 'do') but not their distinctive meanings.

That semantic confusion increases with the rise of proficiency has been found in several studies (e.g. Agustin Llach, 2011; Ringbom, 1987; 2001). In a developmental study of the kinds of lexical errors

that appeared in the written production of young Spanish learners at two different stages, Agustin Llach (2011) observed a statistically significant increase in semantic lexical errors at the higher proficiency level, which included calques (literal translation of the word from the L1 to the L2) and semantic confusion (the confusion of semantically related words, e.g. * my bedroom is *great* (*great* for *huge* or *big*)). These results support that "well-developed lexicons are dominated by paradigmatic associative connections" (Zareva and Wolter, 2012: 60). Psycholinguistic research into the lexical organisation of L2 learners' mental dictionary indicates that "the same class (paradigmatic) connections become more prominent as the proficiency of L2 learners of English increases to an advanced level" (ibid: 59).[57] Through word association tasks, Zareva and Wolter found that with the advance of proficiency, NNSs' lexicon becomes more paradigmatically dominated like NSs'. Synonymy is an important paradigmatic response and L2 learners' mental lexicon becomes organised more like a thesaurus in which words with similar meanings are stored together (Meara, 1978; Zareva and Wolter, 2012). Therefore, the more proficient learners become, the larger this thesaurus is, and the more semantic interference they are confronted with in choosing the right word from a set of semantically related words.

10.2.1.2 Implications for cross-linguistic influence in L2 collocation learning

L2 learners are not only confronted with the semantic interference from vocabulary increase along the paradigmatic relations, their L1 lexical network exercises a considerable influence over the learning and production of collocations. Thus another inference about cross-linguistic influence can be considered from our study.

Firstly, a higher error rate with congruent collocations than non-congruent ones, even for advanced learners, suggests a consistent role of the L1 in producing collocations even for proficient NNSs. As has been discussed in Chapter 3, the active role of the L1 in collocation production is confirmed in psycholinguistic experiments where a 'dual-activation' takes place: an L2 word stimulates not only its collocates, but also its L1 translation equivalent and L1 collocate (Wolter and Gyllstad, 2011). Given that "even for advanced L2 learners, the L1 continues to be active even when performing tasks entirely in the L2" (ibid: 443), it seems that apart from the receptive process in

[57] Paradigmatic relations between words refer to words of the same lexical class that can substitute for another in a syntactic string (e.g. synonyms, antonyms, meronyms, hyponyms, etc.) (Zareva and Wolter, 2012: 44).

primed lexical decision tasks, in the actual collocation production process, L2 learners' L1 still plays a predominant role through interfering and mediating collocation production. The significant traces of the L1 in L2 collocations, and the large number of transfer errors well attest the active role of L1 on the production side.

The consistent role of L1 in collocation production may be closely linked with the asymmetric cross-language connections. As the Revised Hierarchical Model (Kroll and Stewart, 1994; cf. Chapter 3) predicts, the link from L1 to conceptual memory is assumed to be stronger than the link from L2 to conceptual memory, and the lexical link from L2 to L1 is assumed to be stronger than the lexical link from L1 to L2. Then it seems highly likely that in the production process, L1 is firstly activated prior to the production of L2 words (see also the 'dual-activation' in Wolter and Gyllystad (2011). Yet not all aspects of the L1 are easily activated in producing an L2. As is shown in Jiang's (2000) model, the lemma information (containing semantic and syntactic information) of the L1 is copied into the L2 lexical entry. This is a stage called L1 lemma mediation stage where a majority of L2 words fossilise at this stage (Jiang, 2000). Thus based on Jiang's model, L2 lexical information at the lemma level is in turn most likely to be influenced by the L1. For producing L2 collocations, which are word combinations representing syntactic and semantic relationships between lexical items, the L1 thus plays the most significant role for L2 learners. For example, *acquire knowledge* as a word combination involves both the semantics of *acquire* and *knowledge* and the syntactic information of *acquire* as a transitive verb and *knowledge* as an uncountable noun. With the storage of L1 semantics and syntax at the lemma level (e.g. for both the L2 words *acquire* and *knowledge*), production of word combinations involving semantics and syntax in the L2 (e.g. *acquire knowledge*) is easily mediated through L1 semantics and syntax and thus L1 interference occurs in L2 collocation production. Then L2 lexical combinations are the most susceptible to L1 influence compared to other aspects of language acquisition (e.g. morphology and phonology). Additionally, syntactic and phonological constructions are always finite compared with L2 lexical combinations. So building syntagmatic connections between words in an L2 is complicated by the infinite number of collocations, as well as by the influence from L1 collocational knowledge (cf. Wolter, 2006).

Reliance on the L1 lexical network underlies the large number of fortuitous well-formed collocations that share direct translation equivalents between the L1 and L2, but at the same time the occurrence of erroneous congruent collocations. As discussed in Section 9.2, types of mismatches like

'differentiation' and 'coalescing' make direct copying of L1 word combinations to L2 collocations error-prone. For collocations that have no direct translation equivalent between languages, the shared conceptual system may be the same, but features of word combinations differ (cf. Kroll et al., 2010). Both the empirical data in the present study and the experimental data obtained by Yamashita and Jiang (2010) and Wolter and Gyllstad (2011) suggest that once these non-congruent collocations are acquired, they are processed independently of the L1. It is speculated that with regard to the acquisition process for non-congruent collocations, no direct access is gained in the process of exploiting the existing L1 corresponding lexical network, so new connections between words in the L2 will have to be made and thus a restructuring process begins to accommodate the idiosyncrasies between languages (Wolter, 2006: 745). This restructuring may take the form of a 'noticing process' and may accordingly contribute to the memory of these non-congruent collocations as holistic units. Thus in the production process, these lexical combinations are directly produced from the concept to L2 collocations, without mediation of learners' L1. Congruent collocations, however, become non-salient in the learning process and are mediated through the translation of their L1. So in the production process, these collocations are mediated through their L1 and thus either positive or negative transfer takes place. As learners' proficiency rises, the link between L2 and the concept is stronger, resulting in the far fewer non-transfer errors in higher level learners (see Figure 10-1 below for a summary).

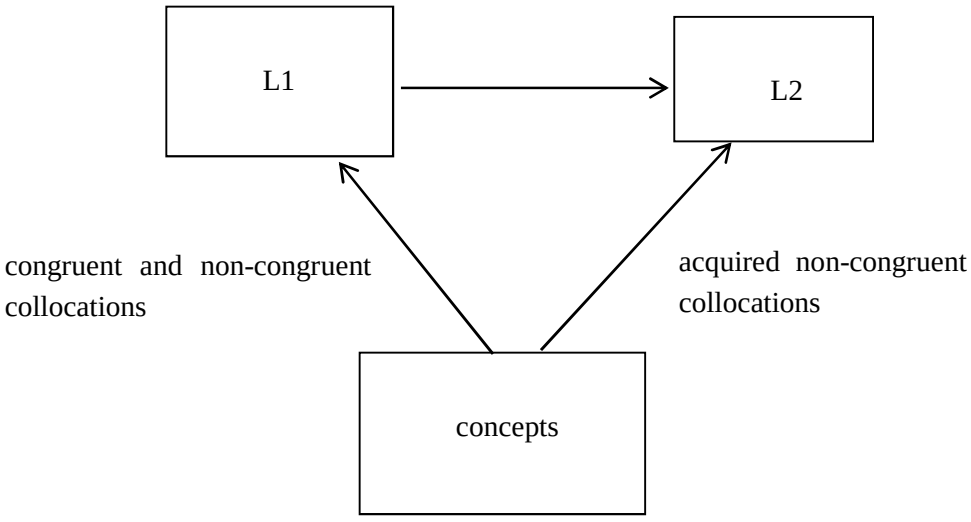


Figure 10-1 Processes for the production of congruent and non-congruent collocations by L2 learners

The above figure synthesises empirical findings of the present study and also conforms to findings from psychological experiments conducted by Yamashita and Jiang (2010) and Wolter and Gyllstad (2011). According to the model, non-congruent collocations that are acquired enjoy a separate and direct route to the production of L2 without word-for-word mediation of the L1. The other collocations are likely to be mediated through a one-to-one translation equivalent in the L1 (though this is especially the case for lower levels, it is probably also the case for proficient NNSs). This is the claim made by Wolter (2006: 743) that learners might bypass the L2 acquisition process through relying on their L1 lexical network when there is a marked overlap between the L1 and L2 lexical networks. For the non-acquired collocations, learners tend to resort to word-for-word translation from the L1 and as a result infelicitous collocations are produced. Where there happens to be a one-to-one correspondence in the L2 (e.g. *answer + question*), the congruent collocation translated from the L1 to L2 is a well-formed one; where there are mismatches (e.g. one-to-many correspondence ('differentiation') and many-to-one correspondence ('coalescing')), collocation errors are likely to occur.

In conclusion, one general inference drawn from our study is that two factors pose particular problems for L2 learners in learning L2 collocations: semantically related words in the L2 and learners' L1 lexical network. Non-natives encounter much greater difficulties in producing collocations than native speakers, since on the one hand, their L1 lexical/conceptual knowledge has a consistent influence on how learners structure connections between words in an L2 (Wolter, 2006); on the other, expanding paradigmatic relations of words (e.g. synonymy relations) in their vocabulary in the course of L2 acquisition means that more and more words are stored in the L2 mental lexicon, thus exerting interfering forces in the word selection process. This is one of the major implications drawn from our study. Our results show that learners are confronted with a dilemma: on the one hand, their production of collocations is characterised with a limited number of collocation types, indicating an inadequate mastery of vocabulary; on the other, the increase in vocabulary in turn inhibits the learning of collocations. In other words, the growth of vocabulary in the paradigmatic relations (i.e. sets in the terminology of Carter and McCarthy (1988: 210)) enables learners to have more varied choices but at the same time produces an interfering effect in learning collocations. Words within sets are in relationships of synonymy, antonymy, hyponymy, etc. Sets are believed to be "powerful organising principles, and have a strong psychological reality for language users and learners" (ibid: 211). In this regard, the synonym sets identified in the learner data are not only the organising principle for

semantically related words, but also the interfering factor in selecting the appropriate word to collocate with another word. Therefore, it is important for learners to acquire not only the shared semantic element of a word in a group of semantically related words, but also to acquire the distinguishing semantic contents of that word in order to differentiate it from its synonyms. The next section will be devoted to a discussion of pedagogical implications for collocation learning mainly in terms of a full mastery of word semantics.

10.2.2 Pedagogical implications

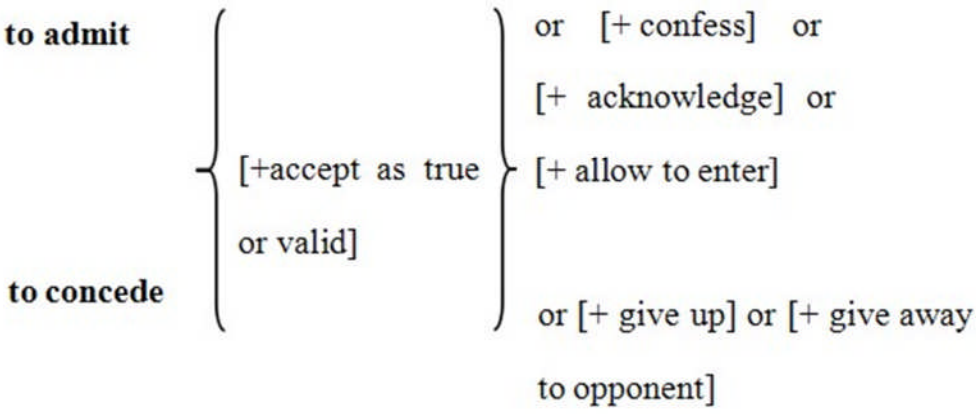
10.2.2.1 Acquisition of verb semantics

How collocations are learnt and what factors interfere with the learning of collocations can shed much light on how collocations are best taught and learnt. Findings in the present research hold a number of important pedagogical implications.

First, verb + noun collocations deserve special attention compared with AN and NN collocations since they are more error-prone. Second, within verb + noun collocations, it is the verb that poses more problems than the noun for L2 learners (Granger, 2014; Nesselhauf, 2005). So verbs deserve more attention in vocabulary learning. As is discussed in the previous section, the misuse of verbs with semantic relatedness (e.g. *implement* and *perform*, *admit* and *concede*) in collocations means that it is important to fully acquire the semantics of verbs: both their basic properties ("semantic markers" in the parlance of Katz and Fodor, 1963) and the features that distinguish them from its semantically related words ("distinguishers", Katz and Fodor, 1963).

An efficient learning of verb semantics has to take into account how the semantics of words is approached. A traditional view of word semantics is to break down word meaning into a number of abstract components or semantic features and identify "those features that will distinguish the meaning of any one word from every other that might … compete for a place in the same semantic territory" (Cowie, 2009: 57). This approach to the meaning of a word is known as Componential Analysis. CA has long been used to describe and distinguish words with semantic relatedness. For example, Rudzka et al. (1981; 1985) presented words in sets whose members have similar meanings and distinguished them through componential grids (and collocational grids as well). With the example of *admit* and *concede*,

which occur in our database as **concede + mistake (admit)*, the componential grid given by Rudzka et al. (1985: 171) is as follows:



The semantic marker of both *admit* and *concede* is "accept as true or valid", but the distinguishers of *admit* are "to confess or to acknowledge or allow to enter" and of *concede* are "to give up or give away to opponent" (ibid.). This way of contrasting the semantic features of semantically related words works well for linguistic analysis but is not suitable as a language-teaching tool, as the features are always abstract (Carter and McCarthy, 1988). As one semantic component of *admit* – "to confess", the word *confess* might be more complex to understand than *admit*. Meanwhile, the decontextualised presentation of meaning components in words or phrases (e.g. "accept as true or valid" for both *admit* and *concede*) makes it hard for L2 learners to comprehend.

So the acquisition of verb semantics is better aided through a contextualised display of its meaning. The learning of word semantics in contexts has been widely advocated (Cobb, 2003; Hanks, 1996; Hoey, 2000; Laufer, 2006). One macro context for learning the semantics of a word is its co-text, as the full sentence definition of the headword adopted by the *Collins COBUILD English Dictionary* (1995). The definition provides "much of the context necessary for the meaning of the word in use in the language, dependent on its environment, to be properly appreciated" (Barnbrook, 2007: 190). For example, the *Cobuild* dictionary defines *implement* and *perform* in the following way:

Implement: If you implement something such as a plan, you ensure that what has been planned is done.
Perform: When you perform a task or action, especially a complicated one, you do it.

Seeing through the meanings of the two verbs, *implement* implies more than "carrying out/do

something"; it also incorporates the meaning of "carrying out what has been planned". When learners are presented with the definition of *implement*, the possibility of them making errors like **implement* + *act* found in our study may be reduced.

Another way beneficial for the learning of verb semantics, especially for learning the semantics of semantically related words is through their collocates (cf. Carter and McCarthy, 1988; Lee and Liu, 2009; Xiao and McEnery, 2006). Collocations contribute to the understanding of the concept of a word through defining its semantic area (Brown, 1974; Nattinger, 1988). The learning of lexical semantics and learning of collocations are mutually beneficial and inseparable. Knowing a word involves knowing which words it usually collocates with, and to know the collocational behaviour of a word is one type of word knowledge necessary for a complete acquisition of that word (cf. Nation, 1990: 31). Learning word meanings through collocates contributes to the comprehension of the semantics of that word and an acquisition of the semantics in turn helps define its co-occurring words. The inseparability of the learning of semantics and collocational behaviour is best manifested in Lewis's (1997: 97) view that "the real definition of a word is a combination of its referential meaning and its collocational field". For semantic sets, display of overlapping collocates and of collocates exclusive to a particular word is helpful for learners to both learn the common meaning of a group of words, and the distinguishing features of each word. Such an approach to learning semantically related words has been advocated by Rudzka et al. (1981; 1985). The following is an example of their presentation of collocational grids for synonymous words:

	one's mistakes	that one is guilty	having misled sb	that one did sth	sb to a place	that sb else is right	a point	sb's claim to sth	an election	to one's opponent
admit	+	+	+	+	+	+				
concede						+	+	+	+	+

Collocational grids like this may not only help learners get the common meanings of verbs in a semantic field, but also the different nuances of meanings. It is beneficial for learners to identify the distinguishing meanings through the individually tailored collocates. This way of learning may be much better than a decontextualised word learning, i.e. memorising the meanings of words in word lists or through translation equivalents in the L1. Learning and teaching in word lists would unavoidably lead

learners to believe that the collocates of synonymous words in a list share many collocates (Hoey, 2000), which further leads to the semantic confusion in producing collocations (e.g. the misleading belief that *implement* and *perform* share the collocate *act*). Likewise, learning words through translation equivalents leads to the same problem of assuming similar collocates of semantically related words. As Meara (1982) points out, learning vocabulary does not just involve pairing L2 words and L1 meanings as the end state of learning that word. With the words *perform* and *implement* as an example, they are translated into the same word in Chinese according to the *Oxford Advanced Learners' English-Chinese Dictionary*, but the distinguishing features are lost in the Chinese translation equivalent. So with the same translation equivalent, the collocational behaviour of semantically related words is highly likely to be believed as the same by L2 learners. Psycholinguistic studies have found that L2 learners have same-translation pairs stored nearby in the mental lexicon and have difficulties distinguishing their meanings (e.g. *hat-cap, problem-question*) (Jiang, 2002). Therefore, it is far from enough to acquire the semantics of an L2 word on the basis of its translation equivalent. Instead, both a full sentence definition of the verb and words in syntagmatic relations with the verb can be presented to L2 learners for a complete acquisition of its semantics.

10.2.2.2 Consciousness-raising

The finding of a stagnant development in collocational knowledge as learners advance to a higher level clearly indicates that collocations deserve more attention from both L2 learners and foreign language teachers. But what is often the case is that L2 learners do not pay attention to collocational relationships between words, as collocations are largely semantically transparent (Bahns and Eldaw, 1993; Laufer and Waldman, 2011; Martelli, 2006; Nesselhauf, 2003; 2005; Wray, 2002). The transparent nature of collocations means that they usually pose no problems for comprehension, but in the production process, semantic interference from a set of semantically similar words leaves learners in a state of not knowing which one to choose to form an appropriate word combination. A lack of awareness of collocations is not only inferred from a large number of collocation errors identified in our study, but also confirmed through Nesselhauf's (2005) finding of no improvement in collocation competence for learners with more exposure to English. That neither the length of a learner's exposure to English, nor the use of a dictionary seemed to have a significant effect on the overall number of

collocations and the number of collocation errors suggests an unawareness of the collocation phenomenon on the part of L2 learners (Nesselhauf, 2005).

An awareness of collocations, on the contrary, facilitates collocation learning. This can be inferred from our study that learners perform better on non-congruent collocations than congruent ones. Non-congruent collocations in the L2 do not have word-for-word translation equivalents in the L1, thus more conscious learning may be involved. In other words, non-congruent collocations may be salient to L2 learners and thus trigger deeper processing. Congruent collocations, on the contrary, are accessed directly in the L1 and require less processing. According to the framework of human memory (Craik and Lockhart, 1972), greater "depth of processing" (i.e. conception of a series of processing stages in the perceptual analysis of stimuli) implies more semantic or cognitive analysis. Accordingly, deeper levels of analysis lead to longer lasting and stronger traces. Therefore, a deeper processing means more chances of a non-congruent collocation to be stored permanently.

Another factor leading to a lack of awareness of collocational relationships may be due to the traditional teaching of vocabulary as single lexical items in English instruction practices (Farghal and Obeidat, 1995; Siyanova and Schmitt, 2008). One consequence of focusing on single words is the combining of individual words completely on the basis of syntactic rules in text production, operating on a principle called the "open choice principle" by Sinclair (1987). It is necessary for learners' attention to be diverted from single lexical items to habitual word combinations. Only after collocations are paid attention to, can they have the chance to become acquired, since "noticing is the necessary and sufficient condition for converting input to intake" (Schmidt, 1990: 129). However, as discussed earlier, the transparent nature of collocations means they often pass unnoticed by learners themselves in language input, thus collocations should be made explicit to learners in teaching materials and in classroom activities.

In the process of raising learners' awareness of collocations, one aspect arising from our results should not be ignored – the influence from learners' mother tongue. Considering the persistent role L1 plays even at advanced levels, it is useful for learners to be aware of L1-L2 differences in learning collocations. As James (1996: 147) proposes, translation is a particularly good way to raise L2 learners' cross-linguistic awareness. The teaching/learning of L2 collocation with reference to L1 collocational patterns has been previously suggested (e.g. Bahns, 1993; Chi Man-lai et al., 1994; Granger, 1998a; Martelli, 2006; Nesselhauf, 2003; 2005; Xiao and McEnery, 2006). Different from the view that

collocations with no translation equivalents in the mother tongue shall be paid particular attention to (cf. Bahns, 1993, Nesselhauf, 2003), our study showed that congruent collocations posed more difficulties in production, since one-to-one correspondences between languages are few. Therefore, special attention needs to be paid to congruent collocations in which differentiation (one-to-many correspondence from the L1 to the L2) and coalescing (many-to-one correspondence from the native language to the target language) exist (cf. Chapter 9). The focus on congruent collocations requires both materials writers and teachers to be familiar with the cross-linguistic differences between collocations before diverting learners' attention to particular collocations. With *acquire knowledge* as an example, teachers can place special emphasis on the verb *acquire* as a collocating verb for *knowledge*, rather than other verbs such as *learn, study, master*, although these verbs are legitimate verb collocates for the Chinese word combination with *knowledge*. In the meantime, learners can be encouraged to apply a "back translation" strategy in learning collocations. When they encounter *acquire knowledge*, they can translate the whole collocation into the mother tongue (as they will always unconsciously do in learning an L2), but this is not the end state. It would be beneficial for learners to translate the previously learnt word combinations from their mother tongue back to English, without looking at the English translation, in order to be more aware of the cross-linguistic differences and ultimately to be conscious of the appropriate L2 collocation. This contrastive analysis of collocations can be facilitating from a purely psycholinguistic perspective: collocation learning requires not only noticing, but also more "cognitive depth" (Craik and Lockhart, 1972). Therefore, collocations may ultimately enter the long-term memory since more retention is gained in the learning process. In instructional practices, contrastive analysis of collocations in terms of L1-L2 similarities/differences has been proved by Laufer and Girsai (2008) to be more effective than teaching methods ignoring these cross-linguistic similarities and differences between two languages.

10.3 Limitations and ways forward

One limitation of this study lies in the learner corpus adopted for data collection and analysis. Results obtained in the study are based on a corpus of the English writings by learners of one mother tongue – the Chinese. So all the generalisations made in this research are on the basis of data restricted to one learner type. Researching collocation performance by learners speaking other L1s would have

been more rewarding, as comparisons can be made between collocation performances by learners of different mother tongues. With data obtained from more learner types, both collocation uses typical to one individual learner type and collocation patterns common to learners of various L1s can be found.

A further limitation regarding the learner corpus adopted is concerned with the properties of the learner data. The Chinese Learner English Corpus is a collection of writings by learners at different learning stages. So it is cross-sectional rather than longitudinal. A longitudinal learner corpus would help us to arrive at more definite conclusions regarding L2 learners' collocational development and the factor of vocabulary growth in collocation learning. Yet due to the unavailability of a learner corpus at the time of beginning this research, a corpus recording the writings by learners of different proficiency levels was used.

Nonetheless, despite adopting a quasi-longitudinal corpus, there is a clear differentiation in proficiency levels. The ST2, ST5 and ST6 learner groups (viz. middle school students, English majors of lower grades and English majors of higher grades), which are assumed to be in a continuous development based on the years of English instruction they received, were found to be in a continuous developmental stage. Several indicators show a continuous rise in proficiency, e.g. the continuous increase in lexical verb + noun collocations, the increase in the number of the overall verbs, adjectives and nouns produced by each level of learners, etc.

Therefore, the investigation into L2 learners' collocational development and the process of how collocations are acquired is far from complete. In future research, more longitudinal studies are needed to identify patterns of phraseological development in L2 learners, and to compare them with those based on quasi-longitudinal data (Paquot and Granger, 2012: 143). In addition, with regard to the pedagogical implications raised in the previous section, experiments can be conducted to test if acquisition of verb semantics as proposed in our study contributes to the learning of collocations in an efficient way in classrooms.

Another area of future enquiry lies in the acquisition of congruent and non-congruent collocations. One of our findings is that learners at lower levels produced significantly more non-congruent collocations than congruent ones. This finding indicates that at early stages of language learning, collocations are learnt as holistic units. However, this claim needs to be further verified through looking at the collocation performance of learners at earlier stages than the ST2 level targeted in our study, in order to get a more comprehensive picture. The recording of learners' written performance at beginners'

and basic levels is scarce and needs to be included, in order to get a fuller picture of L2 learners' production of collocations.

In addition, our finding that non-congruent collocations receive more perceptual salience than congruent collocations in learning and are less susceptible to errors than congruent collocations needs to be further examined in classroom experiments. In this sense, results obtained from spontaneous data in SLA research may be complemented by experimental and intuitional data to capture aspects of competence as well as performance, and to validate results on learners' use of word combinations (cf. Cross and Papp, 2008: 77).

Despite all these limitations discussed above, the findings revealed through a cross-sectional study of the collocations produced by Chinese learners of English contribute to a clearer understanding of the process of second language acquisition, and to a more successful collocation teaching and learning through providing pedagogical implications.

References

Ädel, A. and Erman, B. (2012) Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), pp. 81-92.

Aijmer, K. (2009) “So er I just sort of I dunno I think it’s just because. . .”: A corpus study of “I don’t know” and “dunno” in learner spoken English. In: A. H. Jucker, D. Schreier, and M. Hundt, eds. *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, pp. 151-166.

Aisenstadt, E. (1979) Collocability restrictions in dictionaries. In: R.R.K. Hartmann, ed. *Dictionaries and Their Users. Papers from the 1978 B.A.A.L. Seminar on Lexicography*. Exeter: University of Exeter, pp. 71-74.

Aisenstadt, E. (1981) Restricted Collocations in English Lexicology and Lexicography. *ITL: Review of Applied Linguistics*, 53, pp. 53-61.

Aitchison, J. (2003) *Words in the Mind: An Introduction to the Mental Lexicon*. 3rd edn. Oxford: Blackwell.

Albert, M. and Obler, L.K. (1978) *The Bilingual Brain*. New York: Academic Press.

Algeo, J. (1995) Having a look at the expanded predicate. In: B. Aarts and C.F. Meyer, eds. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, pp. 203-217.

Altenberg, B. (1998) On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In: A.P. Cowie, ed. *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press, pp. 101-122.

Altenberg, B. and Granger, S. (2001) The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2), pp. 173-195.

Altenberg, B. and Granger, S. (2002). Recent trends in cross-linguistic lexical studies. In: B. Altenberg and S. Granger, eds. *Lexis in Contrast: Corpus-Based Approaches.* Amsterdam: Benjamins, pp. 3-48.

Al-Zahrani, M. S. (1998) *Knowledge of English Lexical Collocations Among Male Saudi College Students Majoring in English at a Saudi University*. Ph. D. Thesis. Ann Arbor, MI: UMI.

Agustin Llach, M. P. (2011) *Lexical Errors and Accuracy in Foreign Language Writing*. Bristol: Multilingual Matters.

Bahns, J. (1993) Lexical collocations: a contrastive view. *ELT Journal*, 47(1), pp. 56-63.

Bahns, J. and Eldaw, M. (1993) Should we teach EFL students collocations? *System*, 21(1), pp. 101-114.

Barnbrook, G. (2007) Sinclair on collocation. *International Journal of Corpus Linguistics*, 12(2), pp. 183-199.

Barfield, A. (2007) *An Exploration of Second Language Collocation Knowledge and Development.* Ph. D. Thesis. University of Swansea.

Benson, M. (1985) Collocations and idioms In: R. Ilson, ed. *Dictionaries, Lexicography and Language Learning*. Oxford: Published in association with the British Council by Pergamon, pp. 61-68.

Benson, M., Benson, E., and Ilson, R. (2010) *The BBI Combinatory Dictionary of English: Your Guide to Collocations and Grammar*. 3rd edn. Amsterdam: Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biskup, D. (1990) Some remarks on combinability: Lexical collocations. In: J. Arabski, ed. *Foreign Language Acquisition Papers*. Katowice: Uniwersytet Slaski, pp. 31-44.

Biskup, D. (1992) L1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study. In: P. Arnaud and H. Bejoint, ed. V*ocabulary and Applied Linguistics*. London: Macmillan, pp. 85-93.

Bolinger, D. (1976) Meaning and memory. *Forum Linguisticum*, 1, pp. 1-14.

Bonk, W. (2001) Testing ESL learners' knowledge of collocations. In T. Hudson and J.D. Brown, eds. *A Focus on Language Test Development: Expanding the Language Proficiency Construct across a Variety of Tests*. Technical Report 21. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, pp. 133-142.

Brown, R. (1973) *A First Language: The Early Stages.* Cambridge, MA: Harvard University Press.

Brown, D. (1974) Advanced vocabulary teaching: the problem of collocation. *RELC Journal*, 5(2), pp. 1-11.

Carter, R. and McCarthy, M., eds. (1988) *Vocabulary and Language Teaching*. London: Longman.

Channell, J. (1981) Applying semantic theory to vocabulary teaching. *ELT Journal*, 35(2), pp. 115-122.

Channell, J. (1988) Psycholinguistic considerations in the study of L2 vocabulary acquisition. In: R. Carter and M. McCarthy, eds. *Vocabulary and Language Teaching*. London: Longman, pp. 83-94.

Channell, J. (1994) *Vague Language*. Oxford: Oxford University Press.

Chi Man-lai, A., Wong Piu-yiu, K., and Wong Chau-ping, M. (1994) Collocational problems amongst ESL learners: A corpus-based study. In L. Flowerdew and A.K. Tong, eds. *Entering Text*. Hong Kong: University of Science and Technology, pp. 157-165.

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Church, K., Gale, W., Hanks, P. and Hindle, D. (1991) Using statistics in lexical analysis. In U. Zernik, ed. *Lexical Acquisition: Exploring On-line Resources to Build a Lexicon*. Hillsdale: Erlbaum, pp. 115-164.

Church, K., and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16, pp. 22-29.

Church, K., and Hindle, D. (1990) Collocational constraints and corpus-based linguistics. In *Working Notes of the AAAI Symposium: Text-Based Intelligent Systems*.

Cobb, T. (2003) Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, 59(3), pp. 393-423.

Collins COBUILD English Dictionary. 2nd edn. (1995) London: HarperCollins.

Conklin, K. and Schmitt, N. (2008) Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), pp. 72- 89.

Contemporary Chinese Dictionary, 5th edn. (2005) Beijing: The Commercial Press.

Cowie, A.P. (1981) The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics*, 2(3), pp. 223-235.

Cowie, A.P. (1991) Multiword Units in Newspaper Language. In: S. Granger, ed. *Perspectives on the English Lexicon*: *A Tribute to Jacques Van Roey.* Louvain-la-Neuve: Cahiers de I'Institut de Linguistique de Louvain, pp. 101-116.

Cowie, A.P. (1992) Multiword Lexical Units and Communicative Language Teaching. In: P. Arnaud and H. Bejoint, eds. *Vocabulary and Applied Linguistics*. London: Macmillan, pp.1-12.

Cowie, A.P., ed. (1998) *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press.

Cowie, A.P. (2009) *Semantics*. Oxford: Oxford University Press.

Craik, F.I.M. and Lockhart, R.S. (1972) Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), pp. 671-684.

Cross, J. and Papp, S. (2008) Creativity in the use of verb + noun combinations by Chinese learners of English. In: G. Gilquin, S. Papp and M.B. Diez-Bedmar, eds. *Linking up contrastive and learner corpus research*. Amsterdam: Rodopi, pp. 57-81.

Crossley, S.A. and Salsbury, T. (2011) The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Teaching*, 49, pp. 1–26.

Crystal, D. (1997) *A Dictionary of Linguistics and Phonology*. 4th edn. Oxford: Blackwell.

Davies, A. (2003) *The Native speaker: Myth and Reality* (Bilingual Education and Bilingualism). Clevedon: Multilingual Matters.

Dechert, H.W. (1983) How a story is done in a second language. In: C. Faerch and G. Kasper, eds. *Strategies in Interlanguage Communication*. London: Longman, pp. 175-196.

De Cock, S. (2011) Preferred patterns of use of positive and negative evaluative adjectives in native and learner speech: an ELT perspective. In: A. Frankenberg-Garcia, L. Flowerdew, and G. Aston, eds. *New Trends in Corpora and Language Learning.* London: Continuum, pp. 198-212.

De Cock, S., Granger, S., Leech, G. and McEnery, T., (1998) An automated approach to the phrasicon of EFL learners. In: S. Granger, ed. *Learner English on Computer*. London: Longman, pp. 67-79.

Durrant, P. and Schmitt, N. (2009) To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, 47, pp. 157-177.

Durrant, P. and Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second language research*, 26(2), pp. 163-188.

Ellis, R. (1987) *Second Language Acquisition in Context*. Hertfordshire: Prentice Hall.

Ellis, R. (1994) *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Fan, M. (2009) An exploratory study of collocational use by ESL students - A task based approach. *System*, 37(1), pp. 110-123.

Farghal, M. and Obeidat, H., (1995) Collocations: a neglected variable in EFL. *International Review of*

Applied Linguistics in Language Teaching, 33(4), pp. 315-331.

Fellbaum, C., ed. (1998) *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Fellbaum, C., ed. (2007) *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. London: Continuum.

Fellbaum, C. (2010) Wordnet. In: R. Poli, M. Healy, and A. Kameas, eds. *Theory and Applications of Ontology: Computer Applications*. London: Springer, pp. 231-243.

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

Fitzpatrick, T. (2006) Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook 2006* (6): pp. 121-145.

Foster, P. (2001) Rules and routines: a consideration of their role in the task-based language production of native and non-native speakers. In: M. Bygate, P. Skehan and M. Swain, eds. *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing.* Harlow: Longman, pp. 75-93.

Fox, G. (1998) Using corpus data in the classroom. In: B. Tomlinson, ed. *Materials Development in Language Teaching*. Cambridge: Cambridge University Press, pp. 25-43.

Garside, R. (1987) The CLAWS Word-tagging System. In: R. Garside, G. Leech and G. Sampson, eds. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, pp. 30-41.

Garside, R. (1996) The robust tagging of unrestricted text: the BNC experience. In: J. Thomas and M. Short, eds. *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman, pp. 167-180.

Gass, S.M., and Selinker, L. (2008) *Second Language Acquisition: An Introductory Course*. London: Routledge.

Gitsaki, C. (1999) *Second Language Lexical Acquisition: A study of the Development of Collocational Knowledge*. San Francisco: International Scholars Publications.

Granger, S. (1998a) Prefabricated patterns in advanced EFL writing: collocations and formulae. In: A.P. Cowie, ed. *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press, pp. 145-160.

Granger, S. (1998b) The computer learner corpus: A versatile new source of data for SLA research. In: S. Granger, ed. *Learner English on Computer*. London: Longman, pp. 3-18.

Granger, S. (2002) A bird's eye view of learner corpus research. In: S. Granger, J. Hung, and S. Petch-Tyson, eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, pp. 3-33.

Granger, S. (2014) *John Sinclair's idiom principle: An inspiration for learner corpus research*. Talk given at 2014 Annual John Sinclair Lecture, Birmingham, 8 May 2014.

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M., eds. (2009) *International Corpus of Learner English* (V2). Louvain: Presses Universitaires de Louvain.

Greenbaum, S. (1970) *Verb-intensifier Collocations in English: An Experimental Approach*. The Hague: Mouton.

Greenbaum, S. (1974) Some verb-intensifier collocations in American and British English. *American Speech*, 49 (1,2), pp. 79-89.

Gries, S. (2008) Corpus-based method in analyses of second language acquisition data. In: P. Robinson and N. C. Ellis, eds. *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York: Routledge, pp. 406-431.

Gui, S.C. and Yang, H.Z. (2003) *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

Gyllstad, H. (2005) Words that go together well: Developing test formats for measuring learner knowledge of English collocations. In: F. Heinat and E. Klingval, eds. *The Department of English in Lund: Working Papers in Linguistics*, Vol. 5, pp. 1-31.

Halliday, M.A.K. (1966) Lexis as a linguistic level. In: C.E. Bazell, J.C. Catford, M.A.K. Halliday and R.H. Robins, eds. *In Memory of J. R. Firth*. London: Longman, pp. 148-162.

Hanania, E. and Gradman, H. (1977) Acquisition of English structures: A case study of an adult native speaker of Arabic in an English speaking environment. *Language Learning*, 27(1), pp. 75-91.

Handl, S. (2008) Essential collocations for learners of English: The role of collocational direction and weight. In: F. Meunier, and S. Granger, eds. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: Benjamins, pp. 43-65.

Hanks, P. (1996) Contextual Dependency and Lexical Sets. *International Journal of Corpus Linguistics*, 1(1), pp. 75-98.

Hasselgren, A. (1994) Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4(2), pp. 237-258.

Hausmann, F.J. (1989) Le dictionnaire de collocations. In: F.J. Hausmann, O. Reichmann, H.E. Wiegand, and L. Zgusta, eds. *Wörterbücher: ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin: De Gruyter, pp. 1010-1019.

Henriksen, B. (2013) Research on L2 learners' collocational competence and development – a progress report. In C. Bardel, C. Lindqvist and B. Laufer, eds. *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis.* Eurosla, pp. 29-56. Available through: http://www.eurosla.org/monographs/EM02/EM02tot.pdf [Accessed 10 March 2014].

Herbst, T. (1996) "What are collocations: Sandy beaches or false teeth?" *English Studies*, 77(4), pp. 379-393.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2000) A world beyond collocation: new perspectives on vocabulary teaching. In: M. Lewis, ed. *Teaching Collocation: Further Developments in the Lexical Approach.* Hove: Language Teaching Publications, pp. 224-245.

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoffman, S. and Lehmann, H.M. (2000) Collocational evidence from the British National Corpus. In J.M. Kirk, ed. *Corpora Galore: Analyses and Techniques in Describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998)* Amsterdam: Rodopi, pp. 17-32.

Howarth, P. (1996) *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Tübingen: Niemeyer.

Howarth, P. (1998a) The Phraseology of Learners' Academic Writing. In A.P. Cowie, ed. *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 161-186.

Howarth, P. (1998b) Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1), pp. 24-44.

Hsu, J. (2007) Lexical Collocations and their Relation to the Online Writing of Taiwanese College English Majors and Non-English Majors. *Electronic Journal of Foreign Language Teaching*, 4(2), pp. 192-209.

Hui, Y. (2002) *A New Century Chinese-English Dictionary*. Beijing: Foreign Language Teaching and Research Press.

Hunston, S. and Francis, G. (2000) *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins.

Hyland, K. (2006) "The 'other' English: thoughts on EAP and academic writing". *The European English Messenger*, 15(2), pp. 34-38.

Irujo, S. (1993) Steering clear: avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching*, 31(3), pp. 205-219.

Jackendoff, R. (1983) *Semantics and Cognition.* Massachusetts: MIT Press.

Jackendoff, R. (1997) Twistin' the Night Away. *Language*, (73), pp. 543-559.

James, C. (1996) A Cross-Linguistic Approach to Language Awareness. *Language Awareness*, 5(3-4), pp. 138-148.

Jiang, N. (2000) Lexical representation and development in a second language. *Applied Linguistics*, 21(1), pp. 47-77.

Jiang, N. (2002) Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24, pp. 617-637.

Jiang, N. and Nekrasova, T.M. (2007) The processing of formulaic sequences by second language speakers. *Modern Language Journal*, 91(3), pp. 433-445.

Johansson, S. (2007) *Seeing through Multilingual Corpora: on the Use of Corpora in Contrastive Studies.* Amsterdam: Benjamins.

Johansson, S., and Hofland, K. (1989) *Frequency Analysis of English Vocabulary and Grammar.* Oxford: Oxford University Press.

Jones, S. and Sinclair, J. (1974) English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie*, 23(2), pp. 15-61.

Kamps, J., Marx, M., Mokken, R. J., and De Rijke, M. (2004) Using WordNet to measure semantic orientation of adjectives. *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, Vol. IV, pp. 1115–1118, Lisbon, PT.

Kaszubski, P. (2000) *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: a contrastive, corpus-based approach.* Available through:

http://main.amu.edu.pl/~przemka/rsearch.html [Accessed 13 October 2011].

Katz, J.J. and Fodor, J.A. (1963) The structure of a semantic theory. *Language*, 39(2), pp. 170-210.

Kelly, E.F. and Stone, P.J. (1975) *Computer Recognition of English Word Senses*. Amsterdam: North-Holland Publishing Company.

Kjellmer, G. (1987) Aspects of English Collocations. In: W. Meijs, ed. *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, pp. 133-140.

Kjellmer, G. (1990) Patterns of Collocability. In: J. Arts and W. Meijs, eds. *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi, pp. 163-178.

Kjellmer, G. (1991) A mint of phrases. In: K. Aijmer and B. Altenberg, eds. *English Corpus Linguistics. Studies in Honour of Jan Svartvik.* London: Longman, pp. 111-127.

Kjellmer, G. (1994) *A Dictionary of English Collocations: Based on the Brown Corpus.* Oxford: Clarendon Press.

Klotz, M. (2003) Oxford Collocations Dictionary for Students of English. *International journal of lexicography*, 26(1), pp. 57-61.

Kozlowska, C.D., and Dzierzanowska, H. (1988) *Selected English Collocations*. Warszawa: PWN.

Krashen, S. and Scarcella, R. (1978) On routines and patterns in language acquisition and performance. *Language Learning*, 28(2), pp. 283-300.

Kroll, J.F. and Stewart, E. (1994) Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33, pp. 149-174.

Kroll, J.F., Van Hell, J.G., Tokowicz, N. and Green, D.W. (2010) The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3), pp. 373-381.

Laufer, B. (2006) Comparing focus on form and focus on forms in second language vocabulary learning. *The Canadian Modern Language Review*, 63(1), pp. 149-166.

Laufer, B. and Girsai, N. (2008) Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4), pp. 694-716.

Laufer, B. and Waldman, T. (2011) Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*, 61(2), pp. 647-672.

Lee, C.Y. and Liu, J.S. (2009) Effects of Collocation Information on Learning Lexical Semantics for Near Synonym Distinction. C*omputational Linguistics and Chinese Language Processing*, 14(2), pp. 205-220.

Leech, G. (1974) *Semantics*. Harmondsworth: Penguin.

Levin, B. (1993) *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Lewis, M. (1997) *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M., ed. (2000) *Teaching Collocation: Further Developments in the Lexical Approach*. London: Language Teaching Publications.

Li, J. and Schmitt, N. (2010) The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In: D. Wood, eds. *Perspectives on Formulaic Language: Acquisition and Communication*. London: Continuum, pp. 23–46.

Li, W.Z. (2009) A Critical Review of CIA. *CAFLEC*, 127, pp. 13-17.

LÜ, S.X. (2002) *Collection of LÜ Shuxiang, Vol. 5: 800 Words in Modern Chinese*. Liaoning: Liaoning Education Press.

Lombard, R. J. (1997) *Non-Native Speaker Collocations: A Corpus-Driven Characterization from the Writing of Native Speakers of Mandarin (Mandarin Chinese)*. Ph. D. Thesis. Ann Arbor, MI: UMI.

Lorenz, G. (1999) *Adjective Intensification - Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Louw, B, (1993) Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: M. Backer, G. Francis and E. Tognini-Bonelli, eds. *Text and Technology*. Amsterdam: Benjamins, pp. 157-176.

Martelli, A. (2006) A Corpus Based Description of English Lexical Collocations Used by Italian Advanced Learners. In: E. Corino, C. Marello and C. Onesti, eds. *Proceedings XII EURALEX International Congress*. Alessandria: Edizioni dell'Orso, pp. 1005-1011.

Marton, W. (1977) Foreign vocabulary learning as problem No. 1 of language teaching at the advanced level. *Interlanguage Studies Bulletin*, 2, pp. 33-57.

McEnery, T., and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

McIntosh, C., Francis, B., and Poole, R. (2009) *Oxford Collocations Dictionary for Students of English*. 2nd edn. Oxford: Oxford University Press.

Meara, P. (1978) Learners' word associations in French. *Interlanguage Studies Bulletin*, 3(2), pp. 192-211.

Meara, P. (1982) Word Associations in a Foreign Language. *Nottingham Linguistic Circular*, 11(2), pp. 29-38.

Meara, P. (1984) The study of lexis in interlanguage. In: A. Davies, C. Criper and A.P.R. Howatt, eds. *Interlanguage*. Edinburgh: Edinburgh University Press, pp. 225-235.

Men, H. (2010) A Corpus-based Analysis of Chinese EFL Learners' Adverb/Adjective Collocation Errors in English Writing. *Internet fortune* (4), pp. 108-109.

Men, H. (2014) *L1 Influence on the Production of L2 Collocations: A Corpus-based Study of Chinese EFL Learners' Collocation Acquisition.* Paper given at the 9th Newcastle upon Tyne Postgraduate Conference in Linguistics, 4 April, 2014.

Miller, G.A. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39-41.

Miller, G.A., Beckwith R., Fellbaum, C., Gross, D., and Miller, K. J. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), pp. 235-244.

Milton, J. (2009) *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual matters.

Miyamoto, T. (2000) *The Light Verb Construction in Japanese: The Role of the Verbal Noun*. Amsterdam: Benjamins.

Moon, R. (1998) *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon Press.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary*. Boston, Mass: Heinle & Heinle.

Nattinger, J. 1988. Some current trends in vocabulary teaching. In: R. Carter and M. McCarthy, eds. *Vocabulary and Language Teaching*. London: Longman, pp. 62-80.

Nattinger, J.R. and DeCarrico, J.S. (1992) *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nesi, H. (2011) BAWE: An introduction to a new resource. In: A. Frankenberg-Garcia, L. Flowerdew, and G. Aston, eds. *New Trends in Corpora and Language Learning*. London: Continuum, pp. 213-228.

Nesselhauf, N. (2003) The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), pp. 223–242.

Nesselhauf, N. (2004) What are collocations? In D. Allerton, N. Nesselhauf and P. Skandera, eds. *Phraseological Units: Basic Concepts and Their Application*. Basel: Schwabe, pp. 1-21.

Nesselhauf, N. (2005) *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

Obukadeta, P. (2014) *L2 Collocations: A Problematic Linguistic Phenomenon*? Paper given at the Birmingham English Language Postgraduate Conference, University of Birmingham, 7 March 2014.

Olshtain, E. (1987) The acquisition of new word formation processes in second language acquisition. *Studies in Second Language Acquisition*, 9(2), pp. 221–231.

Oxford Advanced Learner's English-Chinese Dictionary, 7th edn. (2009) Beijing: The Commercial Press.

Palmer, F.R. (1981) *Semantics*. 2nd edn. Cambridge: Cambridge University Press.

Palmer, H.E. (1933) *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Paquot, M., and Granger, S. (2012) Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32, pp. 130-149.

Parastuti, A., Said, M. and Wawan, W. (2009) The negative transfers of English collocations written by the students of Gunadarma University. Available though: http://www.gunadarma.ac.id/library/articles/graduate/letters/2009/Artikel_10604015.pdf [Accessed 11 January 2014].

Partington, A. (1998) *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: Benjamins.

Pawley, A. and Syder, F.H. (1983) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: J.C Richards and R.W. Schmidt, eds. *Language and Communication.* London: Longman, pp.

191-226.

Penke, M., and Rosenbach, A., eds. (2007) *What Counts as Evidence in Linguistics: The Case of Innateness*. Amsterdam: Benjamins.

Peters, A. (1977) Language learning strategies. *Language*, 53(3), pp. 560-573.

Philip, G. (2007) Decomposition and delexicalisation in learners' collocational (mis)behaviour. *Online Proceedings of Corpus Linguistics*, pp. 1-11. Available through: http://ucrel.lancs.ac.uk/publications/cl2007/paper/170_Paper.pdf [Accessed 12 January 2014]

Pu, J.Z. (2010) Corpora and Unified Language Studies. *Journal of PLA University of Foreign Languages*, 33(2), pp. 41-44.

Renouf, A. (1987) Lexical Resolution. In: W. Meijs, ed. *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, pp. 121-131.

Renouf, A. and Sinclair, J. (1991) Collocational frameworks in English. In: K. Aijmer and B. Altenberg, eds. *English Corpus Linguistics. Studies in Honour of Jan Svartvik.* London: Longman, pp. 128-143.

Richards, J.C. and Schmidt R. (2010) *Longman Dictionary of Language Teaching and Applied Linguistics.* 4th edn. London: Longman.

Ringbom, H. (1987) *The Role of the First Language in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, H. (2001) Lexical transfer in L3 production. In: J. Cenoz, B. Hufeisen and U. Jessner, eds. *Cross-linguistic Influence in Third Language Acquisition: Psycholinguistic Perspectives.* Clevedon: Multilingual Matters, pp. 59-68.

Rudzka, B., Channell, J., Putseys, Y. and Ostyn, P. 1981. *The Words You Need*. London: Macmillan.

Rudzka, B., Channell, J., Putseys, Y. and Ostyn, P. 1985. *More Words You Need*. London: Macmillan.

Saint-Dizier, P. and Viegas, E., eds. (1995) *Computational Lexical Semantics.* Cambridge: Cambridge University Press.

Salkie, R. (2002) Two types of translation equivalence. In: B. Altenberg and S. Granger, eds. *Lexis in Contrast. Corpus-Based Approaches*. Amsterdam: Benjamins, pp. 51-71.

Scarcella, R., (1979) Watch up!: a study of verbal routines in adults second language performance. *Working papers on Bilingualism*, 19, pp.79-88.

Schmidt, R.W. (1983) Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson and E. Judd, eds. *Sociolinguistics and Language Acquisition*. Rowley, MA: Newbury House, pp. 137-174.

Schmidt, R.W. (1990) The role of consciousness in second language learning. *Applied Linguistics*, 11(2), pp. 129-150.

Schmitt, N., and Carter, R. (2004) Formulaic sequences in action: An introduction. In: N. Schmitt, ed. *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam: Benjamins, pp. 1-22.

Schmitt, N., Dörnyei, Z., Adolphs, S., and Durow, V. (2004) Knowledge and Acquisition of Formulaic Sequences: A Longitudinal Study. In N. Schmitt, ed. *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: Benjamins, pp. 55-86.

Scott, M. (2004) *WordSmith Tools (Version 4.0)*. Oxford: Oxford University Press.

Sinclair, J. (1966) Beginning the study of lexis. In: C.E. Bazell, J.C. Catford, M.A.K. Halliday and R.H. Robins, eds. *In Memory of J. R. Firth*. London: Longman, pp. 410-430.

Sinclair, J. (1987) Collocation: a progress report. In: R. Steele and T. Threadgold, eds. *Language Topics: Essays in Honour of Michael Halliday*, Vol. 2. Amsterdam: Benjamins, pp. 319-331.

Sinclair, J. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1), pp. 75-106.

Sinclair, J. (1998) The lexical item. In: E. Weigand, ed. *Contrastive Lexical Semantics*. Amsterdam: Benjamins, pp. 1-24.

Sinclair, J. (2003) *Reading Concordances: An Introduction*. London: Pearson Education Limited.

Sinclair, J. (2004) *Trust the Text.* London: Routledge.

Sinclair J, and Fox, G., eds. (1990) *Collins COBUILD English Grammar*. London: Collins.

Sinclair, J., S. Jones, and R. Daley. (2004) *English Collocation Studies: The OSTI Report*. London: Continuum.

Siyanova, A. and Schmitt, N. (2008) L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, 64(3), pp. 429-458.

Spooner, A. (2005) *Oxford Dictionary of Synonyms and Antonyms*. Oxford: Oxford University Press.

Spottl, C. and McCarthy, M. (2004) Comparing knowledge of formulaic sequences across L1, L2, L3 and L4. In: N. Schmitt, ed. *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: Benjamins, pp. 191-225.

Stubbs, M. (1995a) Collocations and Semantic Profiles: on the Cause of the Trouble with Quantitative Study. *Functions of Language*, 2(1), pp. 23-55.

Stubbs, M. (1995b) Corpus evidence for norms of lexical collocation. In G. Cook and B. Seidlhofer, eds. *Principle & Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson.* Oxford: Oxford University Press, pp. 245-256.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2001) *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Thierry, G. and Wu, Y.J. (2007) Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Science of the United States of America*, 104, 12530-12535.

Tufis, D., and Stefanescu, D. (2011) An Osgoodian perspective on WordNet. *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*, pp. 1-8. IEEE.

Van Roey, J. (1990) *French-English Contrastive Lexicology: An Introduction*. Louvain-la-Neuve: Peeters.

Wang, D. (2011) Language Transfer and the Acquisition of English Light Verb + Noun Collocations by Chinese Learners. *Chinese Journal of Applied Linguistics*, 11(2), pp. 107-125.

Wang, Y., and Shaw, P. (2008) Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal*, 32, pp. 201-232.

Wen, Q.F., Liang, M.C, and Yan, X.Q. (2008) *Spoken and Written English Corpus of Chinese Learners*. 2nd edn. Beijing: Foreign Language Teaching and Research Press.

Widdowson, H.G. (1979) *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Widdowson, H.G. (2000) On the Limitations of Linguistics Applied. *Applied Linguistics*, 21(1), pp. 3-25.

Willis, D. (2010) Three reasons why. In: S. Hunston and D. Oakey, eds. *Introducing Applied Linguistics: Concepts and Skills*. London: Routledge, pp. 6-11.

Wolter, B. 2001. Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition*, 23(1), pp. 41-69.

Wolter, B. (2006) Lexical Network Structures and L2 Vocabulary Acquisition: The Role of L1 Lexical/Conceptual Knowledge. *Applied Linguistics*, 27(4), pp. 741-747.

Wolter, B. and Gyllstad, H. (2011) Collocational Links in the L2 Mental Lexicon and the Influence of L1 Intralexical Knowledge. *Applied Linguistics*, 32(4), pp. 430-449.

Wong-Fillmore, L. (1976) *The Second Time Around: Cognitive and Social Strategies in Language Acquisition*. Ph. D. Thesis, Stanford University.

Wray, A. (2000) Formulaic sequences in second language teaching: principles and practice. *Applied Linguistics*, 21(4), pp. 463-489.

Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Xiao, R. and McEnery, T. (2006) Collocation, semantic prosody, and near synonymy: A cross linguistic perspective. *Applied Linguistics*, 27(1), pp. 103-129.

Yamashita, J., and Jiang, N. (2010) L1 Influence on the Acquisition of L2 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *TESOL Quarterly*, 44(4), pp. 647-668.

Yip, V. (1995) *Interlanguage and Learnability: From Chinese to English.* Amsterdam: Benjamins.

Yorio, C.A. (1989) Idiomaticity as an indicator of second language proficiency. In: K. Hyltenstam and L.K. Obler, eds. *Bilingualism across the Lifespan.* Cambridge: Cambridge University Press, pp. 55-72.

Zareva, A. and Wolter, B. (2012) The 'promise' of three methods of word association analysis to L2 lexical research. *Second Language Research*, 28(1), pp. 41-67.

Zhang, W.Z. and Chen, S.C. (2006) EFL learners' acquisition of English Adjective-Noun collocations — A quantitative study. *Foreign Language Teaching and Research*, 38(4), pp. 251-258.

Zhang, X. (1993) *English Collocations and Their Effect on the Writing of Native and Non-native College Fresh-men*. Ph. D. Thesis. Indiana University of Pennsylvania.

Zhang, Y. and Gao, Y. (2006) A CLEC-based Study of Collocation Acquisition by Chinese English Language Learners. *CELEA Journal*, 29(4), pp. 28-35.

Appendix I Erroneous VN collocations produced by the three levels of learners (types)

Learners	DeLexVN	LexVN
ST2	23	41
ST5	17	27
ST6	22	71

(Notes: ST2 and ST5: $p = 0.8404$ ns
ST5 and ST6: $p = 0.1038$ ns
ST2 and ST6: $p = 0.1080$ ns)

Appendix II Well-formed and erroneous VN collocations in the 16 synsets (ST2)

Synsets	Well-formed coll.			Erroneous coll.		
	verbs	nouns	Freq.	verbs	nouns	Freq.
verbs of creation	compose	poem	1	make (compose)	poem	2
	draw	picture	1	create (compose)	poem	2
	draw	conclusion	1	create (compose)	song	2
	hold	party	7	make (create)	environment	1
	hold	meeting	10	give (hold)	meeting	1
	hold	game	1	raise (give)	shout	2
	hold	contest	1			
	hold	festival	1			
	hold	ceremony	2			
	launch	war	1			
	set	fire	2			
	set	example	1			
“fulfil” verbs	discharge	duty	1			
	fulfil	wish	1			
verbs of obtaining	achieve	victory	1	make (achieve)	result	1
	achieve	result	1	earn (acquire)	knowledge	1
	earn	money	1	grasp (acquire)	knowledge	3
	gain	knowledge	1			
	gather	strength	1			
	receive	letter	2			
	receive	degree	3			
verbs of putting	lay	foundation	1			
“settle” verbs	settle	problem	2	do (solve)	problem	4
	solve	problem	2			
“learn” verbs				learn (acquire)	knowledge	16
				get (learn)	lesson	1
				study (acquire)	knowledge	5
				know (acquire)	knowledge	2
verbs of transfer of a message	teach	lesson	1	teach (impart)	knowledge	2
	tell	lie	3	take (tell)	joke	1
	tell	story	3	say (tell)	joke	1
				tell (impart)	knowledge	3
				tell (give)	advice	2
“keep” verbs	hold	breath	1			
	keep	record	7			
	keep	pace	1			
	keep	balance	1			
	keep	secret	2			
	keep	promise	1			

“follow” verbs	obey	rule	2	obey (face)	fact	1
	follow	advice	1	observe (obey)	law	1
“play” verbs	play	part	1	play (perform)	play	3
“change” verbs	change	mind	6			
“break” verbs	break	rule	2			
	break	record	3			
“live” verbs	lead	life	3			
	live	life	4			
“wear” verbs	wear	clothes	17	dress (wear)	clothing	1
“drive” verbs	drive	motorcycle	1	ride (drive)	bus	1
	drive	bus	1			
	drive	car	1			
	ride	bike	5			
“pay” verbs	devote	attention	1			
	pay	visit	2			
	pay	respect	1			
	pay	attention	23			

(Note: verbs in brackets are the target verbs for erroneous VN collocations.)

Appendix III Well-formed and erroneous VN collocations in the 16 synsets (ST6)

Synsets	Well-formed coll.			Erroneous coll.		
	verbs	nouns	Freq.	verbs	nouns	Freq.
verbs of creation	arouse	concern	2	arise (arouse)	discussion	1
	chart	course	1	arouse (cause)	trouble	1
	draft	law	1	build (enact)	regulation	1
	draw	conclusion	11	build (establish)	tie	1
	establish	relationship	1	draw (draft)	law	2
	form	habit	1	make (draw)	conclusion	1
	hold	meeting	2	draw (formulate)	theory	2
	hold	conference	3	draw (draft)	treaty	1
	launch	war	1	put forth (enact)	law	2
	raise	consciousness	1	put forward (enact)	law	1
	raise	objection	1	set (enact)	law	1
	raise	question	3	publish (enact)	law	2
	raise	issue	4	take (launch)	career	2
	raise	argument	1	stir (raise)	consciousness	1
	raise	alarm	1	raise (arouse)	discussion	1
	set	goal	1			
	set	example	4			
	set	fire	1			
	enact	law	1			
"fulfil" verbs	apply	principle	1	accomplish (commit)	crime	2
	enforce	policy	1	carry out (realise)	value	1
	enforce	law	2	ensure (enforce)	law	1
	exercise	power	2	carry on (enforce)	law	1
	exercise	judgment	1	exert (demonstrate)	ability	1
	exercise	right	1	exert (demonstrate)	competence	1
	exert	influence	3	fulfil (demonstrate)	ability	1
	fulfil	role	1	implement (perform)	act	1
	fulfil	ambition	1	attend (perform)	military service	1
	fulfil	wish	1	take (perform)	military service	1
	implement	principle	1	carry (perform)	function	1
	implement	law	1	make (conduct)	exam	1
	implement	policy	1	take (conduct)	survey	2
	perform	military service	2	conduct (commit)	murder	2
	perform	act	1	conduct (commit)	crime	3
	perform	function	1	make (commit)	crime	1

	realise	value	2	do (commit)	crime	4
	realise	dream	6			
	realise	goal	1			
	conduct	survey	2			
	commit	crime	120			
	commit	homicide	2			
	commit	suicide	17			
	commit	offence	1			
	commit	murder	2			
	commit	act	1			
verbs of obtaining	achieve	aim	1	receive (achieve)	success	1
	achieve	dream	1	cause (catch)	attention	1
	achieve	goal	11	reach (catch)	attention	1
	achieve	purpose	2	meet (earn)	praise	1
	achieve	success	3	get (reach)	conclusion	1
	catch	attention	1	approach (reach)	conclusion	1
	earn	money	27	reach (receive)	recognition	1
	earn	living	7	receive (undergo)	operation	1
	gain	knowledge	4			
	gain	victory	1			
	grasp	opportunity	2			
	reach	agreement	1			
	reach	goal	3			
	reach	target	2			
	reach	conclusion	1			
	receive	award	1			
	receive	training	3			
	receive	education	24			
	receive	treatment	3			
	receive	attention	1			
	receive	reward	2			
	receive	punishment	5			
	receive	warning	1			
	seize	opportunity	2			
verbs of putting	attach	importance	8	give (impose)	burden	1
	fix	eye	2	do (impose)	punishment	1
	impose	fine	1	impose (pose)	threat	1
	impose	burden	1	lay (impose)	burden	1
	impose	punishment	2	lay (cast)	eye	1
	lay	emphasis	2	lay (assign)	role	1
	lay	foundation	1	give (put)	end	1
	place	emphasis	1	put (pay)	attention	3
	put	value	5			
	put	emphasis	3			

	put	hope	1			
	put	end	19			
	put	blame	2			
	put	priority	1			
“settle” verbs	solve	problem	35	charge (tackle)	problem	1
	solve	dispute	2			
	resolve	problem	3			
	tackle	problem	1			
	undertake	duty	1			
	undertake	task	2			
“learn” verbs	acquire	knowledge	4	learn (acquire)	knowledge	11
				have (learn)	lesson	1
				get (learn)	lesson	1
				master (acquire)	knowledge	2
				study (acquire)	knowledge	3
verbs of transfer of a message	teach	lesson	1	teach (impart)	knowledge	2
	tell	truth	1	instruct (communicate)	idea	1
	tell	story	5	push (impart)	knowledge	1
	impart	knowledge	1			
“keep” verbs	hold	opinion	3	reflect (hold)	prejudice	1
	hold	position	3	cast (hold)	prejudice	1
	hold	belief	2			
	hold	post	1			
	hold	view	5			
	hold	attitude	3			
	keep	watch	1			
	keep	distance	2			
	keep	eye	3			
	keep	balance	6			
	keep	promise	2			
	keep	pace	1			
	maintain	order	3			
	maintain	balance	1			
	maintain	dignity	1			
“follow” verbs	obey	law	7	obey (adhere to)	principle	1
	obey	rule	2			
	follow	principle	1			
	follow	rule	1			
	adopt	attitude	4			
	adopt	method	2			
	adopt	policy	3			
	adopt	law	1			
“play”	play	role	54	serve (play)	role	1

verbs	play	part	3	act (play)	role	1
				lead (play)	role	1
“change” verbs	shift	focus	2	change (rehabilitate)	criminal	1
	change	mind	1			
“break” verbs	break	law	26	break (violate)	regulation	1
	break	rule	2			
	break	promise	1			
	violate	regulation	1			
	violate	law	6			
“live” verbs	lead	life	37			
	live	life	30			
“wear” verbs	wear	clothes	1	dress (wear)	clothing	1
“drive” verbs	drive	car	2			
“pay” verbs	pay	attention	53	pay (give)	praise	1
	pay	heed	25			
	pay	respect	3			

(Note: verbs in brackets are the target verbs for erroneous VN collocations.)

Appendix IV Frequencies of well-formed and erroneous VN collocation types in the 16 synsets (ST2 and ST6)

Types	Synsets	ST2		ST6	
		WFC	EC	WFC	EC
1	verbs of creation	12	6	19	15
2	“fulfill” verbs	2	0	26	17
3	verbs of obtaining	7	3	24	8
4	verbs of putting	1	0	14	8
5	“settle” verbs	2	1	6	1
6	“learn” verbs	0	4	1	5
7	verbs of transfer of a message	3	5	4	3
8	“keep” verbs	6	0	15	2
9	“follow” verbs	2	2	8	1
10	“play” verbs	1	1	2	3
11	“change” verbs	1	0	2	1
12	“break” verbs	2	0	5	1
13	“live” verbs	2	0	2	0
14	“wear” verbs	1	1	1	1
15	“drive” verbs	4	1	1	0
16	“pay” verbs	4	0	3	1

(Notes: WFC stands for well-formed VN collocations; EC for erroneous VN collocations.)

Appendix V Well-formed and erroneous VN collocations in the 16 synsets (ST5)

Synsets	Well-formed coll.			Erroneous coll.		
	verbs	nouns	Freq.	verbs	nouns	Freq.
verbs of creation	arouse	admiration	1	raise (arouse)	discussion	1
	build	building	4	hold (stage)	race	3
	conduct	experiment	2	hold (stage)	match	2
	draw	conclusion	6	do (enact)	law	2
	draw	picture	1	do (enact)	regulation	2
	establish	relationship	2	play (perform)	dance	1
	form	habit	7			
	hold	party	6			
	hold	meeting	2			
	hold	debate	1			
	launch	campaign	3			
	perform	play	1			
	produce	effect	2			
	publish	book	7			
	raise	question	3			
	set	fire	1			
“fulfil” verbs	apply	principle	2	fulfil (implement)	plan	1
	enforce	law	1	practice (implement)	policy	1
	perform	operation	1			
verbs of obtaining	achieve	purpose	3	catch (seize)	chance	3
	achieve	aim	3	catch (seize)	opportunity	1
	achieve	success	2	grasp (acquire)	skill	1
	earn	money	9			
	earn	salary	2			
	earn	living	2			
	gain	knowledge	6			
	gain	independence	2			
	reach	agreement	2			
	reach	goal	3			
	receive	letter	42			
	seize	opportunity	1			
verbs of putting	attach	importance	4	put (turn)	ear	1
	lay	stress	1			
	place	emphasis	2			
	put	emphasis	2			
	put	stress	1			
	put	end	2			
	set	foot	1			

“settle” verbs	resolve	problem	1	do (solve)	problem	5
	solve	problem	45			
“learn” verbs	learn	lesson	3	learn (acquire)	knowledge	25
	master	skill	3	study (acquire)	knowledge	6
verbs of transfer of a message	teach	lesson	2	teach (impart)	knowledge	7
	tell	story	14	have (tell)	joke	1
	tell	lie	4			
	tell	truth	2			
	tell	joke	3			
“keep” verbs	hold	opinion	2			
	keep	touch	1			
“follow” verbs	adopt	method	1	obey (adopt)	method	2
	adopt	policy	1			
	follow	instruction	2			
	obey	law	1			
	obey	rule	2			
“play” verbs	play	role	23	act (play)	role	2
	play	part	4	occupy (play)	role	1
				do (play)	role	1
				lay (play)	role	1
“change” verbs	change	mind	2			
“break” verbs	break	law	1			
	break	rule	1			
	break	record	1			
	violate	rule	2			
“live” verbs	live	life	17	make (live)	life	1
	lead	life	13			
“wear” verbs	wear	clothes	11			
“drive” verbs	ride	bike	16			
“pay” verbs	pay	attention	25			
	pay	respect	1			

(Note: verbs in brackets are the target verbs for erroneous VN collocations.)

Appendix VI Frequencies of well-formed and erroneous VN collocation types in the 16 synsets (ST2, ST5 and ST6)

Types	Synsets	ST2		ST5		ST6	
		WFC	EC	WFC	EC	WFC	EC
1	verbs of creation	12	6	16	6	19	15
2	“fulfil” verbs	2	0	3	2	26	17
3	verbs of obtaining	7	3	12	3	24	8
4	verbs of putting	1	0	7	1	14	8
5	“settle” verbs	2	1	2	1	6	1
6	“learn” verbs	0	4	2	2	1	5
7	verbs of transfer of a message	3	5	5	2	4	3
8	“keep” verbs	6	0	2	0	15	2
9	“follow” verbs	2	2	5	1	8	1
10	“play” verbs	1	1	2	4	2	3
11	“change” verbs	1	0	1	0	2	1
12	“break” verbs	2	0	4	0	5	1
13	“live” verbs	2	0	2	1	2	0
14	“wear” verbs	1	1	1	0	1	1
15	“drive” verbs	4	1	1	0	1	0
16	“pay” verbs	4	0	2	0	3	1

Appendix VII Adjectives categories in the ST2 and ST6 AN collocation databases

	ST6	ST2
Qualitative adjectives	(69): active, adverse, bad, breaking, bright, broad, clean, clear, close, common, controversial, convincing, dark, deadly, deaf, deep, dense, distant, effective, fair, fatal, fertile, fierce, fresh, full, good, great, guilty, hard, heated, heavy, high, hot, infectious, irresistible, keen, key, leading, lethal, light, long, mass, narrow, near, nice, polluted, practical, primary, primitive, privileged, professional, promising, rapid, remote, rural, scientific, sharp, small, solid, sore, strong, torrential, unexpected, urban, urgent, vicious, warm, weak, wide	(34): active, bad, bright, cheap, classical, close, common, correct, crisp, dark, deep, fair, fast, firm, foul, fresh, full, glib, good, great, happy, hard, heavy, high, long, loose, loud, low, open, popular, rapid, soft, strong, warm
Classifying adjectives	(70): academic, annual, arable, armed, associate, atomic, biochemical, bodily, boiling, broken, capitalist, chemical, compulsory, consequential, corporal, criminal, cultural, curable, daily, developed, developing, domestic, economic, electric, endangered, environmental, ethical, everyday, feminist, financial, final, five-star, flared, foreign, human, illegal, incurable, individual, industrial, initial, international, juvenile, latest, liberal, literal, living, medical, middle, military, monetary, moral, naked, native, national, natural, nuclear, personal, physical, plastic, political, presidential, promissory, public, racial, sexual, social, solar, spoiled, territorial, top	(25): boiled, British, botanical, capitalist, civil war, closing, criminal, daily, developed, developing, double, everyday, extracurricular, final, foster, founding, historic, living, Lunar, military, natural, physical, political, public
Emphasising adjectives	(1): absolute	(1): blue
Colour adjectives	(2): black, blue	(0)

Appendix VIII Well-formed and erroneous congruent and non-congruent collocations in the ST6 (types)

Types	Congruent coll.	Non-congruent coll.	Total
Well-formed coll.	221 (75%)	127 (87%)	348
Erroneous coll.	74 (25%)	19 (13%)	93
Total	295 (100%)	146 (100%)	441

(Note: $\chi^2 = 7.84$, $p = 0.0051$ **)

Appendix IX Well-formed congruent and non-congruent VN collocations in the ST2 and ST6 (types)

Types	ST2	ST6	Total
Congruent coll.	117 (53%)	221 (64%)	336
Non-congruent coll.	104 (47%)	127 (36%)	231
Total	221 (100%)	348 (100%)	569

(Note: $\chi^2 = 6.42$, $0.01<p<0.05$ *)

Appendix X Erroneous congruent and non-congruent VN collocations in the ST2 and ST6 (types)

Types	ST2	ST6	Total
Congruent coll.	51 (80%)	74 (80%)	125
Non-congruent coll.	13 (20%)	19 (20%)	32
Total	64 (100%)	93 (100%)	157

(Note: $\chi^2 = 0.00$, $p > 0.05$ ns)

Appendix XI Positive and negative transfer between VN and AN collocations in the ST2 (types)

Types	VN coll.	AN coll.	Total
Positive transfer (C, W)	117 (75%)	80 (92%)	197
Negative transfer(I, N and C, N)	40 (25%)	7 (8%)	47
Total	157 (100%)	87 (100%)	244

(Note: $p = 0.0007$ ***)

Appendix XII Positive and negative transfer between VN and AN collocations in the ST6 (types)

Types	VN coll.	AN coll.	Total
Positive transfer (C, W)	221 (89%)	204 (99%)	425
Negative transfer(I, N and C, N)	28 (11%)	2 (1%)	30
Total	249 (100%)	206 (100%)	455

(Note: $p < 0.0001$ ***)