

**A Corpus-driven study of Chinese translators' use of
English collocations in commercial
Chinese to English translation**

Haoda Feng

**A thesis submitted to
Auckland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)**

2014

School of Language and Culture

Co-supervisors: Ineke Crezee and Lynn Grant

Contents

List of Tables	vi
List of Figures	viii
Attestation of authorship	x
Acknowledgements	xi
Abstract	xii
Chapter One Introduction	1
1.1 Research background	1
1.2 Research objectives	4
1.3 Translator training programmes in China	7
1.4 Rationale for the research	9
1.5 Questions addressed by this research	10
1.6 Chapter overview	12
Chapter Two Literature review: Collocation and Translation Studies	13
2.1 Introduction	13
2.2 Researching the nature of collocations with a corpus approach	13
2.2.1 Collocations as formal co-occurrences	14
2.2.2 Collocations as extended units of meaning	22
2.2.3 Collocations as form-function composites	26
2.3 Role of collocation	30
2.3.1 Collocations facilitate language development	30
2.3.2 Collocations help achieve native-like language proficiency	33
2.4 Corpus approaches for Translation Studies	37
2.4.1 Translation universals	39
2.4.2 Collocations' association and translation universals	46
2.5 Summary	50
Chapter Three The preliminary study: Setting the stage	52
3.1 Introduction	52
3.2 Rationales for researching collocations in a learner corpus	52
3.2.1 A model of L1 collocation learning	54

3.2.2 L2 collocation learning.....	57
3.3 Distinction between translators and L2 learners	64
3.3.1 PACTE Group's model.....	66
3.3.2 Pym's model.....	68
3.3.3 Translators in this study.....	71
3.4 Theoretical framework of the present study	72
3.4.1 Operational definition of collocation in the present study	73
3.4.2 Mapping a theoretical framework of collocation in translation.....	74
3.5 Summary	82
Chapter Four Research design.....	83
4.1 Introduction.....	83
4.2 Research methodology and research method	83
4.2.1 Research methodology: the corpus-driven approach.....	83
4.2.2 Research method: the Contrastive Interlanguage Analysis	86
4.3 Corpora employed in this study: the NECCD and the TECCTC	92
4.3.1 The Native English Corpus of Commercial Discourse	92
4.3.2 The Translational English Corpus of Commercial Translation from Chinese.....	93
4.3.3 General information on the corpora	94
4.4 Data processing	94
4.4.1 The Bigram Model in collocation retrieval	94
4.4.2 Statistical measures	96
4.4.3 Software tools for data retrieval	100
4.5 Filtering devices	106
4.5.1 Frequency filtering device.....	107
4.5.2 Form filtering device.....	108
4.5.3 Semantic filtering device.....	109
4.6 Collocation retrieval results	109
4.7 Summary	110
Chapter Five Data analysis: Features of Chinese translators' use of English collocations in the commercial register (Part I)	111
5.1 Introduction.....	111

5.2 Features of collocation density and collocation distribution regarding overall frequency..	111
5.3 Features of collocation distribution regarding statistical values	117
5.4 Lexical analysis regarding deviation of collocation use in translational English.....	123
5.5 Summary	130
Chapter Six Data analysis: Features of Chinese translators' use of English collocations in the commercial register (Part II).....	131
6.1 Introduction	131
6.2 Formal features of collocation use in the corpus of translational English.....	131
6.3 Semantic features of collocation use in the translational corpus.....	139
6.4 Functional features of collocation use in the translational corpus.....	147
6.5 Summary	156
Chapter Seven Translation universals in Chinese translators' use of L2 English collocations	158
7.1 Introduction	158
7.2 A model of the control mechanism between features of collocations and translation universals.....	158
7.2.1 Simplification	160
7.2.2 Explicitation	165
7.2.3 Normalisation	172
7.3 Factors that may be responsible for the deviation in Chinese translators' production of L2 English collocations	178
7.4 Summary	184
Chapter Eight Implications of findings	186
8.1 Introduction	186
8.2 Theoretical implications.....	186
8.2.1 Clarifying the role of collocation in translation	187
8.2.2 Re-evaluation of theoretical models.....	191
8.2.3 Providing evidence for the hypothesis of translation universals	194
8.3 Practical implications	195
8.4 Pedagogical implications.....	197
8.5 Summary	204
Chapter Nine Conclusion	205

9.1 Introduction	205
9.2 Summary of major findings.....	205
9.2.1 Overall frequencies	206
9.2.2 Frequency and statistical values	206
9.2.3 MI score of high-frequency collocations.....	206
9.2.4 Formal analysis	207
9.2.5 Semantic analysis	208
9.2.6 Functional analysis	208
9.2.7 Contrastive Interlanguage Analysis.....	209
9.3 Limitations of the present study	210
9.4 Directions for future research.....	212
References	215
Glossary	225
List of appendices	226
Appendix A: Assessment of Translators in China	227
Appendix B: Sample Texts from the Corpora	230
Appendix C: The <i>FoxPro</i> Programme for Retrieving Bigrams (exemplified with the TECCTC)...	233
Appendix D: Top 30 Most Frequently Used Collocations from the Two Corpora.....	234
Appendix E: Source codes for the <i>Perl</i> programme of text chunk segmentation and computation of vocabulary growth	236
Appendix F: Source codes for the <i>Perl</i> programme of lemmatisation	241
Appendix G: Source codes for the <i>Perl</i> programme of computing keyword growth.....	243

List of Tables

Table 2.1	Key word <i>accuse</i> in a five-word span in the random sampling from the British National Corpus	18
Table 2.2	Translation universals and the linguistic indicators in the present collocation study	49
Table 2.3	Translation universals and computational operators in the present collocation study	49
Table 4.1	General information of the corpora in the study	92
Table 4.2	Bigram information in the TECCTC and the NECCD	101
Table 4.3	Keyword information in the TECCTC and the NECCD	102
Table 4.4	General collocation information across the NECCD and the TECCTC	107
Table 5.1	Comparison of collocation use across the NECCD and the TECCTC	111
Table 5.2	T-test regarding the comparison of collocation use between the NECCD and the TECCTC	112
Table 5.3	Top 10 overused collocations in the TECCTC	115
Table 5.4	Top 10 collocations with highest MI scores across the two corpora	117
Table 5.5	Top 10 collocations with lowest MI scores across the two corpora	117
Table 5.6	Partial results of keyword growth data across the two corpora	128
Table 6.1	General information of collocations regarding form	131
Table 6.2	Chi-square tests for comparing tokens of free combinations, bound collocations and idioms between the two corpora	132
Table 6.3	Chi-square tests for comparing types of free combinations, bound collocations and idioms between the two corpora	135
Table 6.4	General information of collocations with regard to meaning	140
Table 6.5	Chi-square test results for comparing collocation tokens with regard to semantic features	142
Table 6.6	Chi-square test results for comparing collocation types with regard to semantic features	144

Table 6.7	General information of collocations regarding semantic prosody	148
Table 6.8	Chi-square test results for comparing collocation tokens with regard to semantic prosodies	150
Table 6.9	Chi-square test results for comparing collocation types with regard to semantic prosodies	152
Table 7.1	Frequency profiles of the NECCD and the TECCTC	160
Table 7.2	Meanings realised by <i>call</i>	162
Table 7.3	Concordance lines of <i>stack up</i> in the NECCD	166
Table 7.4	Top 20 most frequently used collocations with neutral semantic prosodies in the two corpora	173

List of Figures

Figure 2.1	Interface of pragmatic competence and linguistic competence in collocations (Nattinger & DeCarrico, 1992, P. 16)	27
Figure 2.2	Relative proportions of holistic and analytical involvement in language processing from birth to adulthood (schematic representation) (Wray & Perkins, 2000, p. 20)	31
Figure 2.3	Relationship between translation universals	45
Figure 3.1	The model of Working Memory for language acquisition (Ellis, 2001, p. 36)	55
Figure 3.2	L2 model of collocation output	61
Figure 3.3	Model of Translation competence (Reprinted from PACTE, 2003, p. 60)	67
Figure 3.4	Collocation in translation process	75
Figure 4.1	Integrated Contrastive Model (Gilquin, 2008, p. 8)	87
Figure 4.2	The CIA model in this study	89
Figure 4.3	Statistical results for <i>deficit</i> and its collocates in the NECCD	102
Figure 4.4	Collocation retrieval procedures	105
Figure 5.1	Distribution of collocation tokens in the NECCD and the TECCTC	114
Figure 5.2	Distribution of collocation types in the NECCD and the TECCTC	114
Figure 5.3	Distribution of collocation tokens regarding MI score	118
Figure 5.4	Distribution of collocation types regarding MI score	119
Figure 5.5	Comparison of type token ratios regarding MI score grouping	120
Figure 5.6	Keyword growth in the NECCD	126
Figure 5.7	Keyword growth in the TECCTC	127
Figure 6.1	Comparison of collocation tokens regarding formal classification	132
Figure 6.2	Comparison of collocation types regarding formal classification	135
Figure 6.3	Comparison of collocation tokens regarding semantic features	142
Figure 6.4	Comparison of collocation types regarding semantic features	143

Figure 6.5	Comparison of collocation tokens regarding semantic prosodies	149
Figure 6.6	Comparison of collocation types regarding semantic prosodies	151
Figure 6.7	Correspondence analysis regarding semantic prosodies between the two corpora	153
Figure 7.1	A model of the control mechanism between features of collocations and TUs	157
Figure 8.1	Relative proportions of L2 learners' holistic and analytical involvement in language processing from beginning level to advanced level	186
Figure 8.2	Examples of collocation pairs in MultiTerm	195
Figure 8.3	'Four E's' strategies in L2 collocation learning and acquisition	166
Figure 8.4	Knowledge map of <i>economy</i> and its collocates in the NECCD	202

Attestation of authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institution of higher learning, except where due acknowledgement is made in the acknowledgements.

Haoda Feng*August, 2014*

Acknowledgements

I owe several debts of gratitude to my supervisors, Dr Ineke Crezee and Dr Lynn Grant, who have been shepherding me in the exploration of Translation Studies and linguistic theories, and they have provided invaluable comments on my present research in the past four years. Their rigorous academic supervision, warm-hearted research support, and their inexhaustible interest, perspicuity and perseverance in corpus linguistics and Translation Studies have always been a great encouragement to me. My individual research conducted in this thesis also benefited from the discussions with them, without whose expertise in this research area, this thesis would have been far weaker. Thanks are also due to Professor Fengxiang Fan, who offered practical and sound academic advice on writing computer programmes for data processing and assessing data accuracy. I also owe a special debt to my friend, Alan Feng Shi, for his criticisms and suggestions on the present study.

I owe sincere thanks to my parents, Kelun Feng and Yuhua Han, who have been supporting me in various ways through out my postgraduate career. In particular, I would like to thank my beloved father, who is my first foreign language teacher and the most important person in my life. I also owe thanks to my wife Ivy Liu, my son Justin Feng and my daughter Charlotte Feng, who provide me with a harmonious environment for academic research.

Finally, thanks are due to the administration staff in the Faculty of Language and Culture for their excellent services and helpful advice. Thanks are also due to Auckland University of Technology for funding this research project through a three-year Vice-Chancellor Doctoral Scholarship, whose support, scrutiny and meticulous attention to detail have been indispensable at all the stages of writing this thesis.

Abstract

The issue of differences between translational language and native-speaker language has become a topic of increasing interest in linguistics and Translation Studies (TS). One of the primary tasks in this research area is to employ a corpus approach and analyse collocations with authentic language data by comparing comparable corpora consisting of translated and native-speaker texts. Collocation in linguistics and TS refers to the relationship of co-occurrence between lexical items. The present study shows that examining the use of collocations plays a very significant part in assessing the naturalness of second language (L2) use, and therefore can be a valid measure to make a distinction between translational language and native-speaker language. Nevertheless, the role of collocation has not been given enough attention or discussed systematically in TS and, to date, no translation theorist has clarified the mechanism of collocation in TS, by which translators acquire receptive and productive knowledge of collocations in their L2. In addition, previous research in this area is largely confined to Indo-European languages, resulting in a lack of empirical evidence involving Asian languages.

This thesis concentrates on the nature of collocation and explores collocation distribution patterns by comparing native-speaker English and English translated from Chinese in the commercial register. It focused on five main research questions. Firstly, it attempts to propose a conceptual framework of collocation to explain the nature of collocation in language operations. Based on relevant literature review, this study shows that collocation studies can be carried out on a multi-dimensional basis from the formal, semantic and functional perspectives. Therefore, the comparison between Chinese translators and native English speakers in terms of collocation use is carried out from these perspectives.

Secondly, this thesis attempts to describe the role of collocation in translation and demonstrate the key factors that might influence translators' use of English (L2) collocations. This study proposes a theoretical framework which clarifies the role of

collocation in influencing the relationship between translation universals and the native-like rendition of the target texts. It also discusses how translators' implicit knowledge and explicit knowledge of collocations may influence the use of their L2. The results show that collocation appears to be very important and directly determines the natural use of language, because Chinese translators' inappropriate use of L2 English collocations has made them introduce some translation universals into the target texts and produce foreign-sounding-ness in their L2. This also indicates that Chinese translators in the current research may not have reached the stage of implicit knowledge in terms of using L2 English collocations.

Thirdly, this thesis designed two comparable corpora and attempts to propose a method of retrieving collocations. The English texts in both corpora are first segmented into two groups of bigrams respectively, and then are 'screened' to provide collocation pairs with statistical measures. Finally, these collocation pairs are 'refined' according to the three filtering criteria established in this study to obtain qualified collocations in the commercial register.

Fourthly, this thesis employs the Contrastive Interlanguage Analysis approach and attempts to investigate the features of the variation, alternatively the deviation in collocation distribution patterns, in Chinese translators' use of L2 English collocations in L1-to-L2 translations. The results show that, when compared with native commercial English, Chinese translators' translation outputs depend heavily on the repeated use of high-frequency collocations. This is evidenced from the comparatively lower type-token ratio and the slower keyword growth rate in the corpus of translational commercial English. In addition, this study demonstrates that Chinese translators' translation outputs tend to have more free combinations but fewer bound collocations and idioms; more collocations with literal senses but fewer collocations with delexicalized senses; more collocations with neutral semantic prosodies but fewer collocations with positive or negative semantic prosodies.

Finally, this thesis attempts to briefly offer constructive suggestions to translators who

are L2 users of English, based on the findings from statistical and explanatory analyses. It is suggested in this study that the ‘real-life’ language-learning strategy, or situated learning, would appear to be a useful method for helping translators to identify L1-L2 differences and overcome their shortcoming in using L2 collocations in L1-to-L2 translations.

Chapter One Introduction

1.1 Research background

The issue of differences between translational language and native-speaker language has become a topic of increasing interest in linguistics and Translation Studies (TS). Many researchers have conducted relevant research to identify the features of translational language and have suggested solutions to the existing challenges and difficulties that translators are confronted with. As Xiao (2010) notes, “[t]he distinctive features of translational language can be identified by comparing translations with comparable native texts, thus throwing new light on the translation process and helping to uncover translation norms...” (p. 8). In addition, Baker (2004) also points out that “... questions relating to how one selects the features to be compared and, more importantly, how the findings may be interpreted, invite us to elaborate our methodology far more explicitly than in other types of research” (p. 167). Such viewpoints appear to indicate that the investigation into the features of translational language cannot be comprehensive but can only be carried out through employing particular measures. Only in such a way can researchers in this area specify from what aspect they would need to compare translational language and native-speaker language, and clarify how they would effectively outline a suitable methodology and construct a theoretical model to conduct the comparison. In this respect, the use of accurate collocations plays a very significant part in identifying non-native speakers’ second language competence (see for instance Wray, 2002), which implies that collocation can be a valid measure to make a distinction between translational language and native-speaker language. Previous studies (e.g. Baker, 2004; Xiao, 2010) have demonstrated some theoretical frameworks to distinguish translational language from native-speaker language which, to a certain extent, contribute to both the studies of translational language and the studies of collocation. Particularly, along with the advent of large-scale corpora, researchers are provided with more opportunities to explore collocation patterns in translational language by examining authentic language materials. The findings of these studies

appear to reflect a similar generality, that is, L2 learners deviate from native speakers in the way they learn and use collocations. In other words, this type of research aims to uncover how receptive linguistic skills can be turned into productive linguistic skills, resulting in native-like collocation use in translations. This is also the rationale underpinning the present study.

Nevertheless, the role of collocation has not been given enough attention or discussed systematically in Translation Studies (TS). To be more specific, most previous research of collocation merely focuses on L2 acquisition, rather than comparing translational language and native-speaker language to identify the role of collocation in translation. In addition, no studies have, up to date, shown that the inappropriate use of collocations may cause some translation universal features in translational language, or have discussed the relationship between collocation and the indicators of these translation universal features. From the angle of collocation learning, there are no studies documented in the literature which attempt to suggest a pedagogical model with regard to how translators should effectively learn collocations in translator training and use them appropriately in translation practice. Instead, most researchers in this area have to rely on some theoretical models in linguistics regarding second language (L2) collocation learning if they are to explain translators' use of collocations in their research. Needless to say, there are at times discrepancies in the mechanism of L2 collocation learning. For instance, Ellis (2001, 2003) believes that L2 collocation learning can follow the native speakers' (L1) model, largely relying on the memory system, and the chunking in formulaic sequences is the main factor developing the language acquisition process. In contrast, Wray (2002) claims that L2 learners do not follow the same strategy as L1 learners and L2 learners basically adopt a 'non-formulaic' approach. On account of this discrepancy, Durrant and Schmitt (2010) carried out a lab-based study based on three different training conditions of encountering L2 adjective-noun collocations, specifically, single exposure, verbatim repetition and varied repetition. The results from their study demonstrate that L2 learners do retain information about co-occurring words to which they are exposed. Nevertheless, this model is constructed mainly on an ideal learning environment. That is

to say, this model fails to present some other factors that may interfere with L2 collocation learning, such as L1 transfer. For this point, Crezee and Grant (2013) argue that L2 learners need to, first of all, recognise such idiomatic phrases for what they are but such collocations often ‘fly under the radar’. Therefore, no conscious recognition means no conscious acquisition. In other words, their statement largely indicates that that even though L2 learners hear the collocations, which also occur as ‘chunks’ or ‘strings’ of language, they do not really ‘hear’ them or recognise them. This situation can also be explained in terms of L2 learners’ pre-existing knowledge. When L2 learners tend to produce new collocations that they have never come across, their pre-existing knowledge will serve as a screening device and select collocational candidates more from their L1 than from their L2. Then those candidates will be combined according to the conceptual association to form so called ‘collocations’ which may, or may not, be acceptable in their L2. In this sense, this pattern largely deviates from the memory-based chunking mechanism. This is the reason why Durrant and Schmitt (2010) limit their suggestions merely to the investigation of the “words that they [L2 learners] are already assumed to know” (p. 181). This is also the reason why they did not talk too much in their model about the learning pattern of unknown collocations, but simply mentioned that “[i]t is possible that somewhat different processes will be involved for collocations of previously unknown words” (Durrant and Schmitt, 2010, p. 181).

From the perspective of TS, no translation theorist has, to date, systematically clarified the mechanism in TS by which translators acquire receptive and productive knowledge of collocations in their L2. In other words, most previous studies in this area appear to ignore the importance of L2 collocation learning and acquisition, but concentrate more on other factors to seek the evidence of the difference between translational language and native language from other aspects, such as collocation type-token ratio, degree of collocability, delexicalization and semantic prosody. In respect to this, researching different collocation patterns produced by L1 users and L2 translators can also provide researchers with reliable evidence to identify the features of translational language. However, it would appear that this has not been done in previous studies, thus leaving a

gap in the literature.

More importantly, few collocation studies in this area look at the practical merits of collocation studies regarding how to utilise theoretical achievements and findings to help language users and translators enhance their L2 proficiency. In addition, most previous studies only involve languages of the Indo-European language family, which would call for more reliable empirical evidence from investigating ‘inter-family’ language pairs, such as English and Chinese. Furthermore, language varies in different areas in which it is used, and these varieties of language used for a particular purpose or in a particular social setting are termed as ‘registers’ in language studies. In this sense, register, as a significant factor influencing the variation in the formation of collocations, should also be considered in this type of research.

Therefore, it appears that researchers in this area would need to establish a valid theoretical framework to demonstrate the importance of collocations in translation and clarify the issue of how collocation is associated with the indicators of translation universal features in translational language. Researchers would also need to employ an appropriate research approach, such as a corpus approach, to examine the validity of the established theoretical framework with empirical evidence. Researchers would also need to identify the mechanism of learning collocations from both the L1 and the L2 perspectives, based on which they can suggest a pedagogical and practical model of learning L2 collocations to help facilitate translators’ translation tasks.

1.2 Research objectives

In respect to the issues mentioned above, the present study will use a corpus-driven approach to investigate Chinese translators’ use of English collocations in commercial translation. Firstly, this study aims to clarify the role of collocation in the process of translation and investigate the difference between translational English and native English with regard to collocation use, thus bridging a gap in the literature. From this

angle, this study will discuss the relationship between translators' control of collocation use and the presentation of translation universal features in the target text, and attempt to uncover how translators' collocation knowledge would influence their production of L2 in the target language. This study will use a corpus-driven approach (see 4.3.1) to investigate the use of English collocations from a corpus made up with translated English from Chinese and a corpus compiled with native-speaker English. The data (collocations) collected from these two corpora will be integrated as wholes to provide empirical evidence to support the hypothesis of translation universals (see 2.4.1) and examine whether the statements described in the theoretical framework are fully consistent with the findings from the empirical research.

Secondly, this study aims to clarify the intrinsic relationship between collocation and the indicators of translation universals, and attempts to identify the different collocation patterns produced by Chinese translators and native speakers of English through comparing two designed corpora. Based on the rationale of this study, that is the L1-L2 difference in learning collocations, this study will generalise the features of English collocations used by Chinese translators from a number of perspectives, such as form, meaning and function, and investigate the Chinese translators' deviation in using English collocations with the Contrastive Interlanguage Analysis (CIA) research method. It should be noted that the CIA method in this study combines the traditional Contrastive Analysis approach and the tools of corpus linguistics, aiming to discover the non-native expressions in commercial Chinese-to-English translation. Furthermore, this study also intends to re-assess some previous theoretical models (e.g Wray, 2002) regarding L2 collocation learning and using, and analyse the possible cause leading to Chinese translators' deviation in using L2 English collocations.

Thirdly, this study will only look at the collocation use in commercial Chinese-to-English translation. This is because collocation patterns in this register might demonstrate completely different features when compared with those in other registers, or in a general sense. In other words, some high-frequency collocations used in a particular register might not be of high frequency or might not occur at all in

another register. These high-frequency collocations restricted to a particular register can be simply regarded as specialised collocations. This indicates that register-based research of collocations would bring more convincing and detailed evidence to this research area. In addition, translation is normally classified in terms of register, such as commercial translation, medical translation and legal translation. Accordingly, translation practice, in most cases, requires translators' expertise in one or more areas. This indicates that the accurate use of specialised collocations in a particular register calls for the familiarity with the relevant knowledge in this register. Needless to say, commerce is an area of growing international importance and commercial translation covers a large proportion in translations, so there is a large-scale Chinese-English commercial translation available in the public domain, to which researchers will find it easy to get access. Furthermore, I am a qualified translator, working in the Chinese-English language pair, and a researcher specialising in applied linguistics. My enthusiasm in translational language motivated me to conduct this study by 'marrying' these two disciplines (Translation Studies and applied linguistics).

Last but not least, this study aims to utilise the theoretical findings and attempts to suggest an acceptable pedagogical method for teaching non-English speaking background (NESB) translators how to identify and use appropriate collocations in commercial Chinese-to-English translation. This objective can be two fold: on the one hand, this study will demonstrate Chinese translators' weakness in their use of L2 English collocations and suggest from what aspect they would need to improve their ability of collocation use, so they can come closer to native-like selection and fluency; on the other hand, this study will present a pedagogical model, in which translators can not only add declarative knowledge to what they have known but also integrate procedural knowledge into the L2 collocation knowledge system they have already constructed. Thus, it is hoped that translators can enrich their L2 collocation knowledge 'database' and essentially turn their receptive knowledge into productive knowledge.

1.3 Translator training programmes in China

This study is relevant because it aims to show the recurring problems in using L2 English collocations by the translators who are undertaking commercial Chinese-to-English translations in China. It also attempts to provide solutions to these problems and offer suggestions to the current translator training programmes, particularly in China. Therefore, this section will briefly introduce the background of translator training in China.

In response to the increasing needs for inter-cultural communication, there are several translator training programmes established in China. Translator training strategies at China's higher education institutions, such as universities, are part of the pedagogical activities aimed at the development of versatile graduates (Bai, 2014). The graduates (including postgraduates) have become the majority of the large translation force of nearly one million translators at foreign affairs offices, research institutes, colleges and universities, international trade enterprises, travel agencies, publishers, translation companies, and other workplaces in China (Bai, 2014).

In mainland China, for trainee translators who deal with the English-Chinese translation, the duration of the training session is based on trainee translators' experience, study programme and professional pursuits. The timeframe of the training session could be designed on different levels. For undergraduate students of non-English majors, translation training is carried out with their L2 English learning over their first two years of university study. Trainee translators' translation skills will be improved through in-class tutorials, exercises and practice, and will be assessed in the College English Test (abbreviated as CET, two bands available: CET-4 and CET-6) provided by the Ministry of Education. For undergraduate students of English majors, a separate translation practice course is offered in their third year and/or fourth year of university study. It normally takes 16 to 32 weeks to finish the whole course. Their translation skills will be improved with teachers' instruction of translation strategies and in-class practice, and will be assessed using the Test for English Majors (abbreviated as TEM,

two bands available: TEM-4 and TEM-8) which are also provided by the Ministry of Education. At postgraduate level, translation courses are only compulsory for the students with English-related majors, such as Translation Studies, Applied Linguistics and English Literature, even though they may be available to some students of non-English majors who show their interest in Translation Studies. In the 159 universities that are entitled to grant the degree of Master of Translation and Interpreting, the translation course is given over two years, including both translation theory and translation practice. This normally covers two semesters of the two-year Masters programme, introducing trainee translators to translation theories, translation technology and the code of ethics in translation, and helps them do research topic-based research projects.

Freelance translator training or on-the-job translator training is also available for novice translators across nearly 4,000 translation companies and organisations in mainland China, during a 3-month probationary period for the profession (Bai, 2014). Trainee translators are expected to improve their translation skills through interactive communication with their teachers and intensive workshops where senior translators share their hands-on experience. They may also, at times, have opportunities to be sent to some educational organisations for continuing training programmes. For example, the Translators Association of China may provide, jointly with some international specialists and researchers, an additional two-week training on translation skills, technology, project management, and teaching theories to individual freelance translators and developing translators from higher institutions, social organisations and domestic enterprises.

Generally speaking, translation training programmes are offered at various levels in China and all these translation education institutions are trying to improve their teaching in the hope that they could provide effective and practical translation strategies to those who want to start a career in translation. In this respect, the findings from theoretical research would be particularly important for translation teachers and trainers to consider because they might find them very useful to improve their curriculum design and

provide more sound suggestions to enhance trainee translators' translation skills. This is one of the most important reasons for me to conduct this research.

1.4 Rationale for the research

As mentioned above, there was little research on collocation in translation. Therefore, I decided to carry out a formal study focusing on Chinese translators' use of L2 English collocations in commercial Chinese-to-English translations, in the hope that this study would inform translation educators and translators about the important role of collocation in producing native-like texts in an L2 context. It is also hoped that this study will make significant contributions to both linguistics and Translation Studies in a number of domains as outlined below:

1. It will contribute to theory, by adding to the limited literature on the topic of collocation in translation.
2. It will contribute to the evaluation of previous theoretical models regarding L2 collocation learning and using, by providing more empirical evidence.
3. It will potentially contribute to the research on the distinction between translational language and native-speaker language, by providing new insights and data on different collocation distribution patterns discovered from comparable corpora.
4. It will potentially contribute to the development of corpus linguistics and Translation Studies, by introducing the use of a corpus-driven approach in seeking reliable language data for explaining particular linguistic phenomena in translations.
5. It will contribute to the development of the studies on translation universals, by exemplifying the instances occurring in authentic language materials.
6. It will potentially contribute to translator training, by showing translators' weaknesses in using L2 English collocations and suggesting the use of authentic language materials in translation pedagogy.

1.5 Questions addressed by this research

The present study aims to investigate the actual use of English expressions by Chinese translators when translating commercial documents during the time of this research. To be more specific, it essentially looks at how successfully L2 users of English have utilised collocations in the translation of written documents in the commercial register. Therefore, in the light of the findings and issues from the Literature Review section, this study specifically addresses the following five questions:

Is it possible to define collocation in the commercial register and propose a conceptual framework of collocation? This question looks at the necessity of formulating a reasonable model to explain the role of collocation in language operations. Chapter Three will show that collocation can be analysed on a multi-dimensional basis. After reviewing the literature, I have made an operational definition of collocation (see 3.5.1) for this study and I have also set three criteria for retrieving collocations from raw language materials (see 3.5.1). To construct direct links to the theoretical framework and the explanatory section of the study, I have proposed that the comparison of English collocation use between a translational corpus and a corpus of native English should be carried out on a multi-dimensional basis from quantitative, formal, semantic and functional perspectives. More specifically, this proposal serves as a part of the theoretical framework and will be examined and analysed fully in the explanatory section.

Is it possible to develop a theoretical framework which reflects the role of collocation in translation and key factors that might influence translators' use of (L2) English collocations? This question revolves around how to construct a theoretical model to illustrate the functions that collocation performs during the course of translation. With regard to this point, Chapter Three will clarify the proposed theoretical framework of the study, which makes explicit the role of collocation in influencing the relationship between translation universals and the native-like rendition of the target texts (see 3.5.2). In this section, I have also discussed how translators' implicit knowledge and explicit

knowledge of collocations may influence the use of their L2.

What is the most applicable data retrieval method? This question involves data collection and the data retrieval method best suited to the proposed method of analysis (see 4.4). Data in this thesis refers to collocations extracted from the two designed comparable corpora of translational English and native English in the commercial register. I will outline a detailed procedure of collocation retrieval in Chapter Four by using *FoxPro* programming and *BFSU (Beijing Foreign Studies University) Collocator*. I will also employ the Mutual Information test and the Log-likelihood test as the statistical measures in this study to examine the statistical significance of collocations.

What are the distinctive features of the variation in the use of existing collocations and how can these features be explained within the proposed theoretical framework? This question concerns the analysis of the features of Chinese translators' use of English collocations. This study will examine the deviation of using collocations in a corpus of English translated from Chinese by L1 Chinese translators. These distinctive features of the variation will be analysed based on the results of data analysis and will be employed as the evidence of examining the intended theoretical framework.

How would the findings from this study be useful for enhancing Chinese translators' skills of translating the L2 English output? This question looks at offering constructive suggestions to translators who are L2 users of English based on the findings from statistical and explanatory analyses. This study will provide a collocation list of native-speaker English to help overcome the weakness that Chinese translators have exposed while using their L2 English in the commercial register. It will also show L2 translators of English how they can learn to reduce the unnaturalness in their L2 production and create increased awareness of possible weaknesses based on the findings of this study. This will be helpful if translators intend to turn their explicit knowledge into implicit knowledge, and receptive skills into productive skills, while they are handling translation tasks.

1.6 Chapter overview

In response to the research objectives, this thesis is organised into eight chapters. Following this introductory chapter, Chapter Two will present a review of the study of collocation from the linguistic perspective and emphasise the importance of the research approach of using corpora in Translation Studies. This chapter will also make a distinction between general collocations and specialised collocations based on the existing taxonomies in classifying collocations, and examine the role that collocation plays in language acquisition and development. Chapter Three will describe the rationale of researching collocation in a learner corpus, which underpins the present study, and construct a theoretical framework which is pertinent to the difference between translational language and native-speaker language and emphasises the role of collocation and discusses the inter-relationship among collocation, translation units, translation universals and translators' potential knowledge in language operations. Chapter Four will introduce the methodology of this study, that is, the corpus-driven Contrastive Interlanguage Analysis approach. This chapter will also introduce two designed corpora and elaborate on a collocation retrieval procedure, in which the Mutual Information test and the Log-likelihood test will be employed to examine the statistical significance. Chapter Five and Chapter Six will carry out quantitative analyses regarding the collocation patterns produced by Chinese translators in producing L2 English in terms of amount of use, form, meaning and function. Chapter Seven will outline the role of the control of L2 collocations in translations and demonstrate, with examples, the translation universals which Chinese translators brought in their translations due to their lack of adequate understanding of the features of collocations in their L2. In response to the findings from both the quantitative and qualitative perspectives, this chapter will also discuss the reasons for the deviations in Chinese translators' use of L2 English collocations from the aspect of L1 transfer. Chapter Eight will expand upon the theoretical, practical and pedagogical implications based on the quantitative and qualitative analyses. Chapter Nine will summarise the major findings, present the limitations of this study with regard to research design and suggest the directions for future research in this area.

Chapter Two Literature review: Collocation and Translation Studies

2.1 Introduction

The present chapter attempts to build the intrinsic links between the features of collocation and translation universals, which will greatly help identify the role of collocation during the process of translation and clarify how the corpus methodology has contributed to Translation Studies. Based on previous studies, Section 2.2 will provide an overview of collocation studies with corpus approaches and explore the nature of collocations from the perspectives of form, meaning and function. Section 2.3 will look at the role of collocation in language learning and teaching, and will explain the role in terms of language development and native fluency. Section 2.4 will review the significant impact that corpus approaches have on translation practice and translation theories, and will emphasise the importance of researching translation universals in this study area. In particular, this section will examine translation universals in terms of simplification, explicitation and normalisation, and will discuss how the features of collocations are associated with the indicators of these translation universals.

2.2 Researching the nature of collocations with a corpus approach

Recent decades have seen the rapid development of collocation studies since Firth first stated that “[y]ou shall know a word by the company it keeps” (1968, p. 179). Particularly with the advent of large-scale corpora, researching collocations with a corpus (a corpus in this context can be defined as a collection of authentic and computer-readable texts used for linguistic research) approach has breathed new life into the traditional research methodology and made a significant contribution to this research area. Language researchers using a corpus approach can access the corpus

resources they need to test theoretical hypotheses or examine theoretical frameworks on collocations with empirical evidence. In this respect, the nature of collocations has become a widely discussed topic for any researcher who shows their interest in collocation studies using a corpus approach. There have been a number of researchers (e.g. Kjellmer, 1991; Sinclair, 1991; Stubbs, 1995; Nesselhauf, 2005) carrying out in-depth investigations in the hopes of a more salient presentation in respect to the nature of collocations.

In corpus linguistics, collocation is seen as the combinational relationship between or among lexical items and is forged according to the actual use as reflected in native-speaker language rather than maintained merely through grammatical restrictions. This means that researchers working with a corpus approach believe that language operations are driven by lexis rather than rules, and that collocating words (collocations) are natural occurrences (see for instance Sinclair, 1991). This is an obvious challenge to some traditional rule-based theories (e.g. Chomsky, 1965) which hold that grammatical rules are universal in governing the formation of language with the lexicon being composed of nothing but the elements to ‘fill in the slots’ in the grammatical structure. Furthermore, researchers with a corpus approach believe that lexical items forming collocations can constitute larger continuums of linear symbolic structure in language operations, such as phrases and chunks, so they indicate a kind of relationship of formal, semantic and functional independence. Therefore, the present study will look at the nature of collocations in terms of formal co-occurrence, extended semantic unit, and form-functional composite.

2.2.1 Collocations as formal co-occurrences

Collocation studies from a formal perspective mainly focus on the concept of co-occurrence, that is, the tendency of some particular words to occur together. These studies attempt to investigate the syntagmatic relationship between those words and believe the recurrence of this relationship is a feature in language. For instance, in the

word combination *commit suicide*, *commit* and *suicide* are thought to form a verb-noun collocation. This is to say, whenever *suicide* occurs as a noun in a particular context, *commit* would mostly occur as a verb to ‘initiate the action’. This study examined a total of 10 occurrences of *suicide* in verb-noun combinations retrieved from a one-million-word random sample of the British National Corpus, and revealed that *commit* and *suicide* co-occur 8 times which account for 80% of the total verb-noun combinations containing *suicide*. Therefore, in most cases, words that constitute a collocation predict one another. Studies adhering to this perspective are usually said to employ a statistically oriented approach (see for instance Herbst, 1996, p. 380) or a frequency-based approach (see for instance Nesselhauf, 2005), which is also the underpinning of a corpus approach. The use of a frequency-based approach on collocation studies was first employed by Firth (1957, 1968) and Halliday (1961), and subsequently further developed by Sinclair (1991).

Firth emphasised the formal co-occurrence of words or “mutual expectancy” (1968), and he regarded this as the most important extrinsic relationship within collocating words. According to the previously mentioned statement that “you shall know a word by the company it keeps” (Firth, 1968, p.179), it appears that the meanings of words largely depend on how the words are used (1968) instead of explained in a general sense as “idea” or “concept”. In other words, Firth’s viewpoint about collocations reveals that lexical items in native-speaker language are not isolated but exist in a semantic mode in which they mutually ‘expect’ and predict each other. This is what we call ‘formal co-occurrence’ of collocations. In addition, Firth (1957, 1968) emphasised the importance of employing “attested data” (1957), that is, authentic instances in language, when researching collocations. This was echoed by Sinclair’s proposal of “naturally occurring data” and has become a meaningful guide for the subsequent collocation studies with a frequency-based approach. With regard to collocability, that is, the probability of words co-occurring with each other, Firth (1957) distinguished different types of collocations, such as “habitual” collocations, “more restricted technical” collocations, and “unique” or “a-normal” collocations. This classification was also explained by Sinclair (1991) with his “two principles”, that is, “the open-choice

principle” and “the idiom principle”.

In line with Firth (1957, 1968), Halliday (1961) also views collocation as the tendency of words to co-occur and the probability of each individual word to become the filler of next slot. In his words, “any given [lexical] item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable” (1961, p. 276). This standpoint makes it clear that the meanings of lexical items in collocations are actually reflected in the syntagmatical structure rather than the paradigmatical one. In addition, he argued that “[t]he formal criterion of collocation is taken as crucial because it is more objective, accurate and susceptible to observation than the contextual criterion of referential or conceptual similarity” (Halliday, McIntosh and Stevens, 1964, p. 34). This indicates that collocation should be interpreted or examined on the lexis end of the grammar-lexis spectrum, in which grammar and lexis are considered as two poles which mutually fade away from each other.

This viewpoint of lexis-grammar spectrum underpins the frequency-based research method in corpus linguistics. When discussing the distinction between lexical choice and grammatical, he argued that “[i]n lexis, not only are there more items to choose from at any given point, compared with...grammar; also there is no line to be drawn between those that can and those that cannot be chosen” (Halliday, McIntosh and Stevens, 1964, p. 34). This viewpoint was reflected and developed in his later work with Hasan (1976), in which collocation is regarded as co-occurrence of lexical items, having a lexicosemantic relation and indicating a cohesive role in language operations. In this respect, texts can thus be developed in a linear structure which consequently forms a meaning continuum. Therefore, the linear structure entailed in the combination of lexical items should be regarded as the most important element for researching collocations. For instance, in the verb-noun phrase *over the top*, *over* and *top* form a typical and conventional idiomatic combination if they are used to indicate ‘something far more than usual or expected’. However, if the nominal position was replaced by some other words, such as *peak* (i.e. *over the peak*), in terms of meaning association (paradigmatically), the phrase would be also grammatically and semantically acceptable,

but it would not carry any association indicating ‘something far more than usual or expected’. Therefore, Halliday’s (1961) argument regarding collocations constructs a solid basis for the studies that attempt to explore the nature of collocations.

Sinclair’s views on collocations are a departure from those proposed by Firth and Halliday. Sinclair first distinguished “casual collocations” from “significant collocations” (1991, p. 115) with a frequency-based approach, and this distinction restricts his definition of collocation to the significant collocations only. In this respect, Sinclair defines significant collocations as those that “co-occur more often than their respective frequencies and the length of text in which they appear would predict” (Jones & Sinclair, 1974, p. 21). For instance, in the following sentence:

The Ministry of Collaboration of Foreign Trade Economy undertook an investigation to dumping and dumping profit margin[s] jointly with General Office of Customs.

profit and *with* cannot be considered to form a significant collocation because the word *with*, as a grammatical or functional word, is ‘over active’ and universally collocates with lexical words and pronouns in nearly every register of texts. This word pair is simply a casual collocation and should not be considered as the focus in collocation studies. In contrast, the word combination *profit* and *margin* in the above example make sense in the commercial register, so this word pair can be regarded as a significant collocation in the commercial register. In this respect, the present study will set a language filter to rule out word combinations which do not qualify as a significant collocation. This will be discussed in full in the section of data retrieval (see 4.6.2).

Another contribution Sinclair made to collocation studies is that he clarified a number of important notions, such as “node”, “collocate” and “span” (Sinclair, 1991, p. 170). He developed his definition of collocation as “the occurrence of two or more words within a short space of each other in a text” and collocation “in its purest sense, recognizes only the lexical co-occurrence of words” (Sinclair, 1991, p. 170). In this

definition, a collocation may contain two or more words such as *Chamber of Commerce*, *free on board*, *head office* and *control the budget*, provided they occur within a short space (a span). Most collocations contain a keyword or headword, which is often referred to as a ‘node’ (see Sinclair, 1991, p. 170 for an example). The words which co-occur with nodes within a span are termed as ‘collocates’ (see for instance Sinclair, 1991). Therefore, a short space in the foregoing definition refers to the number of words (span) on each side of the node even though the number may vary. For instance, in the following table, if *accuse* is regarded as the node investigated the number will be five and the node will be considered to occur in a span of ± 5 words.

Table 2.1 Key word *accuse* in a five-word span in the random sampling from the British National Corpus

[1]	that had been quoted. He ACCUSED the newspaper of making them
[2]	with public ownership. He also ACCUSED the Tories of double standards
[3]	negotiating machinery. The Health Secretary ACCUSED the unions of posturing and
[4]	the Leader of the Opposition ACCUSED us of ignoring was regional
[5]	I said before Well, I ACCUSED you of reacting as if
[6]	party together. Its left wing ACCUSES him of collaborating with tainted
[7]	those schools. The hon. Gentleman ACCUSES me of doing that, but
[8]	wrestlers he was, in effect, ACCUSING Downing Street of being out
[9]	telephone calls from his colleagues, ACCUSING him of plotting to destroy
[10]	come under strain, with Virgin ACCUSING Island of putting their own

This guideline of specifying a span is particularly meaningful in collocation studies with a corpus approach. Therefore, this study will adopt a ± 5 word span (see also 3.5.1) with regard to the retrieval of collocations.

Based on the clarification of collocation, Sinclair (1991) also advanced two different and conflicting interpretive principles to account for the mechanism of co-occurring words producing rich linguistic continuums, namely ‘the open-choice principle’ and ‘the idiom principle’. He believes that “[t]he preponderance of usage lies between the two. Some features of language patterning tend to favour one, some the other” (2004, p. 29).

In the open-choice principle, language texts are envisaged as “the result of a very large number of complex choices”, at each point of which “a unit is complete”, such as a word, a phrase and even a clause, and “a large range of choice opens up and the only restraint is grammaticalness” (Sinclair, 1991, p. 109). Sinclair sees this principle as some kind of ‘prototype’ of describing language, or in his own words “probably the normal way of seeing and describing language” (p. 109), based on which virtually all grammars are constructed. Therefore, according to the open-choice principle, language texts result from mutual choices from a large number of lexical items, and the aforementioned ‘grammaticalness’ is also the result of those choices.

However, what is problematic is that words do not occur randomly in a text. Therefore, only employing the open-choice principle is not able to account for all the restraints or deal with the various cases of non-random nature in language use. This can be reflected in Sinclair’s own words:

It is clear that words do not occur at random in a text, and that the open choice principle does not provide for substantial enough restraints on consecutive choices. We would not produce normal text simply by operating the open-choice principle (Sinclair, 1991, p. 110).

For instance, the phrase *a piece of cake* in English is used to refer to ‘an easy task’ or ‘something easy to deal with’. The two words *piece* and *cake* form the relationship of mutual choice to indicate the sense of ‘easiness’, which does not allow other choices, such as *a piece of pancake*, *two pieces of cake* and so forth. Therefore, Sinclair (1991) also proposed ‘the idiom principle’ which allows for register variations and “accounts for the restraints that are not captured by the open-choice principle” (p. 110). Sinclair (1991) believes that things that occur physically together in the world and concepts conceived in the same philosophical area are more likely to be mentioned together in language, thus producing lexical co-occurrences in language operations. This is also the formation of collocations because collocations are coined based on the association between co-related concepts repeatedly conceived in the natural world rather than instantiated

coincidentally.

In contrast, the idiom principle holds that “a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair, 1991, p. 110). Barnbrook (2007) summarised seven features of Sinclair’s idiom principle: “indeterminate extent”, “internal lexical variation”, “internal syntactic variation”, “variation in word order”, “strong collocational attraction for other words”, “co-occurrence with certain grammatical structures” and “tendency to occur in certain semantic environments” (p. 186). Barnbrook explained further that the first four features “allow phrases to vary significantly while still being considered as subject to the idiom principle” whilst the latter three “lead to restrictions of independent choice rather than a complete set of constraints” (p. 186). This summary emphasises the importance of the idiom principle to a large extent and echoes Sinclair’s (1991) claim, “the overwhelming nature of this evidence leads us to elevate the principle of idiom from being a rather minor feature, compared with grammar, to being at least as important as grammar in the explanation of how meaning arises in text” (p. 112).

It is obvious that the idiom principle concentrates on the mutual selection of two or more words based on their regular occurrences. Researching collocational tendencies of lexical items can be simply conducted under the idiom principle. In this respect, Sinclair (1991) stressed that the idiom principle is normally preferred as a default mode because “most of the text will be interpretable by this principle” (p. 114). His viewpoint implies that language actually operates on the basis of those pre-constructed or prefabricated phrases, such as collocations, rather than a set of grammatical rules. The distinction between the open-choice principle and the idiom principle is very useful because it is directly relevant to the issue of how a language should be learnt or taught. Notwithstanding the distinction, these two principles are also complementary, or in Sinclair’s (1991) words, “[w]henever there is good reason, the interpretive process [the idiom principle] switches to the open choice principle, and quickly back again” (p. 114). Furthermore, “[l]exical choices which are unexpected in the environment will

presumably occasion a switch; choices, which if grammatically interpreted [the open choice principle], would be unusual are an affirmation of the operation of the idiom principle” (p. 114).

Based on Sinclair’s (1991) two-principle model, this corpus-driven (see 4.3.1) study will divide collocations into three main categories regarding the degree of collocability from the formal perspective, specifically free combinations, bound collocations and idioms, and investigate Chinese translators’ use of these three categories of collocations in commercial English in comparison with native speakers of English (see 6.2). Free combinations are compositional (see Grant & Bauer, 2004; Grant, 2005) and keep the literal sense of each constituent of the co-occurring words and allow the maximal openness of mutual selection. Free combinations are not formulaic and can include any word pairs which meet the statistical requirements and constitute a collocation based on the data retrieval procedure in Chapter 4, such as *senior officials*, *unfair regulations*, *check details*, *accept_examination* (the use of an underscore “_” means that this collocation allows lexical intervention between its constituents) and *adopt_policies*. Bound collocations are largely non-compositional (see Grant & Bauer, 2004; Grant, 2005) and restrict word choices to a minimal extent and only allow substitution of constituents to a limited set, which would include technical terms and business terminologies used in the commercial register, such as *promote_growth*, *boost_economy*, *trade pact*, *budget deficit* and *tax invoice*. Idioms in the present investigation include both those whose constituents are completely or partially figurative in sense and those whose constituents are completely non-compositional (see Grant & Bauer, 2004; Grant, 2005) and show a strong fixedness in structure and contextual determination, such as *catch up*, *pull out*, *catch_eye*, *tech boom*, *golden rule*, *bear market* and *bull market*. Idioms defined in this study are completely formulaic and do not allow substitution of constituents.

Nevertheless, as previously mentioned, texts are developed in a linear structure which consequently forms a meaning continuum. This indicates that the nature of collocation cannot be captured merely from the formal perspective. In respect to this, Sinclair (2004)

stated that “[t]ending towards idiomaticity is the *phraseological tendency*, where words tend to go together and make meanings by their combinations” (p. 29). He exemplified this point with the following case:

both *door* and *window* have room as a significant collocate - here language does little more than correlate with the world, and adds little distinctive pattern, unlike *slammed* with *door* or *seat* with *window*, where collocational selectivity is evident (p. 29).

This case shows clearly that it is hard to identify the boundary between “a relatively independent item” and an item “with a strongly determining environment” (e.g. a collocation), for which Sinclair hypothesised this should resort to the perception of “an extended unit of meaning” (pp. 29-30). He also proposes that “considering the corpus data, we shall begin in an area of patterning that on intuitional grounds should be relevant — the area of very frequent collocations, idioms, fixed phrases and the like” and this is essential “if we are to find evidence of extended units of meaning” (p. 30). Therefore, the next section will look at the nature of collocation in terms of extended unit of meaning.

2.2.2 Collocations as extended units of meaning

Studies from the semantic perspective mainly look at the “close relationship between the different senses of a word and the structures in which it occurs” (Sinclair, 1991, p. 53). This indicates that the variety of a word’s senses allows for a valid grammatical or syntactical relationship with the different words it collocates with. For instance, *kill* is normally combinable with any nominal lexeme indicating living animates, such as *a man*, *a sheep*, *a bee* and so forth, while not otherwise, such as *a corpus*, *a table* and *a toy*. Traditionally, it was thought that individual words correspond to the primary units of meaning. However, the findings from a large number of collocation studies in the recent decades (e.g. Hudson & Francis, 2000; Sinclair 1991, 2004; Stubbs, 2001) have shown that this might not be the case. Although there are many cases where individual words coincide with their units of meanings, there are still more situations where those units of meaning are entailed merely in word combinations, such as a phrase, a clause or

even a whole sentence.

Sinclair (1991) investigated the word *yield* by employing the central corpus of the Birmingham Collection of English Texts (now The Bank of English). Normally, in this type of research words are lemmatised (a lemma means a group of words in the cases where their inflectional differences are irrelevant). For instance, *yield*, *yields*, *yielded*, *yielding* will be regarded as the same lemma *yield*. Based on the total of 125 occurrences of *yield*, he found three major senses of this word: ‘to give way’, ‘to produce’ and ‘to lead to/to provide’; and three minor senses: ‘to surrender/to collapse’, ‘soft’ and ‘the boundary between districts in a city’. Sinclair discovered a significant tendency that sense and syntax are strongly associated. To be more specific, these three major senses of the word *yield* demonstrate a definite pattern, that is, the first sense is “realized by an intransitive verb”, the second “realized by a noun” and the third by “a transitive verb” (p. 56). Based on the 75 occurrences of *yield* being used in one of its three major senses, he concluded that “all the potential counter-examples to the first tense...do not constitute a case of any strength against the basic coincidence of sense and structure” and the word *yield* in the sense ‘to produce’ can be used occasionally as “a transitive verb” (p. 60) whereas “this meaning is overwhelmingly realized as a noun” (p. 63). Furthermore, with regard to the syntax-semantic relationship he believes that “it is folly to decouple lexis and syntax, or either of those and semantics” and “[t]he model of a highly generalized formal syntax, with slots into which fall neat lists of words, is suitable only in rare uses and specialized texts” (p. 108). Most texts are “made of the occurrence of common words in common patterns, or in slight variations of those patterns” and most words “do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text” (p. 108), such as collocations. In this sense, collocations can be considered as ‘meaning carriers’, which create direct relevance to texts, organise the textual information and predict the development of texts.

Based on the corpus evidence, Sinclair (2004) also proposed that “the notion of a linguistic item can be extended... so that units of meaning are expected to be largely

phrasal” (p. 30). This is because the meaning of a word combination always deviates or differs from the mechanical conjunction of those meanings of its individual constituents. According to Sinclair (2004), “words cannot remain perpetually independent in their patterning unless they are either very rare or specially protected” even though, sometimes, “this is not a formal criterion” (p. 28). He exemplified this situation by using the word *game* (p. 27). The word *game* has no independent existence at all while collocating with other words, such as *give the game away*, *new to a game*, *on the game*, *at their own game*, *all part of the game*, *play games* (from an idiomatic consideration). In respect to this, Sinclair (2004) pointed out that collocation “does not have a profound effect on the individual meanings of the words, but there is usually at least a slight effect on the meaning, if only to select or confirm the meaning appropriate to the collocation, which may not be the most common meaning” (p. 28).

In line with Sinclair (1991, 2004), Hudson and Francis (2000) also emphasised the association between form and meaning. They claimed that “most words have no meaning in isolation, or at least are very ambiguous, but have meaning when they occur in a particular phraseology” (p. 270). This can be evidenced from Stubbs’s (2001) examples, in which *dint*, *kith* and *spick* basically have no independent existence, rather their meanings can only be realised in the phrases such as *by dint of*, *kith and kin* and *spick and span*. This also holds true for grammatical or functional words. Therefore, Hudson and Francis (2000) proposed a similar notion ‘pattern’, that is, “a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups and clauses that follow the word” (p. 3). They set three criteria to identify a pattern: “if a combination of words occurs relatively frequently”, “if it is dependent on a particular word choice”, and “if there is a clear meaning associated with it” (p. 37). The reason why they emphasise the relationship between pattern and sense can be twofold: on one hand, the variety of different senses of words will not be realised unless they occur in different patterns; on the other hand, words that constitute a pattern are also assigned to a part of the sense of the pattern (Hudson and Francis, 2000). In this respect, a pattern is regarded as a syntactic-semantic whole that makes no distinction between form and meaning, and it tends to select its words or elements with the sense it requires.

This is relevant for the current study in that the semantic analysis may lead to the investigation of how the meanings of words are changed, transferred or even lost while collocating with each other.

Stubbs (2001) focused on lexical polysemy and stated that “some [individual] words do not have independent existence at all, but occur only in one combination” (p. 31). He emphasised that words are not always the units of meaning, and provided more corpus evidence by examining a total of 82 occurrences of *bank* and 28 occurrences of *banks* based on the linguistic contexts in the one-million-word Lancaster-Oslo/Bergen Corpus of British English (LOB). The word *bank* can be used to refer to a “place where you keep money” such as *Bank of New Zealand*, or refer to an “area of sloping, raised ground or ramp” such as *river bank* (p. 14). Stubbs discovered that “[m]any occurrences [of *bank*] were in fixed phrases which signalled unambiguously the ‘money’ [such as *bank account, bank balance, bank robbery*] or ‘ground’ sense [such as *canal bank, river bank*]” (p. 15) with only very few cases left ambiguous (such as *the Worthing bank murder case*). He concluded a number of principles pertinent to lexical meaning: firstly, lexical meaning is invisible and is impossible to observe, but it can be inferred from observing the words around, especially in repeated co-occurrences; secondly, lexical meaning is dependent of the context, in which a word may predict the occurrence of another; thirdly, “invented and decontextualized examples may exaggerate difficulties” (p. 16) that go beyond the normal the interpretations from a semantic theory; finally, the findings obtained with a corpus approach are reliable and can afford re-testing.

Since meanings of words are hard to observe and strongly associated with the contexts in which they occur, researchers in this area would inevitably take account of the linguistic tendency of what Sinclair (1991) terms “a progressive delexicalization” or a “reduction of the distinctive contribution made by that word to the meaning” (p. 113). This tendency is so frequent that it greatly exceeds what is expected. The word *have* is just a case in point, because the meaning of *have* mostly gets lost in the uses such as *have a rest, have a talk, have a chat, have a meeting, have a look, have a try, have a go* and so forth. In respect to this, Stubbs (2001) investigated the lemma pair *take a* by

employing a 2.3-million-word corpus of contemporary English. He discovered that only around 10% of more than 400 occurrences of *take* possess a literal meaning of “grasp with the hand” or “transport” (p. 32). Therefore, he generalised that the most commonly delexicalized use of *take* occur in word combinations such as *take a close look at*, *took an interest in*, *take a deep breath*, *takes a photograph*, *take a decision* and so forth. Almost all the meanings of these observed combinations are carried by the nouns. Stubbs’s findings provide more evidence to support the standpoint that meanings of lexical items are not only realised by individual words but are largely dependent on the collocations which they constitute. In this sense, collocations can be best summarised as ‘extended units of meaning’. In addition, Stubbs’ (2001) findings regarding delexicalized words in English are particularly meaningful for investigating delexicalization in learner language. For instance, whether L2 learners tend to under-use collocations with delexicalized meanings in a particular register (e.g. commerce) is a topic that appears to be worth researching. Therefore, this study will discuss delexicalization in more detail and examine Chinese translators’ use of collocations with delexicalized meanings in commercial English in comparison with native speakers of English (see 6.3).

2.2.3 Collocations as form-function composites

Apart from the studies from the formal and semantic perspectives, some language researchers (e.g. Nattinger & DeCarrico, 1992) also examined collocations from the functional perspective and discovered that collocations are strongly related to communicative situations so collocations perform pragmatic functions. According to Nattinger and DeCarrico (1992), collocations, or in their term “lexical phrases”, are assigned “functional meanings” and can be referred to as form-function composites (p. 11). They claimed that collocations are not only syntactically structured but also capable of “performing pragmatic acts” (p. 11), and that collocations serve as basic forms for “speech acts” such as promising, complimenting, asserting, and so on (p.11). Their claim demonstrates clearly that pragmatic competence involves “restrictions on the

choice of syntactic rules that are allowed to apply to these form/function units in context” (p. 15). This can be illustrated in the following figure:

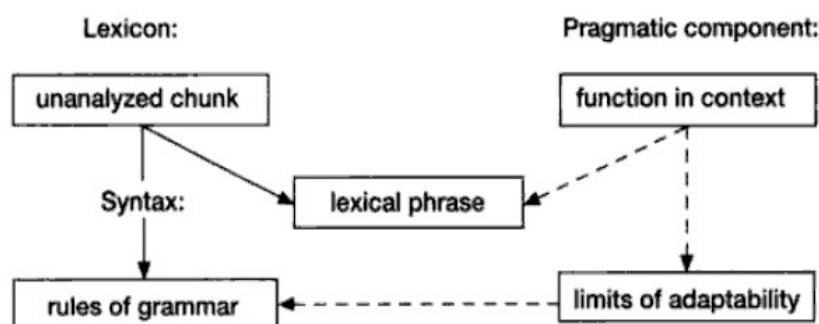


Figure 2.1 Interface of pragmatic competence and linguistic competence in collocations (Nattinger & DeCarrico, 1992, P. 16)

In Figure 2.1, the solid lines refer to the process involved in grammatical competence, while the broken ones refer to pragmatic competence. Collocations (lexical phrases as in the figure) in this model are believed to be largely conventionalised and are explained as form-function composites that are involved in pragmatic principles and restricted to the selection of grammatical rules. It is also noticeable that in this model pragmatics interacts with other components of the grammar. In other words, pragmalinguistic competence fills the gap between grammar and the general rules of language use. That could also be a precise explanation for the use of the word ‘composites’ rather than ‘units’. Therefore, pragmatic competence “is positioned on a continuum somewhere between strict grammatical competence on the one hand, and performance factors such as processing, memory limitations, false starts, etc. on the other” (p. 8), and chooses “form/function composites [collocations] required for particular circumstances” (p. 11). For instance, the collocating words *pronounce (you) husband and wife* serves a declarative function and it is mostly used when a priest or minister is officiating a wedding ceremony. This is relevant to the current study in that the functional analysis may help examine whether translators have used appropriate collocations in appropriate places.

Nattinger and DeCarrico’s study (1992) indicates that in language learning the

knowledge of collocations plays an important part in building up language users' communicative competence in social interactions with regard to how language should be used 'correctly' in 'correct' places. In this sense, collocations, from Yorio's (1980) view, "offer social support to deal with situations that are awkward or stressful" and "make communication more orderly because they are regulatory in nature", thus "reducing the complexity of communicative exchanges" (p. 438). Therefore, when second language (L2) learners accumulate their L2 knowledge, it is important to identify the functional meanings when they are learning a wide range of collocations.

Nevertheless, as Wray (2002) noted, "the relationship between a linguistic form and its function is rather unpredictable" in language learning and using because "it is virtually impossible to predict precisely what form the linguistic unit used for that purpose will take" (p. 53). This causes problems for L2 users who may have mastered a large repertoire of English collocations but do not know exactly how to use them all correctly in socio-communication to complete the expected functions, because these functions in language often vary with a lot of factors, such as time, place, register and stance, even when trying to express the same meaning. Therefore, how to use English collocations to express the 'correct' meaning in the 'correct' place still remains a core problem not only for EFL (English as a Foreign Language) learners, such as Chinese translators in this study, but also for those researchers who specialise in the study of EFL. This is also one of the rationales underpinning the research regarding the difference in using collocations between native (L1) and second language (L2) users.

Functional categories have been formulated from various perspectives (e.g. Aijmer, 1996; Butler, 1997; Cowie, 1988; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Wray, 2002; Yorio, 1980). This study aims to look at whether in translational English Chinese translators have achieved the functional effects which appear in native-speaker English. It also aims to examine whether unconventional English collocations in native commercial English are still used to perform the desired functions in translational commercial English. Therefore, this study employs semantic prosody as the indicator across the two corpora to explore the functional features of the Chinese translators' use

of English collocations. It should be noted that even though the study of semantic prosody is often carried out in the semantic domain, semantic prosody is also an important indicator to examine functional meanings in language users' socio-communication, thus serving attitudinal, evaluating and communicative functions of lexical items (see for instance Louw, 2000; Sinclair, 1991; Stubbs, 2001). Sinclair (2004) stressed that "the initial choice of semantic prosody is the functional choice which links meaning to purpose" (p. 34). In this respect, semantic prosody can serve as a measure to investigate the functional features of Chinese translators using English collocations in comparison with native speakers in the present study.

Semantic prosody can be defined as a "consistent aura of meaning with which a form is imbued by its collocates" (Louw, 1993, p. 57) or "a form of meaning which is established through the proximity of a consistent series of collocates" (Louw, 2000, p. 57). Sinclair (1991) exemplified this with the phrasal verb *set in* (pp. 74-75), which normally collocates with words with negative meanings, such as *rot*, *decay* and *rigor mortis*, so it is assigned a negative aura. Similarly, the word *achieve* can possess a positive semantic prosody, because it often collocates with words with positive meanings, such as *success*, *stability* and *progress*. In addition, semantic prosody is not restricted to individual words, it can also exist in larger language units, such as phrases and lexical chunks. In this respect, collocations can also have semantic prosody because semantic prosody, according to Xiao and McEnery (2006), is "at least as inaccessible to a speaker's conscious introspection as collocation is" (p. 106). For instance, collocations such as *break out*, *set in* (see for instance Sinclair, 1991), *bent on*, *build up of*, *end up (doing)* (see for instance Louw, 1993, 2000), *cause distress* and *cause damage* (see for instance Xiao & McEnery, 2006) have negative semantic prosodies, whilst collocations such as *build up a* (see for instance Louw, 1993), *cause pleasure*, *understanding (and) development*, *growth (and) development* (see for instance Xiao & McEnery, 2006), *successful stories* and *great achievement*.

Nevertheless, it is noteworthy that, as Xiao and McEnery (2006) pointed out, "collocation patterns and semantic prosodies can vary across text categories [registers].

The difference is more distinct between texts in general domains and technical or specialized texts” (p. 126). The present study concentrates on collocations of commercial English, so the collocation patterns found in respect to functional features may differ from those obtained in other studies (e.g. Louw, 1993; Partington, 1996, 2004; Xiao & McEnery, 2006) which focus on general English. In this sense, it is meaningful to view collocations as form-function composites and investigate different semantic prosodies in the commercial register. This will be examined in more detail in Section 6.4.

2.3 Role of collocation

Since collocation is a key part in language formation and language operations, it would be worth reviewing what role it plays during the course of language acquisition and language teaching. Therefore, this part will primarily look at the role in terms of language development and native fluency.

2.3.1 Collocations facilitate language development

It is obvious that formulaic language plays a vitally important role during the whole process of language acquisition and language development. To clarify this point, a number of researchers (e.g. Brown, 1973; Peters, 1977) proposed that native speakers start learning a language by incorporating unanalysed and unglossable structures at a very early stage, in which they generalise productive rules and reanalyse afterwards. This claim largely confirms the role of continuing words, such as collocations, even though it is not very systematic. According to Peters (1977), native (L1) learners are either more analytical or more holistic and that the two mentioned approaches are not basically exclusive.

Wray and Perkins (2000) explored further and took the proportions of holistic and analytical processes into consideration. They claimed that the balance between those

two processes varies during the course of L1 language acquisition and development, and proposed a theoretical model that specifically consists of four phases as illustrated in Figure 2.2.

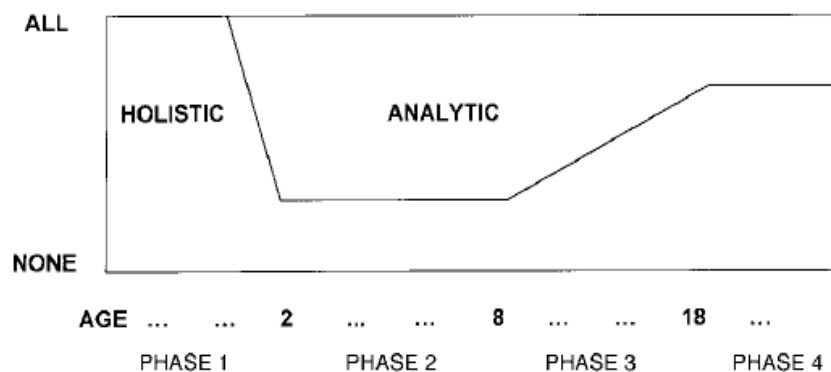


Figure 2.2 Relative proportions of holistic and analytical involvement in language processing from birth to adulthood (schematic representation) (Wray & Perkins, 2000, p. 20)

In Figure 2.2, Phase 1 (from birth to around 20 months) is nearly dominated by holistic processing. This indicates that a native speaker’s “earliest goal is one of social integration and the meeting of its physical needs” and “requires the accumulation of a set of formulaic sequences that successfully achieve that end” (p. 21). At this stage, speakers are not capable of accessing the internal structure of linguistic units. Therefore, they are largely dependent on memorised vocabulary in communication, which may include individual words and some sporadic formulaic sequences. At Phase 2 (from 20 months to the age of 8), the speakers’ grammatical awareness (or in Wray’s term “grammatical analysis module”) starts to operate and they are able to select and store “a sufficient and requisite number of salient formulaic linguistic items to activate a specifically language-oriented analytical mechanism which, through identifying commonalities among the stored formulae, begins the process of creating a generative grammar...” (p. 21). This means that formulaic continuums start to function as a device for developing speakers’ ability to seek universal rules in language whereas individual words can hardly perform this function. Furthermore, “this stage of development is marked by a preference for analytic over formulaic language processing” (p. 21) even though the holistically-processed language is still on the increase. During Phase 3 (until the age of 18), formulaic language (e.g. collocations) again becomes more prominent.

For L1 speakers, organising the language system “becomes progressively more formulaic” and the analytical mechanism plays a significant part in “constantly readjusting the formulaic continuum by deciding whether a given item is unique, or else shares sufficient properties in common with other items to justify subsequent collapsing and re-storage as a single, partly productive formulaic frame” (p. 21). In addition, “language production increasingly becomes a top-down process of formula blending as opposed to a bottom-up process of combining single lexical items in accordance with the specification of the grammar” (p. 21). The balance of the holistic and analytical processes keeps changing constantly before they are developed into adult patterns. In Phase 4 (aged 18 or above), the holistically-processed language and the analytically-processed language are in a comparatively stable state of balance, which means a fully equilibrated system starts to take effect.

However, according to Wray and Perkins (2000), this balanced status cannot be achieved until late childhood. This model indicates how the memory-based mechanism operates in L1 acquisition and development could be useful for investigating, and, to a great extent, implies the role of formulaic sequences in the process. Based on Wray and Perkins’s (2000) model, the present study will briefly suggest a model of relative proportions of L2 users’ holistic and analytical involvement in language processing from the beginning level to the advanced level (see 6.1), and will attempt to discuss how collocations are produced in translators’ L2 knowledge system.

In general, this model shows clearly that formulaic language, such as collocation, functions during the whole process of language acquisition and greatly facilitates language development. It also helps trigger the activation whereby “the child [L1 speaker] is afforded the luxury of developing the analytic grammar by being protected, during these vital years, from the need to accumulate the wide range of formulaic sequences that it will ultimately need in order to function as a normal social adult” (p. 22). This point is very important because it is directly pertinent to the issues of how to understand correctly the role of collocation in language development and design a suitable model of situated learning (see 6.3.2) in respect to teaching collocations in an

L2 environment. This is because, as documented for instance by Nesselhauf (2005), even adult advanced learners of English have difficulties at times in dealing with collocations, which is mostly caused by learners' lack of 'automation of collocation' (see for instance Kjellmer, 1991).

The present study involves Chinese-English bilingual translators, some of whom may be said to still find themselves somewhere along the L2 learner continuum. In this sense, it is necessary to show these difficulties that translators may come across and the errors they may have made in translation practice, and emphasise the important role of collocation facilitating language development (see 3.5.2). Therefore, Wray and Perkins's (2000) model appears to be very useful when researchers in this area attempt to break through the restraints of traditional pedagogical frameworks and suggest valid strategies, such as Davies's (1998, 2004) proposal of "situated learning" and Crezee and Grant's (2013) recommendation of learning collocations using authentic texts, to help translators smooth away difficulties and avoid making errors in learning/using L2 collocations. This will be discussed in more detail in Section 8.3.2.

2.3.2 Collocations help achieve native-like language proficiency

Collocation can not only facilitate language development but also play an essential role in producing fluent native-like language. It is very important to emphasise the importance of native norms in this context because the present study aims to investigate how translators can learn to ensure that their translations appear more native-like in terms of the use of collocations. In other words, if L2 translators were to achieve native-like language production, collocation should always be a core part determining the result. Neurolinguistic and psycholinguistic evidence (e.g. Paradis, 2004) has shown that the human brain is more functional in memorising information than processing information, although this viewpoint might need more reliable evidence. In this respect, the availability of a great number of memorised language chunks, such as collocations, is very helpful in reducing the effort of processing information (Pawley & Syder, 1983,

2000; Partington 1996; Nesselhauf, 2005).

Pawley and Syder (1983) elaborated two notions regarding linguistic competencies, i.e. “nativelike selection” and “nativelike fluency” when investigating memorised language. The former of the two notions concerns “the ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also nativelike”, whilst the latter involves “the native speaker’s ability to produce fluent stretches of spontaneous connected discourse” (p. 191). Based on that distinction, they also elicited two corresponding puzzles: on the one hand, how “he [the native speaker] selects a sentence that is natural and idiomatic from among the range of grammatically correct paraphrases, many of which are non-nativelike or highly marked usages”; on the other hand, “human capacities for encoding novel speech in advance or while speaking appear to be severely limited, yet speakers commonly produce fluent multi-clause utterances which exceed these limits” (p. 191).

These two puzzles are pertinent to exploring the mechanism of native speakers forming formulaic linguistic sequences and are therefore also relevant to this study, which explores the same mechanism for non-native speakers. Pawley and Syder (1983) also stated that the ability of nativelike selection depends on a large body of “sentence stems” that native speakers have mastered. The sentence stems are “institutionalised” or “lexicalised”, which means the expression is “a conventional label for a conventional concept, a culturally standardized designation (term) for a socially recognized conceptual category” and the usage “bears the authority of regular and accepted use by members of the speech community” (p. 209).

Pawley and Syder (1983) concluded that if a language learner/speaker is to be accepted as a native speaker, he must acquire “a generative grammar”, i.e. “a system of rules which enumerates the infinite set of sentences in the language, assigns correct structural descriptions to these sentences, and distinguishes them from ungrammatical sequences” (p. 104). This point is now universally accepted in this research area. In addition, a language learner/speaker must also master an enormous number of lexicalised sentence

stems and learn “a means for knowing which of the well-formed sentences are nativelike”, whereby he can distinguish “those usages that are normal or unmarked from those that are unnatural or highly marked” (p. 194). For instance, in English, the word combinations *capital punishment* and *death penalty* are both regarded as acceptable collocations but they cannot be ‘mixed and matched’, such as **capital penalty* and **death punishment*. This implies that institutionalised sentence stems such as collocations are fixed or semi-fixed word combinations with clear meanings and are recognised by the speakers in a particular speech community. Therefore, native selection is to a great extent culturally institutional and is one of the most important factors contributing to the production of native language. This is relevant to this study which examines the L2 English production by L1 Chinese translators.

Based on a range of analyses of situational cases, Pawley and Syder (1983) found that when native speakers deliver their speech they do not rely heavily on grammatical rules (even though they have unconsciously acquired them). In other words, native speakers do not normally commit themselves to “grammatical constructions” that call for the account of “the structure of an earlier or later clause when formulating a current one” (p. 202). In contrast, native speakers prefer a “clause chaining” style, in which they tend to string together “a sequence of relatively independent clauses, clauses which show little structural integration with earlier or later constructions” (p. 202). This point is echoed by Cowie’s (1994) observation that “native-like proficiency of a language depends crucially on knowledge of a stock of prefabricated units” (p. 3168). In a “clause chaining” style, “a speaker can maintain grammatical and semantic continuity because his clauses can be planned more or less independently, and each major semantic unit, being only a single clause, can be encoded and uttered without internal breaks” (p. 203). In this sense, the “clause chaining” style largely increases the possibility of some words co-occurring and constituting memorised language sequences in native speakers’ language system. Therefore, the use of collocating words (or the kind of memorised ready-made expressions) reduces speakers’ encoding work to some extent and enables them to have enough planning time to organise other activities in speech, such as widening a conversation topic or structuring a larger discourse by expanding the

existing constructions. In return, the saved time for organising these activities can also help them increase their fluency in language communication.

On the whole, Pawley and Syder's (1983) proposal of 'nativelike selection' and 'nativelike fluency' is of particular importance because this study aims to identify the role of collocation in the process of translation and will attempt to suggest an approach to teaching collocations in an L2 environment based on the observed corpus evidence. Their proposal indicates that collocations may affect language output through a chunking mechanism, which can be reflected in what speakers say and how they say it. In terms of the present study, this might suggest that the effective use of L2 collocations will help translators organise translation units smoothly and produce native-like target texts. This not only enables translators to facilitate their translation tasks but also makes them come closer to native speakers in their use of the L2. In contrast, if translators are not able to use L2 collocations effectively they might find it difficult to produce native-like translation units and may bring some features of translation universal (see 2.4) into the target texts. This will be discussed in full in Section 3.4.2.

Pawley and Syder's (1983) proposal of 'nativelike selection' and 'nativelike fluency' also implies that teaching L2 collocations is strongly associated with authentic and reliable language materials. This indicates that the traditional 'presentation-practice-production' approach might not be able to achieve the expected outcomes (see for instance Xiao & Xu, 2008) and there is a need for more suitable models contributing to this research area. Carter and McCarthy (1995) proposed a 'three I's' (illustration-interaction-induction) model. In this model, 'illustration' means observing the authentic data from native-speaker language; 'interaction' indicates exchanging opinions about what has been observed; and 'induction' (see also Robinson, 2003 from the perspective of translator training) means generalising rules from facts. This model indicates that when L2 learners are exposed to collocations they can build on any rules they may have induced and apply those rules in language use to seek more data worthy of observing in native-speaker language. Thus, the rules obtained can also be assessed over and over again until they are essentially applied automatically. This point is

also reflected in Paradis's (2004) differentiation between implicit and explicit types of language knowledge. Paradis's (2004) claimed that implicit knowledge refers to "the knowledge inferred from individuals' systematic verbal performance" (p. 7) and a thoroughly learnt implicit rule is used without awareness and without effort. Contrary to implicit knowledge, explicit knowledge refers to the knowledge that "individuals are aware of" and that "they are capable of representing to themselves and verbalizing on demand" (p. 8). In other words, an old explicit rule is used consciously, but with relative speed. The present study has shown that even though some translators have already reached the implicit knowledge stage, many more developing translators still need to think about what they are doing when they are handling translation tasks. In this sense, it is necessary to clarify translators' knowledge system of L2 collocations and suggest a suitable model providing strategies regarding how to help translators to reach the implicit knowledge stage of English collocations. This point will also be discussed in more detail in Section 3.5.2 and Section 8.3.2 respectively.

2.4 Corpus approaches for Translation Studies

The field of Translation Studies (TS) is concerned with the methodological examination of relevant theoretical models and hypotheses of translation through valid methods, such as objective description and case analysis. Traditional research conducted in TS, in Baker's (1993) words, is "concerned exclusively with the relationship between specific source and target texts, rather than with the nature of translated text as such" (p. 234). In other words, TS is still seen as a research field rather than an independent discipline, and, in return, relevant research in TS is still conducted in terms of equivalence, correspondence, and shifts of translation, which "betray a preoccupation with practical issues such as the training of translators" (pp. 234-235).

However, in the last two decades, corpus approaches (see 4.3.1) have made a significant contribution to Translation Studies. With regard to translation practice, there is an increasing demand of corpus resources from translators who employ computer-aided

methods, such as translation software tools, to deal with their translation tasks. With easy access to large-scale corpora, particularly parallel corpora, translators can align the existing translation materials with a translation memory system (a database which stores translated language pairs for future reference in translations) so that they can manage terminologies more effectively and efficiently (see also 8.3.2). In this sense, corpus-based translation techniques have greatly facilitated translators' work. With regard to translation theories, more and more researchers (e.g. Gellerstam, 1986, 2005; Baker, 1993, 1996, 2004; Laviosa, 1998a, 1998b, 2002; Maurenen, 2004; Xiao, 2010) who specialise in Translation Studies (TS) have noticed the importance of using corpus approaches to expound, test or exemplify translation theories in that they are able to obtain reliable linguistic evidence from authentic texts.

Baker (1993) proposed that researchers could elucidate "the nature of translated text as a mediated communicative event" (p. 243) by comparing translational language and native-speaker language, which is also the most important task of using corpus approaches in TS and the rationale underpinning the present study. This is also reflected in Zanettin's (2013) statement that "corpus-based translation studies (CTS or CBTS) is an established subfield of the descriptive branch of the discipline, and includes a number of different lines of inquiry" (p. 21). These viewpoints indicate that the scope and research targets in TS need re-defining in that the focus of research in TS has moved to the discussion of the nature of translational language, and has led to the consideration of how theoretical research can bring the practical merits into translator training. All of these paradigm shifts greatly help TS develop as an independent discipline.

With corpus approaches, a great number of researchers (e.g. Baker, 1993, 1996; Pym, 1995; Toury, 1995; Gellerstam, 1986; Laviosa, 1998a; Kenny, 1998, 2000, 2001; Olohan & Baker, 2000; Chesterman, 2004; Maurenen, 2007) in this research area see translational language as a distinctive language variety (e.g. Frawley, 1984 and his 'third code') and believe that there are some invisible universal rules (translation universals) hiding behind the linguistic variants produced by translators. In respect to

this, Mauranen and Kujamäki's (2004) claimed that "the idea of linguistic translation universals has found a place at the centre of discussion in translation studies" (p. 1). This is also echoed by Zanettin (2013), who believes that the main research strand with a corpus approach is the hypothesis of translation universals, which is also one of the most significant and challenging areas in Descriptive Translation Studies (DTS). It should be noted that translation universals can be defined as the inherent features revealed in the translated texts, independent of source language, which can essentially distinguish translational language from native-speaker language (e.g. Baker, 1993). Researchers have proposed a wide range of TUs which include explicitation (e.g. Blum-Kulka, 1986; Baker, 1996; see also 3.4.1), simplification (e.g. Baker, 1996; Laviosa, 1998b; Olohan & Baker, 2000), normalisation (e.g. Baker, 1996; Mauranen, 2007), sanitisation (Kenny, 1998), convergence (e.g. Baker, 1996; Laviosa, 2002) and so forth.

Nevertheless, as Low (2003) puts it, "one should logically expect that some focus on function and purpose would help a translator to decide which features to prioritise in a given case and which may be sacrificed at less cost" (p. 93). In this respect, the present study will focus only on the first three categories, i.e. explicitation, simplification and normalisation because they are the most obvious features which would suffice to distinguish translational language from native-speaker language. In addition, this study concentrates on collocations, therefore, it will discuss these translation universals in more detail and attempt to clarify the association of the features of collocations with the indicators of translation universals.

2.4.1 Translation universals

As outlined in the previous section, translation universals are discussed in terms of simplification, explicitation and normalisation. Studies involving simplification show that the language in translated texts is simpler than that in the same target language. Translators tend to "unconsciously simplify language or message or both" (Baker, 1996,

p. 176) when generating target texts. A number of researchers (e.g. Baker, 1996; Laviosa-Braithwaite, 1997; Laviosa, 1998b; Olohan & Baker, 2000; Xiao, 2010) have attempted to provide corpus evidence with regard to simplification in translational language. These studies indicate that simplification can be observed and examined on lexical, syntactical or even stylistical representations, and it is particularly evident from the lexical aspect.

Baker (1996) sees lexical density and type-token ratio as indicators of simplification, where the former “relates to the proportion of lexical as opposed to grammatical words in a corpus” and the latter is “a measure of the range of vocabulary that is used in a text or corpus” (p. 183). In respect to this, Baker (1996) proposed that the use of a narrower range of vocabulary is “a feature of text addressed to non-native speakers of a language” because “these texts are easier to process” (p. 183). This is echoed by Laviosa’s (1998b) corpus-based study to some extent: she proposed lexical variety (vocabulary range) and lexical density as indicators of simplification and discovered that non-translational language demonstrates a significantly greater lexical density than translational language. Xiao (2010) was even more specific when he argued that if a translated text of a language shows “a relatively lower proportion of lexical words over functional words”, “a higher proportion of high-frequency words over low-frequency words” or “a higher repetition rate of high frequency words” (p. 29) than the text generated by native speakers of that language, then there is simplification in the translated text. It will be interesting to see if this applies to the current study also.

For collocation studies, simplification can be examined in terms of collocation density by comparing translational language and native language. Therefore, the present study will focus on the type-token ratio of collocation (see 5.2) and use this as the formal operator by comparing a target corpus made up with translational English and a reference corpus of native-speaker English (see 4.4). It will also examine the collocation distribution patterns across the two corpora so as to obtain evidence to support the hypothesis of simplification in translational English. In addition, this study will look at the formal features of collocations and use this as another indicator of simplification,

and will instantiate these features by examining the proportion of free combinations, bound collocations and idioms found in both corpora (see 6.2). Furthermore, this study will provide a qualitative analysis of simplification with examples found from the two corpora (see 7.2.1).

Explication refers to an inherent feature of translators making implicit information in the source language explicit in their translations where such implicit information does not need to become explicit in the target text. Ben-Shahar (1994) found a typical example regarding explication in the Hebrew translation of William Faulkner's *Sanctuary*, in which some particular sentences in the source text were 'explained' in more detail through adding explanatory vocabulary, such as the conjunction *but*. Similarly, Xiao (2010) also found that connectives, such as 以至于 (*yi3zhi4yu2*: so...that...), 换句话说 (*huan4ju4hua4shuo1*: in other words), 虽说 (*suishuo*: though), 总的来说 (*zong3de4lai2shuo1*: in short) and 一来 (*yi1lai2*: first) (p. 25), are used more frequently in translated Chinese than in native Chinese (see also Chen, 2006).

Explication was first examined by Blum-Kulka (1986) who proposed "an observed cohesive explicitness from SL [the source language] to TL [the target language] texts regardless of the increase traceable to difference between the two linguistic and textual systems involved" (p. 19). This was echoed by Baker (1996) who claimed that translators tend to "spell things out rather than leave them implicit" (p. 180) during the process of translation.

With regard to the research methodology, Blum-Kulka (1986) postulated that explication should be investigated based on the empirical evidence from individual sample texts and suggested that this could be realised through "examining different types of interlanguages, from those produced by language learners to the products of both non-professional and professional translators" (p. 19). Blum-Kulka (1986) employed six groups of English (SL)-French (TL) sample translations made by bilingual graduate assistants working on the Harvard Literacy Skills project, in which the TL texts are longer than the corresponding SL texts, and both lexical and syntactical

transformations can contribute to explicitation. Øverås (1998) investigated the English-Norwegian Parallel Corpus by retrieving 50 first sentences from 40 novels and their corresponding translations, among which 20 novels are from English to Norwegian and the remaining 20 the other way round. She found 347 instances of explicitation from English-Norwegian translations with average shifts being 17.3 per text, and 248 instances from Norwegian-English translations with average shifts being 12.4 per text (p. 15). Explicitation is observed in all these 40 texts. In particular, there were 33 texts (among the 40 investigated texts) containing explicitation more than implicitation. Similarly, Chen (2006) focused on popular science writings and found that connectives, i.e. conjunctions and sentential adverbials, are more common in translational Chinese than native-speaker Chinese. He also compared the translated Chinese target texts with the English source texts based on a case study of five identified ‘translationally distinctive connectives’ (TDCs), and found that there was a statistically significant difference regarding the use of these connectives.

The studies documented in the literature bring up convincing evidence from authentic language materials and support Blum-Kulka’s (1983, 1986) explicitation hypothesis to a great extent. Researchers in this area look for linguistic indicators of explicitation from a range of levels, which includes lexis (e.g. Laviosa, 1998b), syntax (e.g. Olohan & Baker, 2000; Kenny, 2001), discourse (e.g. Øverås, 1998) and so forth. However, to my knowledge, no researcher has previously attempted to examine explicitation from the semantic aspect. For collocation studies, explicitation can be examined from the semantic aspect because collocations can be regarded as extended units of meaning. As set out in Section 2.2.2, this study will look at the semantic features of collocations in terms of delexicalization. It will use delexicalization as the indicator to investigate explicitation by comparing translational English and native-speaker English, and will examine the collocation distribution patterns across the two corpora so as to obtain evidence to support the hypothesis of explicitation in translational English. Specifically, this study will instantiate these semantic features by examining the proportion of collocations with a literal sense and those with a delexicalized sense (see 6.3). In addition, this study will provide a qualitative analysis of explicitation with examples

found from the two corpora (see 7.2.2). In this sense, this study attempts to fill a gap in the literature.

Normalisation, also termed as “conservatism” (see for instance Baker, 1996), “sanitisation” (see for instance Kenny, 1998) or “conventionalisation” (see for instance Mauranen, 2007), refers to translators’ “tendency to exaggerate features of the target language and to conform to its typical patterns” (Baker, 1996, p. 183). Normalisation shows that translators’ use of their L2 is, at times, more ‘conventional’ and ‘normalised’ than the target language, “producing more conventional rather than unusual target strings” (Zanettin, 2013). According to Baker (1996), normalisation is strongly associated with the status of the source language, for which she proposed that “the higher the status of the source text and language, the less the tendency to normalise” (p. 183). This explains Toury’s Law of Interference that “tolerance of interference [...] tend[s] to increase when translation is carried out from a ‘major’ or highly prestigious language/culture” (1995, p. 278). In other words, the more prestigious and powerful is the source language or culture, the less the need to comply with the conventions of the target language, and vice versa.

Ben-Shahar (1994) also discovered that marked structures are always ‘normalised’ by translators, for which she claimed that “[t]ranslators’ tendency to formal equivalence translation makes them translate such elements whenever they occur in a source text, even where their use in Hebrew is less frequent than it is in the source language” (Ben-Shahar, 1998, p. 5). Vanderauwera (1985) examined 50 novels translated from Dutch into English and found that “translators of Dutch fiction exhibited reserve in rendering unusual and mannered imagery and word choice in the target text” (p. 108). Malmkjær (1998) provided more evidence that in multiple translations from Danish (SL) into English (TL) only the translations from the minority of translators are preferable because “the original [text]...contravenes a norm for Danish which is equivalent to the norm for English” (p. 5).

These studies largely support the hypothesis of normalisation, and normalisation is

particularly obvious when translators use “typical grammatical structures”, “punctuation” and “collocational patterns or clichés” (1996, p. 183) in the target language. For this point, Xiao (2010) also proposed that “over use of typical features of the genres involved” and “the treatment of the different dialects used by certain characters in dialogues in the source texts” (pp. 10-11) might also serve as significant factors which contribute to normalisation. Indicators of normalisation can be examined at the lexical level, such as degree of lexical and collocational creativity (see for instance Kenny, 2001; Olohan, 2004) and degree of formality (see for instance De Sutter et al., 2012), or at the syntactical level, such as the distribution of typical and atypical register features (Xiao, 2010), or at the semantic level, such as the range of terms used to represent the conceptual domain of colours (Olohan, 2004). Nevertheless, no researcher, up to date, has attempted to examine normalisation from the functional aspect. This study will attempt to address that gap in the literature.

For collocation studies, normalisation can thus be examined from the functional aspect because collocations can be regarded as form-function composites. Kenny (1998) outlined a methodology to identify “TL [the target language] collocations on a more empirical footing” (p. 3) and suggested that corpora could be incorporated into the investigations of collocations to obtain more convincing evidence. In addition, Kenny (1998) focused on discourse prosody (also known as semantic prosody), and has found that translational language is a “somewhat ‘sanitized’ version of the original” (p. 515). In this sense, some collocations which may bear positive or negative meanings are neutralised in translational language. Therefore, as mentioned in section 2.2.3, this corpus-driven study will look at the functional features of collocations in terms of semantic prosody, therefore, it will use semantic prosody as the indicator to investigate normalisation by comparing a corpus made up with translational English and a corpus of native-speaker English, and will examine the collocation distribution patterns across the two corpora so as to obtain evidence to support the hypothesis of normalisation in translational English. Specifically, this study will instantiate these functional features through examining the proportion of positive semantic prosody, neutral prosody and negative prosody in collocations (see 6.4). Furthermore, this study will provide a

qualitative analysis of normalisation with examples found from the two corpora with regard to L1-L2 contrast (see 7.2.3).

The three translation universals, namely, explicitation, simplification and normalisation, are the most distinguishable features in translational language, so this study will mainly look at these three features to investigate translators' use of L2 collocations within its theoretical framework. In other words, this study will not provide a comprehensive review of all translation universals. It should be noted, however, that even though these features are distinguished from each other, they are also associated in some way. This means that when translated texts are simplified by translators, they might also be normalised or made explicit at the same time. I have attempted to illustrate the relationship between these universal features in Figure 2.3.

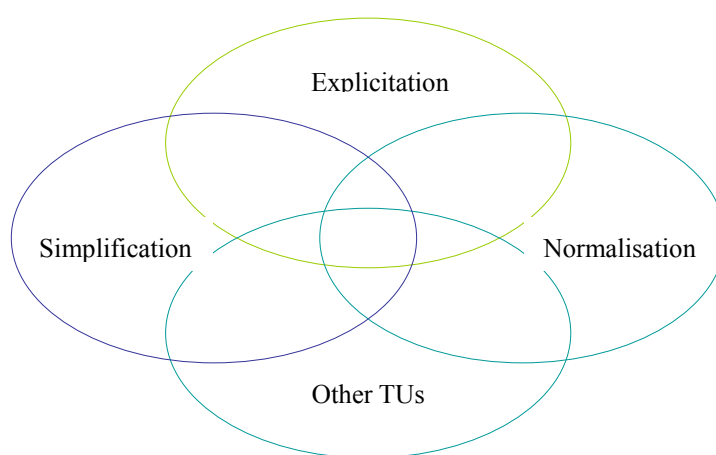


Figure 2.3 Relationship between translation universals

This study will adopt a Contrastive Interlanguage Analysis (CIA) research method (see 4.3.2) to compare the target corpus and the reference corpus, which indicates that these two corpora should be monolingual and comparable. A number of researchers (e.g. Olohan & Baker, 2000; Baker, 2004) employed this research method to seek evidence from comparable corpora and have shown the importance of this research method. The

translated texts and the native texts in comparable corpora might not necessarily be identical in every sense, but they should be equalised in terms of size, genre, register and so forth. Therefore, this study used a rigorous procedure of selecting language materials for the compilation of the corpora, which will be further detailed in Section 4.4. In addition, it should be noted that the qualitative analysis section (Chapter 7) may look at some typical examples of translation universals in the English translations by tracing back to the corresponding Chinese source texts for comparison purposes. However, this does not mean this study will employ a bilingual contrastive approach to assess Chinese translators' use of L2 English collocations. In addition, this study will not take account of the level of 'shining through' (see for instance Xiao, 2010), that is, how and to what extent the universal features of translated texts are transferred from the source language to the target language. Therefore, there was no need to build up a bilingual parallel corpus in this study.

2.4.2 Collocations' association and translation universals

As set out in Section 2.2 and Section 2.3, it appears that the linguistic features in the use of collocations are strongly associated with the indicators of translation universals. Thus it is possible and necessary to outline different levels of linguistic analysis in an attempt to clarify how translation universals should be examined in this collocation study and how this study should be carried out. Zanettin (2013) provided an interpretative model and distinguished four tiers of abstraction with regard to the examination of theories/hypotheses in translation studies with a corpus approach, which can be specified as follows:

- Tier 1 is the tier of theory, in this case the general hypothesis that, as a result of the process of translation, all translated or interpreted texts share certain properties which distinguish them from similar non-translated texts;
- Tier 2 concerns the descriptive features which support the theory;
- Tier 3 is represented by the linguistic indicators which realize a certain feature as concerns different levels of linguistic analysis;
- Tier 4 involves the computational implementation of these indicators, that is the way abstract linguistic features are instantiated through ... computational

operators (p. 21).

This model underpins the rationale of the present study. Therefore, this study is basically in line with this distinction regarding the levels of linguistic analysis and will lay out its research procedure accordingly. On the first tier, this study will demonstrate the rationales of researching collocations in a learner corpus (see 3.2) and construct a theoretical framework (see 3.4) to illustrate that collocations play an important part during the process of translation and that the inappropriate use of L2 collocations may result in translators bringing translation universals into the target texts.

In regard to the second tier, this study has shown that it will look at translation universals in terms of simplification, explicitation and normalisation. This study has also proposed the linguistic indicators for investigating these descriptive features, and will calculate the instances of collocation use in the authentic texts. This will be carried out by employing both a corpus-driven approach (see 4.3.1) and Contrastive Interlanguage Analysis (see 4.3.2) in an attempt to obtain more empirical evidence as a support for the theoretical framework of this study. On the third tier, as mentioned, translation universals can be analysed from formal, semantic and functional perspectives which correspond to the distinctive features of collocations, that is, collocations as formal co-occurrences, extended semantic units and form-function composites. Specifically, this study will employ degree of collocability, delexicalization and semantic prosody as the linguistic indicators of translation universals.

As mentioned previously regarding the fourth tier, this study will carry out quantitative analyses to examine these linguistic indicators of translation universals from four computational operators: the type-token ratio (see 5.2); the proportions of free combinations, bound collocations and idioms (see 6.2); the proportions of collocations with a literal sense and those with a delexicalized sense (see 6.3); and the proportions of collocations with a positive semantic prosody, those with a neutral semantic prosody and those with a negative semantic prosody (see 6.4). In this way, translation universals can be examined in more detail through different collocation distribution patterns found

by comparing a corpus made up with translated English from Chinese and a corpus compiled with native-speaker English.

Because collocations in language use are expected to be largely phrasal (cf. Sinclair, 1991; Wray, 2002), language users need to recognise and use collocations as lexical wholes rather than dividing them further into smaller units. In this respect, it would appear that over-use of free combinations, or under-use of bound collocations or idioms, in translators' L2 would make the TT 'simpler' than the same native-speaker TL, which may result in simplification in their translations. Therefore, the proportions of the three kinds of collocations are employed as computational operators to investigate simplification in regard to collocation distribution from the formal perspective.

In regard to semantic features, it would appear that over-use of collocations with literal senses or under-use of collocations with delexicalized senses in translators' L2 may result in explicitation through comparing translational English with native-speaker English. Therefore, the proportions of collocations with literal senses and those with delexicalized senses are employed as computational operators to investigate explicitation in regard to collocation distribution from the semantic perspective.

Over-use of neutralised language in a particular translated text may indicate that this text to some extent includes over-use of typical features of the genres involved (cf. Xiao, 2010), such as over-use of collocations with neutral semantic prosodies, which therefore shows normalisation in the TT. In this respect, over-use of collocations with neutral prosodies, or under use of those with positive or negative semantic prosodies would make the TT read more 'normalised' than the same native-speaker TL through comparing translational language with native-speaker language. I employed the proportions of collocations with different semantic prosodies as computational operators to investigate normalisation in regard to collocation distribution from the functional perspective. Thus, the intrinsic links between the features of the use of collocations and the indicators and computational operators of TUs in the research procedure can be reflected in Table 2.2 and Table 2.3:

Table 2.2 Translation universals and the linguistic indicators in the present collocation study

Translation universals	Linguistic indicators		
	Formal	Semantic	Functional
Simplification	Degree of collocability		
Explicitation	Delexicalization		
Normalisation	Semantic prosody		

Table 2.3 Translation universals and computational operators in the present collocation study

Translation universals	Computational operators		
	Formal	Semantic	Functional
Simplification	The type-token ratio; proportions of free combinations, bound collocations and idioms		
Explicitation	Proportions of literal sense and delexicalized sense in collocation use		
Normalisation	Proportions of positive semantic prosodies, neutral semantic prosodies and negative semantic prosodies in collocation use		

In addition, the explanatory section (Chapter 7) of this study will also provide examples where the inappropriate use of collocations may result in translation universals, and attempt to briefly suggest a model of the control mechanism between the features of collocations and translation universals.

2.5 Summary

The present chapter has provided a selective overview of previous collocation studies with a corpus approach and has attempted to explore the nature of collocations from the perspectives of form, meaning and function. This chapter has identified the features of collocations, i.e. that collocation can involve formal co-occurrences, extended units of meaning and form-function composites. This chapter has also examined the role of collocation in language learning and teaching, and explained the role in terms of language development and native fluency. Wray and Perkins's (2000) model of holistic and analytical involvement in language processing and Pawley and Syder's (1983) proposal of 'nativelike selection' and 'nativelike fluency' have shown that collocations play an significant part in facilitating language development and helping second language users achieve native-like language proficiency. The clarification of the role appears to be greatly helpful for establishing a theoretical model in an attempt to frame collocations more systematically, and it would be a noticeable mechanism that accounts for producing rich linguistic continuums. The present study involves Chinese-English bilingual translators, some of whom may be said to still find themselves somewhere along the L2 learner continuum and may come across various difficulties in using L2 English collocation in the commercial register. In this sense, it is necessary to make translators aware of such issues and emphasise the important role of collocation in translation.

In addition, this chapter briefly outlined the significant contribution that corpus approaches have made to the field of Translation Studies and may be able to make to translation practice. In practice, more and more translators prefer to align the existing translation materials with their computer-assisted terminology management system by accessing the corpora they need to facilitate their translation tasks. With regard to translation theories, a great number of researchers (e.g. Gellerstam, 1986, 2005; Baker, 1993, 1996, 2004; Laviosa, 1998a, 1998b, 2002; Maurenen, 2004; Xiao, 2010) who specialise in Translation Studies (TS) have noticed the importance of using corpus approaches to expound, test, exemplify and examine translation theories in that they are

able to obtain reliable linguistic evidence from authentic texts. This means that corpus linguistics has become a valid methodology for Translation Studies. This chapter also demonstrated that the investigation of translation universals has become the main strand of conducting Translation Studies with a corpus approach. Therefore, this chapter has emphasised the importance of researching translation universals and introduced three categories of translation universals which the present study will examine, namely simplification, explicitation and normalisation.

Furthermore, this chapter clarified the linguistic indicators and computational operators of identifying these three translation universals, and discussed how they are associated with the features of collocations. Specifically, translation universals can be found in terms of degree of collocability, delexicalization and semantic prosody, and all these indicators can be assessed by using computational operators in this frequency-based collocation study. Therefore, this chapter has created the intrinsic links, which will greatly help identify the role of collocation during the process of translation and help construct a theoretical framework for the present study.

On the whole, this study aims to formulate an appropriate approach to teaching non-English speaking background (NESB) translators how to identify appropriate collocations in Chinese to English translations in the commercial register for the purpose of acquiring these as part of their implicit language knowledge of their L2 English (Paradis, 2004; Robinson, 2003). This chapter has clarified some key notions with regard to the nature of collocation and brought up some important issues for discussion. I will revisit these key notions in my Discussion and Conclusion chapters. These points may be used to help demonstrate the significance of a memory-based mechanism of language acquisition, and will be discussed briefly in the explanatory section based on the findings from the quantitative research. The next chapter will look at L1-L2 contrast and employ that as the rationale underpinning this study, and outline a theoretical framework for the present study based on a range of relevant models (e.g. Ellis, 2001; Wray, 2002; PACTE, 2003; Pym, 2003; Robinson, 2003; Paradis, 2004) in linguistics and translation studies.

Chapter Three The preliminary study: Setting the stage

3.1 Introduction

This chapter will briefly review the rationale of researching collocations used in a second language (L2) and attempt to establish a theoretical framework. Section 3.2 will describe the rationale underpinning the present study by comparing the different models of learning collocations between native (L1) speakers and L2 learners. Based on the distinction between these two groups of learners, this chapter will also show potential difficulties that L2 learners, especially translators, may be confronted with when producing L2 collocations. Section 3.3 will distinguish translators from common L2 learners and specify the assessment of senior translator competence in China. This indicates that translators' difficulties in handling L2 collocations may result in producing translation universals which essentially distinguish translational language from native-speaker language. Section 3.5 relates to one of the core parts of the thesis, and will construct a theoretical framework for the present study, in which the interaction among collocation, translation units, translation universals and translators' potential knowledge in language operations will be discussed and analysed. Furthermore, this part indicates that the conceptual framework of collocation should be evaluated in terms of quantity, form, meaning and function so that it also paves a way for the ongoing quantitative and qualitative research.

3.2 Rationales for researching collocations in a learner corpus

Learning collocations is widely acknowledged to be one of most challenging fields associated with second language acquisition. However, the mechanism of collocation learning is not fully clarified yet. For researchers in this area, the detailed description of L2 collocation patterns from empirical evidence is the basis of providing constructive solutions to bridge the gap between L2 learners and native speakers. In this sense, identifying the patterns of L2 collocation use is usually central to the investigation of

the difference between L2 learners and native speakers in terms of language use. In other words, this type of research chiefly looks at how, and to what extent, L2 learners' use of collocations deviates from that of native speakers. The recent decades have seen a number of language researchers (e.g. Nattinger and DeCarrico, 1992; Skehan, 1998; Wray, 2000, 2002; Lewis, 2000; Ellis, 2001, 2002, 2003, 2005; Schmitt, 2004; Nesselhauf, 2005; Meunier & Granger, 2008; Barfield & Gyllstad, 2009) construct a range of theoretical frameworks regarding L1-L2 contrast in terms of language acquisition. Even though these natural language-based models vary at times due to different research aims and designs, they appear to reflect the same rationale, that is, the difference in learning and using collocations. Some researchers (e.g. Ellis, 2001, 2003) believe that collocation learning largely relies on the memory system and the chunking in formulaic sequences is the main factor developing the language acquisition process. According to Ellis (2001, 2003), this mechanism does not merely hold true for L1 acquisition but also works with L2 acquisition. This account is also in line with Nattinger and DeCarrico's (1992) claim that adults (L2) and children (L1) can develop language learning in the same way. Others (e.g. Wray, 2002) claim that L2 learners adopt a 'non-formulaic' approach, and tend to memorise and analyse individual words rather than combining ones as wholes. Therefore, L2 learners normally do not retain any information about the co-occurrence of words when they are exposed to their input. In respect to this discrepancy, Durrant and Schmitt (2010), carried out a lab-based study involving 84 participants (non-native speakers of English), and conducted their research based on three different training conditions of encountering L2 adjective-noun collocations, specifically, single exposure, verbatim repetition and varied repetition of collocations. It should be noted that verbatim repetition indicates that "the learner could engage with one piece of language a number of times over" whilst varied repetition means "the repeated use of a target collocation in different sentence contexts" (p. 172). These participants were instructed to read sample sentences (in which pre-designed adjective-noun collocations are tactically embedded) within a timeframe, normally varying from 3 to 7 seconds per sentence. Upon the completion of the training, these participants were required to undertake a test to examine their acquisition of those pre-designed L2 collocations. The results demonstrate that L2 learners can retain

information about co-occurring words to which they are exposed. Furthermore, based on their findings, Durrant and Schmitt (2010) proposed that the shortfall in L2 learners' collocation knowledge is "more likely to be the result of insufficient exposure to the language than of a fundamentally different approach to learning" (p. 182). Hence, this study will select some typical theoretical frameworks to discuss the rationale of learning collocations in more detail.

3.2.1 A model of L1 collocation learning

According to Ellis (2001, 2003), the process of native speakers learning collocations involves largely what is called the 'memory-based system', which is similar to Aitchison's (1987) argument that "humans start by using memory, and routine possibilities. If this proves inadequate, they turn to computation" (p. 14). This indicates that native speakers' instinctive production of language relies heavily on their memory rather than the grammatical rules which need more time in processing language data. This corresponds to Wray and Perkins's (2000) model which shows the holistic involvement in adult native speakers' language processing accounts for the larger proportion when compared with the analytical involvement. Hence, "working memory" (WM) (e.g. Ellis 2001) is the main factor to denote language learners' psychological mechanism to adopt a language in their memory system. According to Baddeley and Hitch (1974), working memory is primarily divided into three components, namely, the central executive, or supervisory attentional system (SAS), the phonological loop which comprises phonological store and articulatory rehearsal, and the visuo-spatial sketchpad. Based on Baddeley and Hitch's (1974) proposal, Ellis further explained that the model of working memory largely acknowledges "the intimate connections and mutual influences" (p. 35) of the three components it represents. This model also includes "different modalities of storage, separations between activated short-term and consolidated [long-term] representations, and the role of attentional process in learning and recollection" (p. 35), and can be illustrated in the following flowchart:

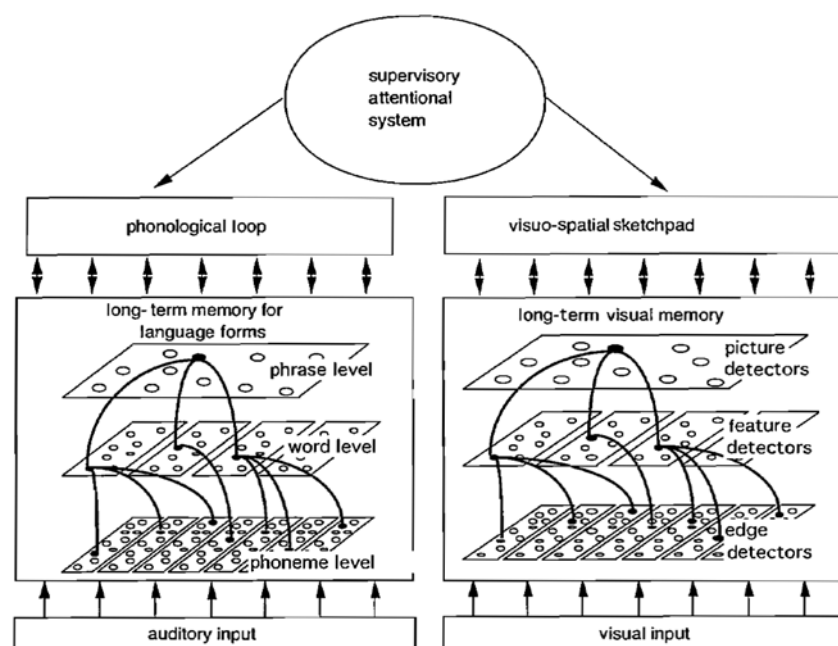


Figure 3.1 The model of Working Memory for language acquisition (Ellis, 2001, p. 36)

The nature of the Working Memory Model, in Ellis's own words, is that "we have specialist systems for perceiving and representing, both temporarily and in the long term, visual and auditory information, along with a limited resource attentional system" (p. 35). Learning collocations can thus be explained in this theoretical model. Frequently co-occurring lexical items in language operations are strongly associated with each other and may be stored in short-term memory in which they become largely phrasal. The whole process, either phonologically or visually, will inevitably lead to their becoming associated in long-term memory as stable linguistic representations. These units of memory organisations are then what are called chunks. This point can be based on James's (1890) "the Law of Contiguity": "[o]bjects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before" (James, 1890, p. 561, as cited in Ellis, 2001, p. 42). In addition, according to Ellis (2001), learning collocations is largely connected with an implicit process of chunk formation, which indicates that collocations are acquired beyond learners' awareness. Instead, learners' ability to use collocations is only reflected in their actual output after they are exposed to adequate knowledge during the process of chunk formation (Ellis, 2001). In other words, "chunking, the bringing together of a set of

already formed chunks (e.g. collocations) in memory and welding them together into a larger unit, is a basic associative learning process which can occur in all representational systems” (2001, p. 40).

This model covers the entire range of L1 learners’ collocation acquisition and, to a great extent, makes explicit the relationships between lexical units, lexical units and chunks, and even between chunks. In addition, Ellis’s (2001) model lends theoretical support for Sinclair’s (1991) ‘idiom principle’ which implies that language users possess a wide range of pre-constructed word combinations making up single choices, and that most texts can be interpreted by the idiom principle. As Ellis (2001) stated, language learning “involves learning sequences of words (frequent collocations, phrases, and idioms) as much as it does sequences within words” (p. 45-46), and “such collocations can simply be viewed as big words — the role of WM [working memory] in learning such structures is the same as for words” (p. 46). In this sense, collocations can be regarded as units calling for valid repetition to consolidate acquisition. It is also suggested that “[m]emory chunks (schema, scripts, fames, stereotypes, etc.) lie at the core of creativity in all domains of cognition” (p. 47) and identifying collocations needs to be interpreted from a complicated philosophical point of view. To be more specific, “identifying the smaller chunks and building up the larger ones” needs a hierarchical process by “repeated cycles of differentiation and integration” and the point about the idiom principle is that “maximally rapid intelligibility is afforded by the use of frequent, pre-existing chunks in the parole” (p. 47). In other words, the more language learners are exposed to and learn to use a particular collocation, the faster they will obtain the stable acquisition of this collocation and take it into their effective capacity of output.

Overall, the Working Memory model, to a great extent, explains how native speakers adopt collocations and how they “process” them into their linguistic competence. On one hand, the chunking of co-occurring words (collocations) undergoes a rather implicit process beyond L1 learners’ attention because L1 learners are not born with the competence of utilising collocations. Instead, they learn to use collocations through their social cognitive skills in the nurturing phase, in which they process complicated

cognition information into a kind of “linear” system provided by biological apparatus (e.g. Bates, Thal & Marchman, 1991). Therefore, the chunking mechanism enables learners to possess vast knowledge about how to transfer the possibilities of word sequences into their formation of linguistic competence. On the other hand, this process cannot merely result in “sequences of language which are potential labels” but also require the establishment of “cross-modal” associations which “typically occur between the highest level of activated node” (Ellis, 2001, p. 42), for which Ellis (2005) proposed a model of “conscious focus of attention” (p. 309). This model indicates that if implicit association formation is envisaged as a natural and heuristic process that would need a long time, the mechanism of consciousness is quite different because it can be triggered instantaneously. Consciousness combines the input of different cognitive modalities and helps learners clarify conceptual structures about the relationship between the phonological/morphological representations and the corresponding referents. Ellis’s (2005) model partially explains why knowledge of collocation is not completely determined by input frequencies (Ellis & Larsen-Freeman, 2006). Therefore, this model indicates that L1 collocation learning involves a complicated process, in which the chunking mechanism plays a vitally important part to help produce frequent formulae in language operations. In other words, L1 learners build up their ‘database’ of collocation gradually and progressively by memorising frequently occurring language pairs along with their development of consciousness in language use. The chunking mechanism and the consciousness mechanism are combined to enhance L1 learners’ competence in achieving native selection and fluency in their use of collocations.

3.2.2 L2 collocation learning

When discussing the Working Memory model regarding L1-L2 contrast, Ellis (2003) proposed that the mechanism of chunking might also hold true for L2 collocation learning. However, this point differs from the findings from Wray’s (2002) study. In addition, a large number of studies (e.g. Bahns & Eldaw, 1993; Farghal & Obeidat, 1995; Grander, 1998a, 1998b; Howarth, 1998a, 1998b; Foster, 2001; Sugiura, 2002; Wray,

2002; Feng, 2010) have shown that Ellis's argument regarding L2 collocation learning is not strong enough, and that the process of acquiring L2 collocations appears to be different from learning L1 collocations. Focusing on English, some researchers (e.g. Grander, 1998b; Howarth, 1998b; Foster, 2001; Sugiura, 2002; Wray, 2002) found that EFL learners tend to underuse collocations that are preferred by native speakers, such as *in the case of* and *on the part of* (Sugiura, 2002, p. 311).

Grander (1998b) and Howarth (1998b) investigated EFL learners' use of restricted collocations in the academic domain and they found that EFL learners have difficulties in identifying collocations so they tend to underuse them in their L2. Howarth (1998b) further noted that EFL learners only employ half of these restricted collocations which are frequently used by native speakers. Similarly, Wray (2002) discovered that L2 learners are able to adopt formulaic sequences, such as collocations, easily at an early stage, but "by the time the learner has achieved a reasonable command of the L2 lexicon and grammar, the formulaic sequences appear to be lagging behind" (p. 182). This is because, according to Wray (2002), "they [L2 learners] often have idiosyncratic grammar or vocabulary" and "[L2] learners cannot know them [collocations] unless they have actually encountered them [collocations] before... at a point in their learning when they have a chance of making sense of them" (p. 182). This indicates that L2 learners fail to identify collocations when they see or hear them. Instead, they are more inclined to notice individual words than memorise formulaic sequences as wholes.

This also echoes Irujo's (1986) proposal that "input without interaction is not sufficient for language acquisition" (p. 237). In other words, interaction in their L2 would enable L2 learners to be more actively involved in the language acquisition process. Nesselhauf (2005) investigated this issue in a more detailed frame which specifically includes verb-noun collocations, noun and prepositional phrases, and so forth. Based on the data retrieved from the German Corpus of Learner English (GeCLE), she found that nearly 25% of the English collocations produced by advanced German learners are wrong and more than 33% are deviant or questionable. The deviation has been "found to occur not only in the verb but also in other elements of the collocation (nouns, determiners, noun

complementation etc.) and in the use of collocations as wholes” (p. 237). Therefore, she suggests that collocations should be worthy of greater attention in language teaching than they have hitherto received. To obtain more evidence from the learners with an Asian L1 background, Sugiura (2002) compared Japanese L2 learners of English and native speakers of English in terms of collocation use. Based on a parallel corpus containing fixed expressions both generated by learners and native speakers, Sugiura also found that Japanese learners of English produced significantly fewer fixed expressions when compared with native speakers of English. Similar findings can also be found in Feng’s (2010) study. Feng (2010) focused on verb-adverb/prepositions collocations used by Chinese translators and discovered that even though Chinese translators, as advanced L2 learners (see 3.3), are able to pick up a considerable number of English collocations, they still find it difficult to master some complicated ones, such as phrasal-prepositional verbs in their L2 English.

Overall, this can be best summarised with Bahns and Eldaw’s (1993) conclusion that “EFL learners’ knowledge of general vocabulary far outstrips their knowledge of collocations” (p. 108). This also supports Kjellmer’s (1991) hypothesis that EFL learners normally produce their L2 from individual words rather than collocating words so their “building material is individual bricks rather than prefabricated sections” (p. 124). Wray (2002) exemplified this with the adjective-noun collocation *major catastrophe*, and claimed that most L1 learners would regard it as an idiomatic expression referring to ‘large disasters’ whereas L2 learners 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” (p. 209). If they happened to come across the situations where they were required to describe *major catastrophe*, “they would have no memory of *major catastrophe* as the pair originally encountered, and any pairing of words with the right meaning would seem equally possible” (2002, p. 209). Therefore, in respect to these findings, Ellis’s Working Memory model might not be completely suitable for explaining L2 collocation learning.

Given that L2 learners’ output of collocations differs significantly from that of native

speakers, the issue of how to model the patterns of L2 collocation learning would be theoretically important. Wray (2002) proposed that adult L2 learners basically employ a non-formulaic approach in language learning and they do not acquire phrasal chunks (e.g. collocations) the way native speakers do. Instead, L2 learners tend to divide chunks into individual words and memorise those words separately when they are exposed to their language input. This strongly echoes Kjellmer's (1991) hypothesis. According to Wray (2002), L1 collocations are “fully formulaic pairings which have become loosened”; in stark contrast, L2 collocations are “separate items which become paired” (p. 211). Therefore, L2 learners seldom establish a kind of “strength of association” (p. 211) while collocating words, even though they might have become aware of that. I have illustrated Wray's (2002) claim in Figure 3.2:

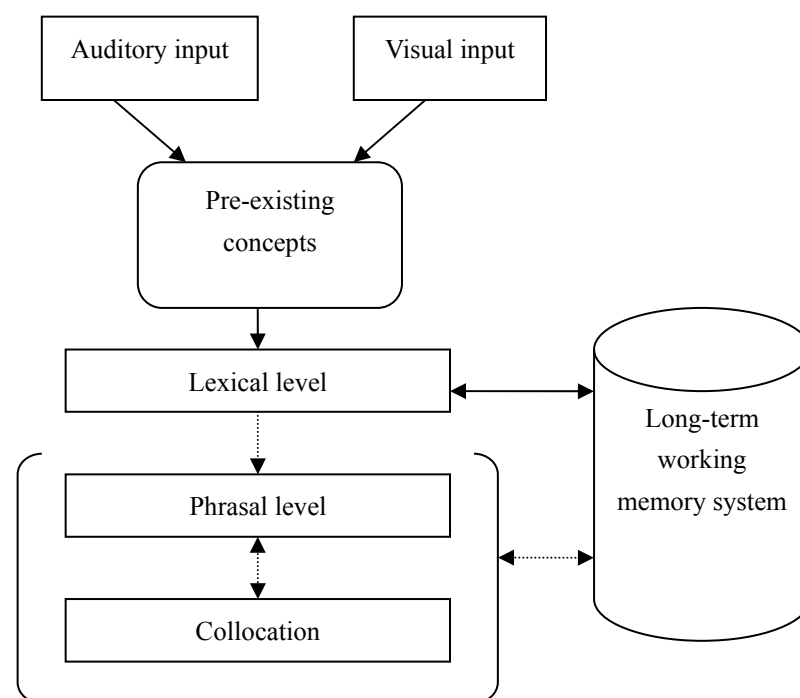


Figure 3.2 L2 model of collocation output

It is clear from this model that L2 learners with pre-existing knowledge are able to conceptualise what they have seen or heard, and transcribe those concepts into linguistic data on a lexical rather than phrasal level. L2 learners may wish to employ particular constructions as “pedagogical practices... tend to encourage learners to practise new constructions with a broad range of lexis” (Durrant, 2008, p. 51). Therefore, L2 learners may fail to produce prefabricated phrases and collocations in

their actual linguistic representations (indicated as a one-way dotted arrow between 'lexical level' and 'phrasal level' in Figure 3.2). Furthermore, they may store constructions with a large range of lexical forms, rather than formulaic sequences, into their long-term memory system. In this sense, the pre-existing concepts in L2 learners' cognitive system, to some extent, decrease their ability to use formulaic sequences by "weakening the imperative to communicate" (Durrant, 2008, p. 52).

In addition, the long-term working memory is directly associated with their linguistic output. This indicates that the memory system will not give any direct feedback at the phrasal level when L2 learners come across linguistic difficulties (indicated as a two-way dotted arrow between 'long-term working memory system' and 'collocation' as per Figure 3.2). As a result, they would tend to circumvent these difficulties in communication either by "adjusting their needs to avoid linguistically difficult situations, or by meeting their needs through non-linguistic means" (Durrant, 2008, p. 52). In other words, they know what they need to say, but that is just beyond their ability of producing linguistic formulae. Unlike L1 learners who are capable of dealing with the linguistic difficulties by using formulaic sequences, L2 learners might have to seek other means to meet their needs in communication, such as breaking formulaic sequences into smaller linguistic units (indicated as a two-way content arrow between 'long-term working memory system' and 'lexical level' in Figure 3.2). This is an important factor contributing to their linguistic output still staying on the lexical level. As Wray puts it, "they [L2 learners] simply avoided the situations in which they might need utterances which they could not produce" (2002, p. 175). This point also backs her claim that "there seems to be a link between the use of formulaic sequences and a need and desire to interact, these two together contributing to the overall achievement of communicative competence" (2002, p. 175).

However, this theoretical model is not entirely exclusive. Wray (2002) claimed that L2 learners might have "some means of building up the store of nativelike formulaic sequences post hoc, probably by residing and fully interacting for some time in the L2 environment" (p. 210). It is obvious that Wray did not point out clearly the reason why

L2 learners are still able to produce native-like collocations to some extent. On one hand, the wording “post hoc” is obscure in this context because Wray does not clarify what prerequisites may facilitate the use of collocations among L2 learners. Instead, she simply gave a couple of seemingly acceptable instantiations by using the word “probably” which adds a level of uncertainty about her assertions. On the other hand, she did not mention the cases where some L2 learners may also be able to produce native-like collocations at times in an L2 environment where their L2 is regarded as the main language in communication and the ‘language pairing’ mechanism in this circumstance would need more explanation. In this respect, some researchers (e.g. Durrant, 2008; Durrant & Schmitt, 2010) carried out further studies and have clarified some important points.

In their lab-based study (2010), Durrant and Schmitt found that L2 learners of English can “retain information about what words appear together in the language to which they are exposed” (p. 182). This point is in line with the findings in their previous corpus-based study (2009) that advanced L2 learners of English use the collocations or formulaic sequences to which they are frequently exposed in their language input. Therefore, Durrant and Schmitt’s conclusion is in contrast with Wray’s (2002) model, but largely supports Ellis’s (2001) Working Memory model that indicates that learners, no matter L1 or L2, acquire their knowledge of collocations through a memory system in which the chunking mechanism can be consolidated. When explaining the shortfall in L2 output of collocations in comparison with L1 Durrant and Schmitt (2010) simply ascribed this difference to “[L2 learners’] insufficient exposure, alternatively inadequate input, to the language than of a fundamentally different approach to learning” (p. 182).

In addition, Durrant and Schmitt (2010) further proposed that “the fluency-oriented” repetition approach for teaching collocations is more suitable for an in-class environment. They suggest that, to help L2 learners to establish valid associations between words or lexical items, teachers should devise materials to repeat what learners have learnt previously within an appropriate timeframe. In terms of the strategies, they emphasised that “verbatim repetition has some advantage over varied repetition” (p. 181)

based on their findings that the effective size (a parameter in their study to assess the validity of repetition conditions) is 0.56 for verbatim repetition and 0.48 for varied repetition (p. 181).

Nevertheless, the results for the validity of these two training conditions are fairly close, for which they conclude that “both the verbatim and varied conditions ... appear to be effective means of establishing initial collocation memory traces, with verbatim repetition being slightly more effective” (p. 181). Furthermore, Durrant and Schmitt (2010) pointed out a very significant issue, that is, L2 learners’ awkward use of collocations results largely from their limited knowledge about collocations. Therefore, they suggest that it is necessary for L2 learners to be exposed to a second language, such as spending more time in an L2 environment, if they want to increase the fluency of using language formulas.

Overall, Durrant and Schmitt’s (2010) argument, to a great extent, has clarified some obscure points in Wray’s (2002) framework and consolidates Ellis’s (2001) claim. Their viewpoint indicates that L2 learners can also rely on memory and the chunking mechanism can also operate to produce native-like collocations if L2 learners are trained appropriately under ‘pre-designed’ controlling conditions of encountering collocations, such as single exposure, verbatim repetition and varied repetition. The present study is basically in line with this perspective and will discuss this in more detail in Section 6.3.2.

However, it should be noted that Durrant and Schmitt’s (2010) model is constructed based on an ideal learning environment. That is to say, this model fails to present some other factors that may interfere with L2 collocation learning. The most obvious factor that contributes to this point appears to be L1 transfer (or L1 interference), that is, L2 learners are, to some extent, influenced by their mother tongue when they use collocations in their L2. This point is particularly true when L2 learners are exposed to previously unknown collocations, or even unknown lexical items that constitute collocations. This can also be explained in terms of L2 learners’ pre-existing knowledge,

that is, when L2 learners tend to produce new collocations that they have never come across, their pre-existing knowledge will serve as a screening device to select collocational candidates more from their L1 than from L2. Then these candidates will be combined according to the conceptual association to form so-called ‘collocations’ which may, or may not, be acceptable in their L2. In this sense, this pattern largely deviates from the memory-based chunking mechanism. This is the reason why Durrant and Schmitt limit their suggestions merely to the investigation of the “words that they [L2 learners] are already assumed to know” (2010, p. 181). This is also the reason why they did not talk too much about the learning pattern of unknown collocations, but simply mentioned that “[i]t is possible that somewhat different processes will be involved for collocations of previously unknown words” (2010, p. 181). This indicates that for translators who are still somewhere along the L2 learner continuum might come across situations, thus making translated texts share some properties which distinguish these translations from non-translated texts.

In respect to this, the present study will briefly suggest a pedagogical model to fill these theoretical gaps (see 6.3.2). In addition, it should also be noted that because this study concentrates on Chinese translators’ use of English collocations in the commercial register and translators’ production of L2 collocation patterns might not be the same as common L2 learners, those aforementioned models might not be completely suitable for explaining this particular case. Therefore, it is necessary to distinguish translators from L2 learners at this stage before it attempts to outline a theoretical model to clarify the role of collocation in translation.

3.3 Distinction between translators and L2 learners

When translators are working with the target language which is not their mother tongue, they need to produce their second language based on their linguistic competence and translation strategies to fulfil the translation tasks. Although translators’ L2 competence may, at times, come very close to the competence they have in their native language,

they might still come across linguistic problems as L2 users/advanced L2 learners. Therefore, even very fluent bilingual translators are still L2 learners in some sense. Nevertheless, this does not mean that translators are ordinary L2 learners. One of the most important factors that contribute to the distinction between translators and L2 learners lies in the fact that translators have translation competence while L2 learners do not. In respect to this, the PACTE research group (2003) holds that translation competence or translation knowledge is “qualitatively different from bilingual competence” and is “expert knowledge in which procedural knowledge is predominant” (PACTE, 2003, p. 60) over declarative knowledge (see 3.3.1). This indicates that translators are essentially a special group of (advanced) L2 learners who are supposed to master not only bilingual skills in translation tasks but also procedural knowledge of translation, which specifically includes their expertise in a particular register, their strategies in translations, their ability to use translation software tools and so forth. To put it another way, if translators are routinely working in one particular register, such as the commercial area, they should be reading in this area in their L2, and thus become increasingly exposed to correct L2 collocations and the use of these collocations much more than ordinary L2 learners.

Nevertheless, in practice, individual differences (see Ellis 1994, p. 472) play a major role in (second) language acquisition, and near-native ability in the use of collocations seems to only be achieved by those who are intrinsically motivated and talented, where others’ L2 skills may plateau. In this respect, all other aspects being equal, the more experienced the translator the more familiar they should be with correct use of collocations. In the present study, the reason I have chosen the collocation use in a particular register (commerce) as the target of research is that translators’ translation competence is built up with years of experience, in which they gain their expertise in a particular area. In other words, those who specialise in commercial translation might not be familiar with medical translation for instance. This means in collocation studies that translators might employ different collocations across different registers and, as mentioned in Section 1.2, different registers may demonstrate different collocation patterns. Therefore, it is more meaningful to observe the features of collocation

distribution in a particular register rather than a general sense when researchers attempt to explore translators' use of collocations. In contrast, L2 learners' use of collocations is usually investigated from general language use though some studies (e.g. Xiao, 2010) look at the linguistic features regarding different genres in an L2 corpus. This is another factor that contributes to the distinction between translators and L2 learners.

A number of attempts and proposals have already been made in Translation Studies (e.g. Bell, 1991; Nord, 1992; Pym, 1992; Hansen, 1997, Hatim & Mason, 1997; Campbell, 1991, 1998; Neubert, 2000) and a few researchers have built up robust frameworks (e.g. PACTE, 2003; Pym, 2003) directly pertinent to the discussion of translation competence, which greatly help clarify the distinction between translators and ordinary L2 learners.

3.3.1 PACTE Group's model

Translation competence refers to the translation knowledge that translators need to possess to meet the needs of translation tasks. In an empirical-experimental study (2003) conducted by PACTE (Procés d'Adquisició de la Competència Traductora i Avaluació, which means "Process of the acquisition of translation competence and evaluation"; it is a research group of translation and interpreting practitioners at the Universitat Autònoma de Barcelona and a member of Grup de Recerca en Competències of the Universitat Politècnica de Catalunya), translation competence is examined as both a process and a product. This means that translators' translation competence is essentially composed of both procedural knowledge and declarative knowledge as stated previously. It should be noted here that, according to PACTE (2003), procedural knowledge refers to the knowledge acquired and built up intrinsically by translators (Robinson, 1997) through translation practice regarding how to utilise the resources or facilities effectively to ensure the quality and accuracy of translation. Procedural knowledge is "difficult to verbalise" and "mainly automatic" (p. 45). In contrast with procedural knowledge, declarative knowledge refers to the knowledge that concerns what translators need to utilise or employ to facilitate their translation work. It is "easily

verbalised” and normally “acquired by being exposed to information” (p. 45). Furthermore, according to the PACTE model, translation competence is made up of five categories of sub-competence, namely bilingual sub-competence, extra-lingual sub-competence, strategic sub-competence, instrumental sub-competence and knowledge about translation sub-competence, which, as a whole, activates a series of psycho-physiological mechanisms (pp. 57-59). The interrelationships among those categories of sub-competence can be illustrated in Figure 3.3:

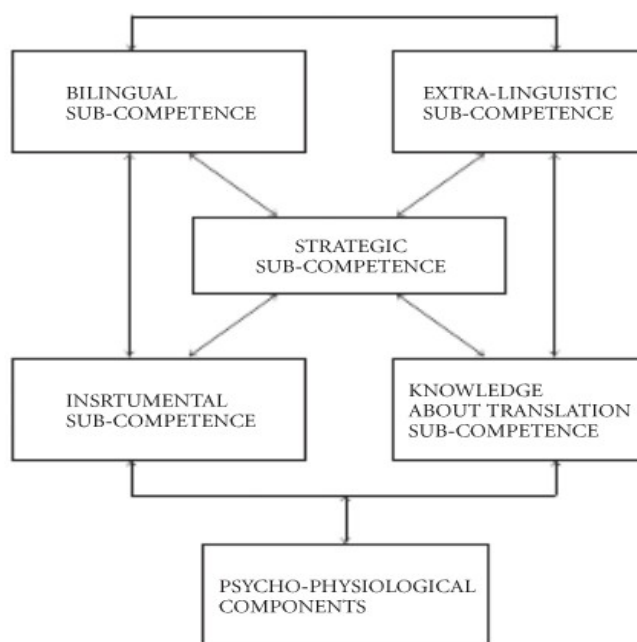


Figure 3.3 Model of Translation competence (Reprinted from PACTE, 2003, p. 60)

PACTE’s (2003) proposal regarding translator competence is essentially a model that entails a multi-component competence and involves a number of translation skills. In this model, bilingual sub-competence is “procedural knowledge needed to communicate in two languages”, which specifically includes “pragmatic, socio-linguistic, textual, grammatical and lexical knowledge”; extra-linguistic sub-competence concerns “declarative knowledge...about the world in general and special areas”; knowledge about translation sub-competence mainly involves the knowledge about functions of translation and translation professionalism; instrumental sub-competence is “procedural knowledge related to the use of documentation sources and information and communication technologies applied to translation”; strategic sub-competence plays a

crucially important role in the whole model because “it controls the translation process”, and it “guarantee[s] the efficiency of the translation process and solve[s] the problems encountered” (p. 59). Those five categories of sub-competence collaborate to trigger “different types of cognitive and attitudinal components and psycho-mentor mechanisms”, such as memory, attention, intellectual curiosity, rigour, creativity, logical reasoning and so forth (p. 59).

Therefore, it is obvious from PACTE’s model that these five categories of sub-competences are interrelated. In particular, this model indicates that translation competence is substantially distinguished from bilingual competence, and bilingual competence is merely considered to be a sub-component of the holistic competence system for translation. Therefore, when translators are carrying out translation tasks, all those competence mechanisms are actually in operations to reflect ‘procedural knowledge’, not merely bilingual competence. Furthermore, translators’ bilingual competence, as illustrated in this model, is interrelated with other categories of competence. In other words, translators’ bilingual competence results from mutual influence within this ‘procedural knowledge’, so it is substantially different from L2 learners’ bilingual competence. This is one of the key points indicating that translators’ linguistic knowledge or linguistic competence in translation is essentially distinguished from that of other L2 learners. In this sense, translators can not be seen as ordinary L2 learners because they possess metalinguistic skills.

3.3.2 Pym’s model

When discussing translator competence, Pym (2003) disagrees with the viewpoint of multi-component competence and argued that “the multicomponential expansions of competence are partly grounded in institutional interests and are conceptually flawed in that they will always be one or two steps behind market demands” (p. 481). Pym (2003) clearly stated a number of shortcomings of multi-component models. First of all, the number of the subcomponents in those models is not definite. That is to say, it is hard to

figure out how many components would constitute the entire system of translation competence. Secondly, the definitions in multi-component models fail to describe learning processes (see also Toury, 1995, p. 238) so they are incomplete and largely referred to as “ideal competence” (Pym, 2003, p. 487). Thirdly, those models might need empirical evidence to assess the validity. In addition, Pym (2003) also argued that multicomponentiality “operates as a political defence of a certain model of translator training” (p. 487) because it concerns mostly theoretical requirements but largely ignores what the translation market calls for.

According to Pym (2003), there have been big changes in the market demands, ranging from the publication of up-to-date translation information (e.g. the creation of localised translation websites) to the development of advanced translation technology (e.g. the utilisation of updated translation software tools). In respect to this, Pym (2003) generalised that “the multicomponential expansions of competence are partly grounded in institutional interests and are conceptually flawed in that they will always be one or two steps behind market demands” (p. 481). Therefore, Pym (2003) proposed a minimalist approach, in which he elaborates a two-fold “minimalist definition” regarding translation competence: a. “[t]he ability to generate a series of more than one viable target text (TT₁, TT₂, ... TT_n) for a pertinent source text (ST)”; and b. “[t]he ability to select only one viable TT from this series, quickly and with justified confidence” (p. 489). In this definition, Pym believes that translators’ competence should cover a process of generating a wide range of potential translations and selecting the most appropriate one. In this sense, translators’ translation practice is, in his own words, neither purely “linguistic” nor “solely commercial”, but just a process of “generation and selection” which mostly “occurs with apparent automation” (p. 489).

Pym’s model indicates that translators are essentially distinguished from common L2 language learners. On the one hand, translators and L2 learners follow different work procedures while using their L2. According to Pym (2003), translators are required to produce more variety of linguistic representation in their L2 for the source text than common L2 learners to meet the needs of translation. That is to say, translators possess

the competence to optimise their L2 production from variable target texts and choose the best one out of the potential candidates. In this sense, translators' production of L2 is a multi-directional process which substantially involves wording, re-wording, verbal addition and deletion on the language level, and checking, editing, obtaining feedback and so forth on the procedural level. This kind of multi-directional process is largely determined by the purpose of translation. Contrary to this, ordinary L2 learners' production of language appears to be a one-way linear process. In other words, they are generating 'one-off' linguistic representation. Even though L2 learners might need to check and edit their L2 production at times, they do not need to undergo a complicated procedure as translators do, such as referring back to the source language and waiting for the feedback from clients. On the other hand, translators and ordinary L2 learners use their L2 differently to achieve different aims. Translators' production of the target language (L2 in this study) involves a complicated generation-and-selection process of transferring information from the source language addresser to the target language addressee, with translators being both text decoders and encoders. In contrast, L2 learners only produce their second language to accomplish the multiple communication purposes, with their role being merely language encoder.

Pym's minimalist model indicates that translators' production of the target language (L2) is essentially different from that of common L2 learners. In collocation studies, when investigating translators' patterns of using L2 collocations in translational language, researchers should not only take account of linguistic features but also associate these linguistic features with the properties commonly shared in translational language and attempt to explain the reasons hiding behind them. The interference of the source language is just a case in point. When translators come across the situation in the source language which might cause them to produce 'marked' expressions in the target language, they will always be able to correct these 'marked' expressions after referring back to the source text and re-checking the target text. For instance, when Chinese translators read *chi1yao4* (*eat medicine, the Arabic numbers indicate the phonetic values of intonations in Mandarin Chinese) and *kan4bao4* (*watch/see newspaper) in Chinese, they are already 'directed away' from 'take medicine' and 'read newspaper' in

English, which is something they would usually correct to avoid this kind of ‘markedness’ and select the most appropriate candidate out of the viable target texts generated. Contrary to this, common L2 learners will also experience L1 interference, but more indirectly, in that they do not have the L1 collocation right there in the source text, thus influencing possible choices in language production. In other words, they would probably choose **eat medicine* instead of *take medicine*. Therefore, translators’ deviation in using L2 collocations may not reflect the same features as that of L2 learners. Collocations in translational language may demonstrate some inherent features that are quite different from both source language and target language (Xiao, 2010). For instance, according to some studies (e.g. Baker, 1993, 1996; Mauranen, 2007; Xiao, 2010), translational language appears to be made explicit, simplified and normalised (see 2.4.1) to some extent.

On the whole, despite the different models employed by different researchers as mentioned above, the common ground is that translators are distinguished from ordinary L2 learners from a number of aspects. Therefore, in this collocation study, it is necessary to construct a theoretical framework to identify the role of collocation in translation and clarify how translators can achieve native-like selection and fluency in the target texts when using collocations in their L2 under ideal working conditions, where translators have time to receive immediate feedback on their translations/outputs.

3.3.3 Translators in this study

Because the present study attempts to examine senior Chinese translators’ use of L2 English collocations in the commercial register, it is necessary to briefly look at their professional qualification in this context. Senior translators, also known as senior translation proofreaders, are deemed to be at the same level as professorship by professional title at a translation workplace (e.g. the Ministry of Commerce, the Ministry of Foreign Affairs, Foreign Affairs Office, Translators Association of China, industrial and commercial enterprises and social communities) or a higher institution in

mainland China (Bai, 2014). In reference to the eligibility for the professional title as senior translator, translators should have built up rich experience in translation, proofreading or finalising translated texts with broad scientific and cultural knowledge (see Appendix A for the assessment of translators in China).

In this respect, Chinese translators involved in the present study should have already qualified as professional translators and have been named “senior translators” when they were given the tasks to translate the articles (e.g. the texts used to compile the TECCTC) to be published in official domains. In addition, they should also have shown their expertise and experience in translating commerce-related texts because these texts may involve a large variety of technical terms and jargon in commercial English. As for the present study, due to the large-scale language samples in the corpus of translational English, which amount to over 10,000 English articles which are translated by more than 5,000 senior translators, it would be very difficult and unnecessary to identify these translators individually. In addition, important commercial texts as employed by this study are translated using a very complicated procedure and are normally finished by teamwork. In most cases, senior translators draft the initial versions, and then have their translations cross-examined by their colleague and checked by specialist translators in regard to technical terms, and finally deliver the translations to proofreaders for double-checking prior to publication. Even though senior translators involved in this study remain anonymous their overall translation outcomes represent the highest level with regard to translation skills and strategies in commercial translation. Therefore, these translators’ individual variation regarding translation competence will not be considered in the current research.

3.4 Theoretical framework of the present study

This study has shown that translators are essentially distinguished from ordinary L2 learners, and that the deviation of translators’ use of their L2 collocations may reflect some universal features in translational language, or the so-called ‘translation

universals'. Therefore, it is necessary to establish some criteria to clarify the notion of collocation and construct a theoretical framework of examining L2 collocations in this study.

3.4.1 Operational definition of collocation in the present study

The nature of collocation from the literature review section can be summarised as such: firstly, collocation reflects the syntagmatic relationship of lexical co-occurrence which can be quantified statistically; secondly, a collocation is the recurrence of a string of prefabricated lexical items or a lexical sequence; thirdly, a collocation is a composite involving form, meaning and function; and finally, collocation indicates the conventional use of language which is influenced by register. In respect to these features, collocation in this study is defined as follows:

A collocation is a prefabricated, structurally coherent and semantically complete lexical combination consisting of at least two words, whose occurrence is more frequent than by chance in the commercial discourse and can show statistical significance quantitatively.

This definition indicates that, to qualify as a collocation, a word combination must be “prefabricated”, “structurally coherent” and “semantically complete” or semantically independent. To meet these three criteria, a collocation must also be conventional in native use of language, and the ingredients that constitute a collocation must be structurally fixed or semi-fixed to fulfil a complete syntactic unit and realise an independent meaning. To be more specific, any violation to these criteria may result in ill formation of a collocation, or even a marked collocation. For instance, substitution is not allowed as in the idiomatic collocation *a piece of cake* (e.g. it cannot be *a piece of cheesecake*).

In addition, this definition indicates that collocations can be examined with quantitative

methods regarding the syntagmatic relationship of lexical co-occurrence. In other words, to qualify as a collocation, a word combination must be able to demonstrate significance in the statistical tests, such as the Log-likelihood test. In this respect, this study establishes three further criteria for the data retrieval procedure to denote the general definition proposed previously:

- a. a collocation should consist of two words (lexical items);
- b. a collocation should demonstrate statistical significance in collocability even if it is an entry defined in a dictionary (e.g. *budget deficit*);
- c. a collocation should allow no more than four continuing intervening words for the continuity purposes within a word combination (e.g. *his* in *conduct his research* thus serves as an intervening word in the collocation of *conduct* and *research*).

Furthermore, as noted previously, this study essentially attempts to discover the features of Chinese translators' use of English (L2) collocations in the commercial register. Therefore, this study will primarily look at the specialised collocations in this register. This study will also examine some relevant general collocations showing statistical significance in the commercial register because it is hard to make a clear distinction between specialised and general collocations. Overall, identifying collocations in a particular register is not only theoretically important but also methodologically motivated.

3.4.2 Mapping a theoretical framework of collocation in translation

This study has clarified the nature of collocation and shown the rationale of researching L2 collocation. Following this, it is important to incorporate these previously mentioned points into a theoretical framework. As with introduction of the role of collocation (see 2.3), the intended framework will demonstrate that collocation essentially plays a pivotal role in facilitating the natural rendition of the target text and consolidating L2

knowledge at both implicit and explicit stages. This is because the production of L2 collocations in the whole process of translation is strongly associated with the conceptual knowledge of the target language and the actual renderings (translations) of the target language. In addition, L2 collocation production is strongly influenced by register and is largely determined by the way translators acquire L2 collocations. Furthermore, it is worth noticing that the control of L2 collocation may serve as a determinant factor resulting in translation universals (see 2.4.1). Therefore, to obtain a clear picture of the complicated relationship among the factors around collocation, I have illustrated this framework in Figure 3.4:

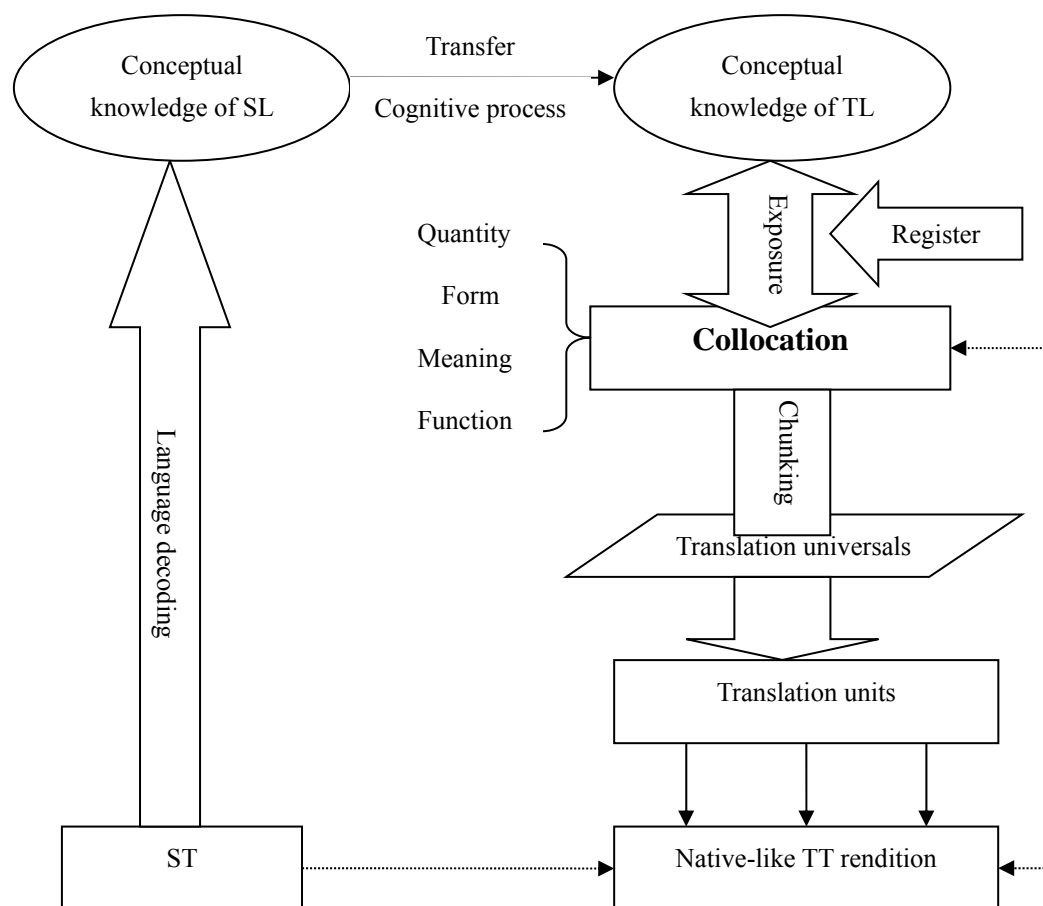


Figure 3.4 Collocation in translation process

Note: This model applies to human translation (HT) and machine-aided human translation (MAHT) only. Human-aided machine translation (HAMT) and machine translation (MT) may require another model.

This model demonstrates clearly that transferring a source text to a native-like target

text requires a rather complicated process. Translators start with understanding the source text and transforming the language information into their conceptual knowledge system of the source language, all of which can be regarded as a language decoding process. Then translators transfer the decoded information into their conceptual knowledge system of the target language through a cognitive process. This process involves an interface between the source and target cultures, as well as translators' knowledge of context, experience based on previous translations, etc. and together these largely determine translators' production of the target text. Finally, translators employ collocation as the strategy to transform the decoded information into linguistic representations, all of which can be regarded as a language encoding process. In this process, collocation serves as a core part when translators intend to use a large variety of appropriate lexical chunks in the hopes of producing native-like target texts. Therefore, the functions of collocation can be analysed from the following aspects.

First of all, in this model collocation is directly associated with the conceptual knowledge of the target language. As mentioned in Section 2.3.2, language knowledge is basically composed of implicit knowledge and explicit knowledge (Paradis, 2004). When translators are exposed to L2 collocations they will accumulate their knowledge to some extent in their 'database' system, implicitly or explicitly. This indicates that the more they are exposed to the target language, the more they will consolidate their target language knowledge. In return, translators' implicit knowledge and explicit knowledge jointly interact with their control of collocations when they render the target text. In this respect, the implicit and explicit proportions in translators' knowledge system directly determine the naturalness of the L2 collocations they produce.

If explicit knowledge is greater than implicit knowledge, translators will be more inclined to produce their L2 collocations consciously and think about what they are doing while handling translation tasks, which may result in deviations from the use of native-like collocations. Contrary to this, if implicit knowledge exceeds explicit knowledge in the system, translators will be more inclined to produce L2 collocations without awareness and consolidate their implicit knowledge of the target language,

which will make them ‘closer’ to native speakers in terms of collocation use. The kind of ‘dialectic relationship’ between the two types of knowledge in motivating the use of L2 collocations can be illustrated in the following schema:

$$\frac{\text{Implicit Knowledge}}{\text{Explicit Knowledge}} = \text{Native-like Collocation}$$

In this sense, to achieve native-like use of collocations, translators will need to enlarge their implicit knowledge or transfer their explicit knowledge to implicit knowledge as much as possible. Only by doing this can they realise, in their translation work, ‘how to do’ rather than ‘what to do’.

Secondly, it is obvious in this model that the use of collocations is influenced by register. As Halliday (1978) stated, “the language we speak or write varies according to the type of situation” (p. 32), and this kind of situation can be regarded as ‘register’. This indicates that the use of collocations is register-specific and that a high-frequency collocation in a particular register may not be necessarily frequent in another register or in a general sense. For instance, the phrase *cash flow* can be seen as a common collocation in commercial language, but it is rarely used in the medical area. Similarly, the frequently used nominal collocation *Coronary Artery Disease* in the medical area is seldom seen in business English. In this sense, researchers need to take note of register while examining collocations because different registers may demonstrate different collocation distribution patterns. Researching specialised collocations in a particular register would bring up more insightful theoretical achievements when compared with a study of general collocations. Only in such a way can language researchers working in this area observe language behaviours in more detail and provide more valid and precise descriptions about the relationships between lexical items.

Thirdly, this model shows clearly that collocation serves as a determining factor in producing translation units and reducing translation universals. A translation unit (see for instance Baker, 2001) refers to a segment of text which translators regard as an

independent cognitive entity during the course of decoding the source language and encoding the target language. A translation unit is not necessarily consistent between the source text and the target text. In other words, a translation unit in the segmentation of the source language might not be the equivalent unit when encoding the target language, which thus implies a kind of ‘shift’ in translation. For instance, *jiaozi* in Chinese (a kind of traditional Chinese food) has no equivalence in English, so it is normally translated as *dumpling* or *stuffed dumpling* which merely reflects the way *jiaozi* is made but fails to present the connotations in its name in the source culture (symbolising good fortune, good health and so forth). In addition, a translation unit varies in length. It can be a word, a phrase, a lexical chunk, a clause, a sentence and even a paragraph. Therefore, when translators are producing translation units, collocations play an important part through the chunking mechanism. The accurate use of collocations will enable translators to ‘unite’ lexical items more smoothly to constitute chunks or formulaic sequences according to the conventions of the target language. In this sense, the larger these units, the more likely they are to achieve native-like linguistic representations and the greater chance there is to render the native-like target texts. For instance, if the collocation *income and expenditure circular flow* in business English (describing a simple reciprocal cash flow circulation between producers and consumers in economics) is regarded as one translation unit, it would greatly reduce the chances of some developing translators replacing the constituent elements of this lexical combination. Otherwise, if translators break this collocation into two translation units, such as *income and expenditure* and *circular flow*, it would increase the probability of replacing *circular flow* with *circulation* (the word combination *income and expenditure circulation* is not a significant collocation in business English). In this respect, appropriate formation/enlargement of translation units would also reduce the deviation from native-like linguistic expressions to a minimum, which is, to a large extent, determined by collocation.

In addition, this theoretical framework depicts translation universals as ‘screening factors’, which counteract the natural formation of translation units. As a result, this kind of ‘interference’ will result in some universal features in translational language,

such as simplification, explicitation and normalisation and would become obvious obstacles when translators attempt to produce native-like target texts. Therefore, I have created another schema to indicate the relationship among translation units, translation universals and native-like rendition of the target text:

$$\frac{\text{Translation Units}}{\text{Translation Universals}} = \text{Native-like Target Texts}$$

This schema indicates that the accurate use of L2 collocations can help translators bring together words to constitute reasonable high-frequency translation units, which are strong enough to break the constraints of translation universals and show naturalness in the target text. To a great extent, this will not merely help enhance translators' L2 proficiency but also facilitate their translation work. Contrary to this, the inaccurate use of L2 collocations would restrict translators from producing appropriate translation units and make them more inclined to present translational universal features in the target text. As a result, translators would be constrained by these features and they would find it very hard to produce native-speaker target language. In this respect, I propose that to render native-like target texts translators should continue to build up a strong database of translation units because new units will evolve all the time; to construct such a database, translators should recognise the importance of acquiring L2 collocations and applying them in their translation practice. This will be discussed in full in the Conclusion chapter.

Last but not least, this theoretical model demonstrates that collocations in language operations, especially translations, can be regarded as multi-dimensional composites and should be analysed from different angles. A number of studies (e.g. Bolinger, 1976; Fillmore, 1979; Pawley & Syder, 1983; Sinclair, 1991, 2004; Hunston & Francis, 2000; Stubbs, 2001; Wray, 2000, 2002) have already constructed conceptual frameworks to incorporate these perspectives. Based on such previous studies, I propose that the use of L2 collocations by translators can be assessed with a quantitative approach, so their L2 collocation patterns can be compared statistically with those by native speakers of that

language. I further suggest that the analyses of the features of L2 collocation patterns can be carried out with regard to form, meaning and pragmatic function (see also 2.2). These factors, in return, constitute a complete conceptual framework of collocation which indicates the research direction as to how collocations will be investigated with a frequency-based approach (see also 2.4.2).

In this conceptual framework of collocation, frequency indicates that researchers can investigate the quantitative features of translators' use of L2 collocations through the comparison of comparable corpora, where one of the corpora involves native-speaker language produced by native speakers and the other involves translational language produced by translators. This perspective looks at overall frequencies and collocation types, taking into account the general proportions of the collocation use by employing authentic language data from the comparable corpora. This perspective also reflects the 'dual mode' system which advocates that the preponderance of language use lies in "the open-choice principle" and "the idiom principle" (Sinclair, 1991, 2004), or that the alternation in language use reconciles between "holistic" and "analytical" systems (Wray & Perkins, 2000), because it attempts to uncover the mechanism of how collocations are learnt based on the comparison between L2 translators and native speakers.

In this sense, evaluating the conceptual framework of the present study from a quantitative perspective will shed some light on the assessment of the existing models of collocation learning. As set out in Section 3.2, there are different viewpoints with regard to L2 collocation learning. Some researchers, such as Ellis (2001), believe that both L1 and L2 learners can adopt a formulaic approach to learning collocations, and collocation learning relies heavily on language users' memory system, in which language data is processed into chunks or formulaic sequences, rather than individual words, to convey meanings. Other researchers, such as Wray (2002), argue that L2 learners (especially translators in this study) basically adopt a 'non-formulaic' approach, which is quite different from L1 learners' memory-based approach. L2 learners tend to ignore collocations when they see or hear them; rather, they are more inclined to

‘notice’ individual words than recognise formulaic sequences or memorise them as wholes. As a result, L2 learners can seldom ‘capture’ any information about the lexical co-occurrence and the intrinsic relationship between the words to which they are exposed, and they normally do not retain any information about collocations. With regard to this discrepancy, evaluating the theoretical framework with empirical evidence in this study can re-assess the validity of these two models (i.e. Ellis’ model and Wray’s model) through the comparison between comparable corpora. In addition, the evaluation from a quantitative perspective can provide researchers with an opportunity to observe translation universals in that statistical results can show how and to what extent translational language is made explicit, simplified and normalised regarding the use of collocations. Such statistical results would make researchers recognise the recurring problems that translators are confronted with, and would motivate researchers to elaborate reasonable pedagogical strategies to solve these problems in translator training. Furthermore, the theoretical knowledge of collocation learning through this conceptual framework would benefit translators when they try to master their L2 production and facilitate their translation work.

On the whole, this intended theoretical framework indicates that the conceptual framework of collocation should be constructed based on quantitative, formal, semantic and functional perspectives. It should be noted, nevertheless, that a proposal of investigating collocation from those perspectives does not mean that researchers in this area can focus merely on one particular perspective but overlook the others. Instead, collocation in language operations results from the interaction or some kind of ‘co-effort’ from those approaches. Collocations can be best described as “gestalts” in this sense (Lakoff, 1987, p. 538). Therefore, the establishment of this conceptual framework of collocation basically holds a holistic point of view in language research and implies that the methodology can be multi-dimensional.

3.5 Summary

This chapter has provided a brief overview of the rationale of researching L2 collocations, which indicates that L2 translators' use of collocations may be different from that of native speakers. Based on this research rationale, this chapter has attempted to define collocation in the commercial register and construct a theoretical framework to identify the role of collocation in translation in response to the first and second research questions. This theoretical framework indicates that, for translators, their inappropriate use of L2 collocations may result in some universal features in translation practice, specifically including explicitation, simplification and normalisation. These translation universals, to some extent, may become potential barriers for translators to achieve native-like selection and native-like fluency regarding the use of L2 collocations. Contrary to this, the appropriate use of L2 collocations will not merely determine the smooth formation of translation units which they employ in the natural rendition of the target language, but also help them reduce, or even avoid, translation universals in their translation practice. In this sense, knowing the mechanism of how collocations operate in language code switching, as well as how the properties of collocations can be recognised, will increase translators' bilingual skills and enable them to construct a consolidated knowledge system, in which they will find it much easier to produce language formulae. This will, in return, benefit them by increasing language proficiency and appropriate use, thus helping facilitate their future translations.

Furthermore, this theoretical framework indicates that collocation can be investigated from the quantitative, formal, semantic and functional perspectives. Therefore, based on this theoretical framework, the next chapter will state clearly the five research questions and outline the research methodology and the research method underpinning the present study in an attempt to clarify the procedure of evaluating the validity of this theoretical framework in the ongoing quantitative research section.

Chapter Four Research design

4.1 Introduction

This chapter will elaborate on a general procedure for retrieving collocations from the two designed corpora used in this study. Section 4.2 will outline the research methodology and research method used in the present study. Section 4.3 will introduce Contrastive Interlanguage Analysis and use that as underpinning for a corpus-driven Contrastive Interlanguage Analysis approach as a model in this study. Section 4.4 will introduce the two corpora (the NECCD and the TECCTC) in detail. Section 4.5 will look at the statistical measures for identifying significant collocations and will include a discussion of the Log-likelihood test and the Mutual Information score test which are the methods used in this study. In Section 4.6, I will explain the rationale for the design of three language filters in an attempt to help reduce inappropriate word combinations in the commercial register and increase the data retrieval accuracy.

4.2 Research methodology and research method

4.2.1 Research methodology: the corpus-driven approach

The present study is descriptive and employs the corpus-driven approach as the methodology. Corpus linguistics is seen by many researchers (e.g. McEnery & Wilson, 1996; Bowker & Pearson, 2002; Meyer, 2002; McEnery et al., 2006) as a methodology for language studies. In particular, as McEnery et al. (2006) note, “corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself” (p. 7f.). Language studies with a corpus-driven approach use large-scale corpus data to examine lexical items (or lexical combinations) and uncover the features of particular phraseological and/or grammatical patterns in which lexical items are distributed. Corpus-driven linguistics essentially adopts a “bottom-up” method (see e.g.

Gries, 2010), in which “bottom” means the data retrieved from authentic language materials whilst “up” means the generalisation of linguistic theories. In other words, corpus-driven studies undergo a process of extraction/retrieval, observation, generalisation and interpretation/explanation. To be more specific, the exploration of any linguistic issue with a corpus-driven approach starts from retrieving relevant language data (normally the target of research) from corpora, with which researchers need to process the retrieved data with appropriate tools and obtain the quantitative information required by their research. Then, researchers need to observe and describe the general characteristics and tendencies of the obtained quantitative information. Finally, they need to further examine the target of research (e.g. collocation) in terms of form, meaning and function and so forth, and explain the target of research appropriately. In this sense, language studies with a corpus-driven approach are descriptive. With regard to the features of this research process, corpus-driven linguistics is believed to be frequency-based and probability-driven (see e.g. Nesselhauf, 2005) in that frequency and probability may reflect some important innate properties in language operation or language use. Based on the high frequency of a particular language form (e.g. a word, collocation and lexical bundle), researchers can discover its morphological, semantic and functional features in the communication of a particular speech community, with which they can effectively uncover a great number of language properties pertinent to formulaicity, lexicalisation, grammatisation and so forth. Therefore, the corpus-driven approach, in nature, is inductive rather than deductive.

The corpus-driven approach, as a descriptive method, is chosen as the most appropriate methodology because the research design of the present study is in line with the research process of corpus-driven linguistics. The target of research in this study is collocation, for which I will compare the collocation distribution patterns in the corpus of native-speaker commercial English and the corpus of translational commercial English from Chinese. The research starts with collocation retrieval and collocation processing, in which statistical tests are used to measure the associative relationships between collocating words. The obtained quantitative information will ‘drive’ me to observe and describe how collocations are distributed and dispersed in language use

with regard to some linguistic indicators (e.g. collocability and semantic prosody). Then I will generalise the features of collocation distribution in the two corpora, and make a comparison in terms of form, meaning and function. Finally, I will explain the findings from the quantitative research within the proposed theoretical framework of this study. In this respect, the descriptive method appears to be appropriate for the research procedure outlined above. For all of these reasons, the corpus-driven approach is employed in this study.

It should be noted that, with regard to a rigid distinction of corpus linguistics, there are two main approaches normally employed in corpus linguistic studies, namely the corpus-based approach and the corpus-driven approach. Even though both of the two approaches emphasise the employment of large-scale collections of authentic language materials (e.g. corpora), the corpus-driven approach is essentially distinguished from the corpus-based approach with regard to “types of corpora used, attitudes towards existing theories and intuitions, and the focuses of research” (McEnery, Xiao & Tono, 2006, p. 8). In the corpus-based approach, corpora are believed to “expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study” (Tognini-Bonelli, 2001, p. 65). In contrast, the corpus-driven approach is thought to view “the integrity of data as a whole” (p. 84) and claim that “[t]he theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus” (p. 85). In this respect, the corpus-driven approach is more radical than the corpus-based approach, and deserves to be “a new paradigm within which a whole language can be described” (McEnery, Xiao & Tono, 2006, p. 11).

In addition, the corpus-driven approach is strongly associated with frequency and does not require corpora to be annotated (that is, to mark up corpora by encoding with additional information so as to specify the ‘values’ of constituents of corpora, such as the grammatical identity of tokens and the beginning of a sentence). In this sense, the corpus-driven approach “makes no distinction between lexis, syntax, pragmatics, semantics and discourse (because all of these are pre-corpus concepts and they combine to create meaning)” (McEnery, Xiao & Tono, 2006, p. 10). Furthermore, the

corpus-driven approach requires larger-scale corpora than the corpus-based approach in language studies and emphasises the full exploitation of corpus evidence. This indicates that researchers in this area can access sufficient language evidence and employ frequency to filter some data irrelevant or unimportant to their research. For instance, when researchers attempt to explore collocation patterns in translational English, they can set 5 occurrences as the minimum threshold of identifying a significant collocation in a particular corpus, so as to uncover the features of frequency distribution under the corpus-driven approach.

The present study aims to investigate Chinese translators' use of English collocations and explore the collocation distribution patterns produced by translators through comparing translational English and native English. In this respect, this study does not intend to expound, test or exemplify the collocation distribution patterns because these patterns are not yet formulated and will not be uncovered until they are fully examined with corpus evidence. This indicates that, from the operational perspective, this study will not hypothesise any particular collocation distribution pattern, but will build up a model which incorporates a conceptual system and computer programming to retrieve (and compute) collocations based on frequency and analyse collocation distributions with statistical measures. This is the reason why it employs a corpus-driven approach as the methodology. It should be noted, however, that even though a corpus-driven approach does not necessarily require a distinction between lexis, syntax, pragmatics, semantics and discourse, it may still, at times, allow quantitative and explanatory analyses from those perspectives.

4.2.2 Research method: the Contrastive Interlanguage Analysis

Essentially, this corpus-driven study will focus primarily on Chinese translators' use of English collocations in commercial translation and will investigate the variation to the existing collocation use in translated English. It also attempts to generalise the features of collocations used by Chinese translators and will identify some of the existing

problems in commercial translation. Hence, this study will employ the Contrastive Interlanguage Analysis (CIA) approach to provide a benchmark as to how translational English is different from native English in terms of collocation use. The aim for this approach is also to provide a solid foundation for the subsequent quantitative analysis and explanatory study regarding what translators are actually doing in translation practice.

Contrastive Interlanguage Analysis, in nature, is a method which combines the traditional Contrastive Analysis approach and the tools of corpus linguistics, aiming to identify unnatural expressions in learner language (specifically Chinese translators' English in this study). The term "interlanguage", as shown in the term CIA, was first introduced by Selinker (1972) when he attempted to uncover the difference of language use between L1 and L2 speakers to express the same meaning in a particular situation. It should be noted in this context that interlanguage basically refers to a linguistic system which is developed by an L2 learner, whose L2 is proficient in production but still influenced by his or her L1 to some extent. The impact resulting in unnaturalness in learners' L2 is constituted by a number of factors, such as L1 transfer (e.g. Nitschke, Kidd & Serratrice, 2010), L2 learning strategies (e.g. Cohen, 2011) and so forth. In this sense, CIA "concerns the varieties of the *same* language" (Hasselgård & Johansson, 2011, p. 38), specifically English in this thesis. CIA basically involves two aspects: the comparison of learner data with native speaker data and the comparison between different types of learner data. Given the characteristics of this project, the former type of comparison will be employed in this study.

The advantages of using the CIA approach are quite obvious in corpus linguistics, with which a large number of researchers have already achieved very insightful findings. For instance, some researchers (e.g. De Cock et al., 1998; Hasselgren, 1994; Ringbom, 1998) have found that the type-token ratio (TTR) in a learner corpus is lower than that in a native corpus, which means L2 learners normally produce less variety of vocabulary in their L2 output when compared to native speakers. Hasselgren (1994) notes that L2 learners are more inclined to use the words/lexical items which they "feel safe with" (p.

237) for fear that they would make a mistake, as a result, they tend to over-produce frequent words in a general sense instead of those equivalent synonyms with more precise meanings. This particularly happens to the translation of verbs and nouns. Similarly, De Cock et al. (1998) also observe that when compared to native speakers L2 learners are inclined to over-produce recurrent word combinations, with learners' high-frequency word combinations being largely different from those of native speakers. This point is echoed by Feng's (2010) investigation of 141 of the most frequently used English multi-word verbs (MWVs) in native and learner corpora, in which he found that in a million-word sampling Chinese learners produced *carry out* 733 times and *set up* 654 times, together accounting for 31.60% of all the occurring MWVs in the learner corpus. However, *carry out* occurs merely 150 times and *set up* does not even occur in top 10 most frequently used MWVs in the corpus of native English. In this sense, with the CIA approach researchers can conclude that L2 learners possess a 'narrower' range of vocabulary when using their L2. It should be noted here that this discovery is merely 'the tip of the iceberg', researchers with the CIA can always compare similar corpora from different perspectives to ponder other issues, such as to what extent L2 learners' written language is influenced by their spoken language, how L2 learners' collocation use is influenced by their L1 and so forth. This study concentrates on collocation and compares the difference between native and translational corpora, which appears to be a typical case of employing the CIA approach.

Establishing CIA models to explain the features of translational language is not unusual in previous linguistic research and Translation Studies. Granger (1996) and Gilquin's (2000/2001) Integrated Contrastive Model (ICM) is just a case in point. As Granger (1996) notes, "the model involves constant to-ing and fro-ing between CA [Contrastive Analysis] and CIA", in which "CA data helps analysts to formulate predictions about interlanguage which can be checked against CIA data" (p. 46). The inter-relationship between CA and CIA can be illustrated in the following diagram:

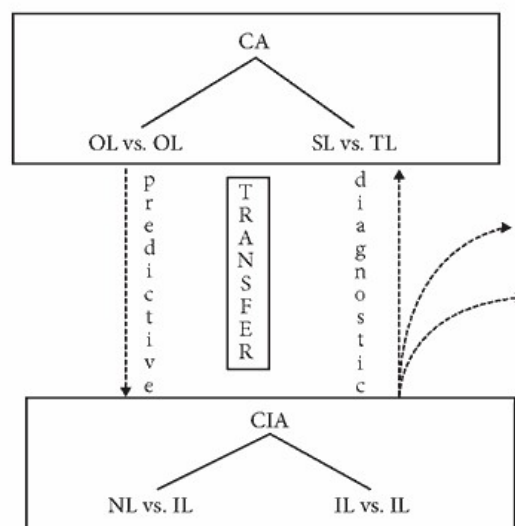


Figure 4.1 Integrated Contrastive Model (Gilquin, 2008, p. 8)

This model provides a new perspective to look at two types of language with accountability to two basic hypotheses, i.e. the predictive hypothesis (also called “CA a priori” or “strong CA hypothesis”), and the diagnostic hypothesis. In the predictive hypothesis, CA looks at the comparison either between original languages, or between the source language (SL) and the translated/translational language (TL). Based on this kind of comparison, researchers can predict L2 learners’ interlanguage, investigate their production with CIA and seek traces of L1 (transfer). Thus, researchers can “test the accuracy of the predictions and thus establish the (potential) presence, or otherwise, of transfer” (Gilquin, 2008, p. 7). According to Gilquin (2008), the rationale of this model appears to be very clear:

[I]n the case of discrepancies between the learners’ mother tongue and the target language, the learner is likely to transfer the L1 pattern to his/her interlanguage, hence producing an erroneous L2 pattern (negative transfer). In the case of similarities between L1 and L2, on the other hand, the learner is expected to produce a correct pattern in L2 (positive transfer) (p. 7).

The diagnostic hypothesis (also called “CA a posteriori” or “weak CA hypothesis”, Gilquin, 2008), however, undergoes the opposite process. CIA looks at the comparison either between the native language (NL) and interlanguage (IL), or between interlanguages (ILs i.e. “learner data” and “data produced by learners from other mother tongue [L1] backgrounds”, Gilquin, 2008, p. 7). This kind of comparison enables

researchers to “notice L1-specific errors and look to contrastive analysis for an explanation” (Gilquin, 2008, p. 7). In addition, this model also shows that “the explanation for an error will not always be found in the relation between the learner’s mother tongue and the target language” and the error “may be due to...L1 influence, development factors” (Gilquin, 2008, p. 7).

Generally speaking, this model is a two-way hypothesis system, in which learners’ problems of using their L2 can be clearly analysed through multi-dimensional comparisons. In particular, this model takes account of language transfer and attributes the difference between NL and IL, as well as between SL and TL, to language transfer. Thus, researchers are able to discover translators’ problems in L2 output, such as collocation, and explain these problems based on the contrastive analysis of native language and the target language within this model. In this sense, this model is very useful for this thesis because it specifies how to carry out a reliable comparison between the two designed corpora in this study, and from what aspect the difference should be interpreted. Therefore, in the methodology section this thesis is basically in line with the CIA approach of Granger (1996) and Gilquin’s (2000/2001) Integrated Contrastive Model. This thesis aims to summarise the features of English collocation patterns in L2 Chinese translators’ commercial translation and uncover how their L2 collocation use deviates from that of native speakers. To be more specific, this study compares native commercial English (NL) with translational commercial English produced by Chinese translators (IL), and from this comparison predicts that there are translation universals (e.g. Explicitation, simplification and normalisation, see 3.4) in translational language (TL) in terms of collocation use. Next it employs statistical measures to test and examine the difference “diagnostically” (Granger and Gilquin’s term) between native English and translational English. This will help with explaining whether translators’ production of collocation deviates from natural use, and if so, to what extent it deviates. This indicates that some Chinese translators, as shown in this study, are, to a greater or lesser extent, influenced by their native language (Chinese) when producing their L2 (English), thus leaving some universal features in their translation. In this respect, the L1 transfer in translators’ L2 (interlanguage) production (translational English) can

simply be explained through translation universals (TUs). This is also in line with the theoretical framework of this study, which can be visualised as per my proposed model:

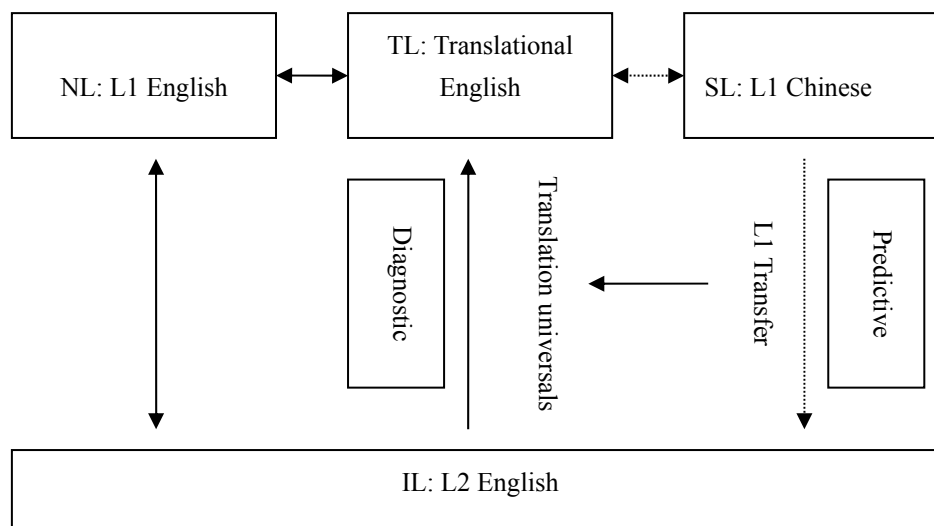


Figure 4.2 The CIA model in this study

This model shows that the comparison between the two designed corpora (NL vs. TL) in this study essentially indicates the comparison between native English and Chinese translators' L2 English (TL). This NL-TL comparison aims to uncover the distinctive features of Chinese translators' TL by investigating the collocations used in the translational corpus, which can help explain the unnaturalness in the translational English they produced. On the one hand, based on the NL-TL comparison (see 5.3.1 to 5.3.3), I can "predict" the potential unnatural use of English collocations in translators' IL, and analyse or speculate about whether this kind of unnaturalness results from L1-specific errors and, if so, how the traces of L1 can be sought (shown as a downward dotted arrow in Figure 4.2). The reason why I have used a dotted arrow in this situation is that the explanation for an error cannot always be found by comparing learners' native language and their L2 English. Nevertheless, it should be noted that this study will not include the comparison between SL and IL in the quantitative research section even though such a comparison may be carried out through strategies such as exemplification in the explanatory study. On the other hand, translators' TL which is

influenced by L1 transfer will have a direct impact on their production of L2 (English) collocations, thus bringing TUs into TL. In respect to this, I propose that the more Chinese translators' TL is transferred from their L1 (Chinese), the more translation universals they will present in their commercial translation; and that the larger this kind of L1 interference the less likely they will be to produce native-like English collocations and lexical chunks.

4.3 Corpora employed in this study: the NECCD and the TECCTC

The language data required in this study is mainly from two designed corpora, specifically, the Native English Corpus of Commercial Discourse (NECCD, the reference corpus) and the Translational English Corpus of Commercial Translation from Chinese (TECCTC, the target corpus). Because this study aims to investigate Chinese translators' use of English collocations in the commercial register, these two specialised corpora are designed to provide language data of commercial discourse. The data to be discussed and analysed in the study mainly comes from the TECCTC whilst the NECCD serves as a valid reference corpus for comparison purposes.

4.3.1 The Native English Corpus of Commercial Discourse

The NECCD, as the native English reference corpus in this study, was established by combining native-speaker English texts, aiming to construct a reliable data base of commercial discourse with authentic and up-to-date language materials. This corpus is designed to expose the characteristics of native English speakers' use of collocations in the commercial register. Therefore, it provides necessary resources for the ongoing comparison in terms of the use of English collocations. These texts were all chosen from a wide range of authoritative organisations, media and news providers, such as The International Chamber of Commerce, British Chamber of Commerce, United States Department of Commerce, New York Times, The Wall Street Journal, Reuters, Business Wire, Entrepreneur, Yahoo Finance, Stockopedia, Telegraph, Fox Business,

MarketWatch, Investopedia and so forth. In this sense, this method of selection also looks at business English within a broader geographical scope, that is, not merely British English nor purely American English. The language materials in the NECCD constitute approximately five million running words of English. They include extracts from a large variety of genres, such as business reports, business correspondences, business conference proceedings, economic analysis, business policies, commercial laws, financial investigations, commercial news, business journal articles, financial periodicals and so forth. Furthermore, it should be noted that all the language materials employed in this corpus can be accessed in the public domain (see Appendix B for the sample text from the NECCD). Therefore, ethical approval was not required in the section of data collection.

4.3.2 The Translational English Corpus of Commercial Translation from Chinese

The TECCTC, as the target corpus, is compiled from translations of commercial documents from the Chinese language and as such is expected to provide characteristics of Chinese translators' use of English collocations in the commercial area. Those translations were selected from texts originating mainly from authoritative media and public record resources, such as the official websites of Ministry of Commerce of the People's Republic of China, China Council for the Promotion of International Trade, State Administration of Foreign Exchange, China Securities Regulatory Commission, China Daily, Xinhua News Agency and so forth. These language materials make up approximately five million English words in size, which mainly corresponds to the NECCD in terms of genre. The corpus includes the commercial documents that were published from the 1970s to 2010, such as business reports, business magazines, business conference proceedings, business letters, broadcast transcripts, company policies, commercial news, journal articles, periodicals and so forth (see Appendix B for the sample text from the TECCTC). Even though some of these commercial documents are translated by anonymous translators, the quality of these translations represents an advanced standard in China because they are all on official authoritative websites.

Therefore, the translators examined in this study are considered as advanced L2 learners of English.

4.3.3 General information on the corpora

Both of the two designed corpora used in this study are largely comparable with regard to size, genre and register, and they can provide reliable language materials for exploring the use English collocations in two different speech communities. The relevant information from the corpora is shown in the following table:

Table 4.1 General information of the corpora in the study

	NECCD	TECCTC
Number of tokens	5,238,867	5,166,993
Number of types	70,532	47,692
Type-token ratio	1.35	0.92

4.4 Data processing

This section will explain the approach of retrieving collocations from the two designed corpora. Data in this context refers to collocation.

4.4.1 The Bigram Model in collocation retrieval

This study employs the Bigram Model during the course of collocation retrieval. This model is derived from the N-gram Model (e.g. bigram and trigram), which is formalised to predict lexical occurrences from the statistical perspective. Built on probability, the N-gram Model is also called the language model according to some researchers (e.g. Jurafsky & Martin, 2000). The prefix “n” refers to any natural number (1,2,3,4...), and particularly in collocation studies, researchers look at the situations where n is equal to, or greater than, 2. In this study, n is equal to 2. The mechanism of the n-gram model in

corpus linguistics is to identify lexical sequences based on word form. It should be noted here, nevertheless, that the word form is “the [full] inflected [or derived] form as it appears in the corpus” (Jurafsky & Martin, 2000, p. 193). This notion is opposed to lemma which enables corpus linguists to generalise about the behaviour of groups of words in cases where their individual differences are irrelevant (such as *dog* and *dogs*). Thus, word form and lemma are known as token (the actual number of running words) and type (the number of distinct words) respectively when they are counted in a corpus. In respect to this distinction, researchers can process collocation candidates and investigate the co-occurrence relationship between actual lexical items based on the n-gram model. For instance, if the Bigram Model is used to identify two-word sequences, the sentence from the TECCTC, *China has been trying to stimulate domestic consumption in a bid to boost the economy*, can thus be segmented as the following:

China has, has been, been trying, trying to, to stimulate, stimulate domestic, domestic consumption, consumption in, in a, a bid, bid to, to boost, boost the, the economy.

From these two-word sequences, *domestic consumption* may be identified as a collocation candidate through *FoxPro* programming. As such, it works the same with the trigrams, quadrigrams/four-grams, pentagrams/five-grams and so forth. The fragmentation of n-grams is sentence-aware, which means it starts from the first word of a sentence and advances one word at a time until it comes to the last word of the sentence. This will exclude the situations where co-occurring words in different sentences are regarded as a collocation. The present study is also in line with this perspective and will only examine co-occurring words within a sentence. In this way, every time a lexical sequence is obtained, it is examined against the previously obtained sequences, and can be stored as an entry of a frequency list in a table generated by *FoxPro* programming.

This indicates that, based on the Bigram Model, I can retrieve all the bigrams from the two designed corpora with *FoxPro* and list them in the frequency-descending order. I

can also determine keywords from these obtained bigrams and establish a solid basis for exploring discontinuous collocations. Furthermore, this model is also the prerequisite to implement statistical measures to retrieve statistically acceptable collocations.

4.4.2 Statistical measures

Statistical tests are acknowledged as reliable measures in frequency-based collocation studies. There are five main testing methods for identifying collocations, namely, the Mutual Information (MI) score test, the Z-score test (or Z test), the T-score test (T test), the Chi-square test, the Log-Likelihood (LL) test. These methods are used to look at whether words tend to co-occur more frequently than expected by chance alone. Among these testing methods, the Z-score test and the T-score test are parametric tests, and the Chi-square test, the Log-Likelihood test are non-parametric tests.

This study will only employ the Log-Likelihood test and the Mutual Information score test as the statistical measures for data retrieval from the two designed corpora, because the other three testing methods have some limitations in identifying collocations. The Z-score test takes expected occurrence as its denominator, so the researcher may obtain a mistakenly high score when they are examining words with an extremely low frequency (Evert, 2004). According to Butler (1985), when the sample size falls below 30, the z-score will not be reliable. The T-score test assumes that “the sampling distribution of the mean is approximately normal even where the distribution within the original population is not normal, provided that the sample size is large” (Butler, 1985, p. 75). The above issues indicate, in collocation studies, that parametric tests take a corpus as a great number of bigrams which may constitute a collocation or may not. They generate a binomial distribution, for which Dunning (1993) states that the “agreement between the binomial and normal distributions is exactly what makes test statistics based on assumptions of normality so useful” (p. 65). In this respect, if the mean number of positive outcomes is comparatively high, the binomial distribution is approximately the normal distribution. However, if the mean number of positive

outcomes is relatively low (e.g. the identification of collocations), the binomial distribution is heavily skewed (positively or negatively), thus violating the assumption of normality, and the probabilities calculated with the normal approximation are, to a large extent, inaccurate (Dunning, 1993). Similar to the Z-score test, the Chi-square test (or Pearson's Chi-square test) may produce considerably high values at times, which is largely due to the underlying normality assumption when it is used for samples of a small size or when the probability is too low. As Snedecor and Cochran (1989) note, the Chi-square test is not recommended to use when the total sample size is smaller than 40 and the expected value in any of the cells is less than 5.

Therefore, a number of researchers (e.g. Dunning, 1993 and Manning and Schütze, 1999) suggest the employment of the Log-likelihood ratio test which is more advantageous in investigating small sample numbers and sparse data, such as infrequent collocations in a corpus, by measuring co-occurrence affinity. In this sense, this statistical measure can fulfil the requirements of retrieving collocations with low frequencies. In addition to the Log-likelihood ratio test, some other researchers (e.g. Church & Hanks, 1990) also recommend the Mutual Information score test because it is useful to estimate word association norms. Therefore, this study will only employ the Log-likelihood ratio test and the Mutual Information score test. I will discuss these two statistical measures in more detail below.

The Log-likelihood test

The Log-likelihood ratio test is a non-parametric test, which does not assume normal distribution of probabilities. The calculation of LL ratio between two adjacent words (i.e. a bigram) is also based on their frequencies. LL ratio can be calculated as follows:

$$\begin{aligned}
 LL &= -2 \log \lambda = -2 \log \frac{b(c_{12}, c_1, p) b(c_2 - c_{12}, n - c_1, p)}{b(c_{12}, c_1, p_1) b(c_2 - c_{12}, n - c_1, p_2)} \\
 &= -2 (\log b(c_{12}, c_1, p) + \log b(c_2 - c_{12}, n - c_1, p) - \log b(c_{12}, c_1, p_1) - \log b(c_2 - c_{12}, n - c_1, p_2))
 \end{aligned}$$

where b stands for a binomial distribution, i.e. $b(k, n, x) = x^k (1-x)^{n-k}$, c_1 the frequency of w_1 , c_2 the frequency of w_2 , c_{12} the frequency of co-occurrence of w_1w_2 , n the corpus size, p : c_2/n , p_1 : c_{12}/c_1 and p_2 : $(c_2 - c_{12})/(n - c_1)$ (see Manning & Schütze, 1999). For ease of calculation, this formula can also be re-written as follows:

$$LL = -2(\log(p)c_{12} + \log(1-p)(c_1 - c_{12}) + \log(p)(c_2 - c_{12}) + \log(1-p)((n - c_1) - (c_2 - c_{12})) - \log(p_1)c_{12} - \log(1-p_1)(c_1 - c_{12}) - \log(p_2)(c_2 - c_{12}) - \log(1-p_2)((n - c_1) - (c_2 - c_{12})))$$

For instance, for the *budget* and *deficit* pair, the word *budget* occurs 342 times, the word *deficit* occurs 307 times and they co-occur 23 times in the 5,166,993-token TECCTC. Thus, for *budget* and *deficit*:

$$LL \text{ ratio}(\text{budget}, \text{deficit}) = -2 \times (\log(5.94156e-5) \times 23 + \log(1 - 5.94156e-5) \times (342 - 23) + \log(5.94156e-5) \times (307 - 23) + \log(1 - 5.94156e-5) \times ((5166993 - 342) - (307 - 23))) - \log(6725.14620e-5) \times 23 - \log(1 - 6725.14620e-5) \times (342 - 23) - \log(5.49679e-5) \times (307 - 23) - \log(1 - 5.49679e-5) \times ((5166993 - 342) - (307 - 23))) = 121.97$$

The critical value of LL ratio is 3.84 at the significance level $\alpha=0.05$ for one degree of freedom, which means that the LL value between two words needs to be equal or higher than 3.84 if they intend to show statistical significance. Since the obtained outcome 121.97 is much greater than the critical value, I can therefore claim that the word pair *budget* and *deficit* constitute a significant collocation in the TECCTC.

The Mutual Information test

Mutual Information (MI) is another means or index of measuring the statistical significance of the association between words. The notion of MI can be theorised as “a symmetric, non-negative measure of the common information in the two variables” (Manning & Schütze, 1999, p. 67), which looks at “how much one word tells us about

the other” (p. 178). In this sense, MI essentially differs from the hypothesis testing methods (e.g. Log-likelihood) because, as Clear (1993) has summarised, “MI is a measure of the strength of association between two words” whilst a hypothesis-testing method is a measure of “the confidence with which we can claim there is some association” (pp. 279-282).

MI is particularly useful to estimate word association norms when researchers are using corpora. The MI score takes expected occurrence as its denominator, and it is computed based on the likelihood of the lexical units occurring together within a particular word span. In collocation studies, MI indicates collocational status between lexical items and measures statistical distinctiveness. To be more specific, any two words in a corpus will attain a MI score --- the higher the MI score, the stronger the association between the two words. The MI score can be calculated using the following formula:

$$MI(w_1, w_2) = \log_2 \frac{p(w_1 w_2)}{p(w_1) p(w_2)} = \log_2 \frac{O}{E}$$

It is obvious from this formula that the MI test actually compares the probability (p) of observing w_1 and w_2 as a whole ($p(w_1 w_2)$) with the probability of observing the two words independently (by chance, $p(w_1)p(w_2)$). Put another way, if w_1 and w_2 co-occur more frequently than by chance, their joint probability, i.e. $p(w_1 w_2)$, will be greater than their independent frequencies, i.e. $p(w_1)p(w_2)$, with their MI score being larger than 0. Contrary to this, if w_1 and w_2 are not statistically associated with each other, $p(w_1 w_2)$ will come infinitely close to $p(w_1)p(w_2)$, with the MI score being equal to, or even smaller than 0 (in the cases where an MI score < 0 , this means w_1 and w_2 are in complementary distribution, that is, when w_1 occurs w_2 tends not to occur, and vice versa) (see Church & Hanks, 1990). In respect to the critical value, Hunston (2002) proposes that an MI score of 3 or higher can be seen as statistically significant. This study employs this criterion when retrieving significant English collocations in the commercial register.

For instance, the MI score between the word pair *budget* and *deficit* in the TECCTC can be calculated as follows:

$$MI(budget, deficit) = \log_2 \frac{\frac{23}{5166993}}{\frac{342}{5166993} \times \frac{307}{5166993}} = 10.14$$

Since the MI score between *budget* and *deficit* is greater than the critical value 3, it is safe to claim that these two words constitute a strong probability of collocation in the TECCTC.

As stated above, the Log-likelihood ratio test and the MI score test are both reliable means of measuring the collocational status between lexical items from the statistical aspect. Further, it should be noted that this study will employ a ‘minimally optimal’ approach when retrieving collocation candidates. That is to say, for any word pair to qualify as a collocation candidate, it will have to show statistical significance in both of the two tests. Both the Log-likelihood ratio test and the MI score test will be carried out with Beijing Foreign Studies University Collocator (see 4.5.3.2).

4.4.3 Software tools for data retrieval

FoxPro Programming

Visual *FoxPro* (hereafter referred to as *FoxPro*) is both a management system and a text-based programming language, which provides researchers a powerful ‘platform’ for retrieving and processing accurate datasets from native-speaker language, and helps them create an integrated relational database system to manage data effectively and efficiently (Fan, 2010a). The robustness of *FoxPro* programming has become increasingly prominent in NLP (natural language processing) studies with the advent of large-scale corpora. *FoxPro* can not only provide language researchers with accurate information to facilitate their data retrieval tasks, but also improve their creativity and arouse their interest when they are handling the tedious data processing tasks. *FoxPro* is

user-friendly, offering a wide set of natural-language-like operators, commands and functions which can be entered in the operation window. Furthermore, *FoxPro* is more advantageous under a corpus-driven approach when compared to other extraction packages, because it allows for mathematical formulas and can perform complicated mathematic operations/computations and string manipulations. This is particularly important for investigating lexical co-occurrences (e.g. collocations) with a corpus-driven approach.

In this study, collocations are observed and retrieved based on the Bigram Model (e.g. Jurafsky & Martin, 2000). Therefore, in order to maintain the accuracy of data processing, this study employs *FoxPro* programming to retrieve bigrams from the two designed corpora. This is a crucial step to obtain the collocation candidates before they are examined through statistical measures. As stated in Section 4.5.1, the two designed corpora will be ‘fragmented’ into lexical sequences of 2 words (according to the definition of collocation in this study) before collocation candidates are further identified. This task can be completed with *FoxPro* programming.

The programme for retrieving bigrams (see Appendix C) consists of six main steps: a. preparation; b. table creation; c. data input; d. scanning; e. data output; f. editing. In this programme, statements 1 to 5 are prerequisite commands to start *FoxPro* programming. Statement 6 creates a temporary table named *wordtable* in work area 1, which only provides a medium for processing data and will be deleted automatically once the programme has finished running. Statement 7 creates a two-field table named *bigram.dbf* (the widths of the two fields are 40 characters and 6 digits), which helps store the retrieved bigrams and their frequencies. Statement 11 assigns an initial value *nothing* to the variable *twowords* which is also employed for storing and computing bigrams. Statements 12 to 13 input all the contents of the text file *TECCTC1.txt* and replace `*()_` with a space respectively. Statements 14 to 15 tokenise the file *TECCTC1.txt* and put these tokenised contents of the file *TECCTC1.txt* into a temporary text file *temp.txt*. Statements 16 to 17 open the created temporary table *wordtable* in work area 1 and append the entire tokenised file *TECCTC1.txt* from the temporary file

temp.txt to the temporary table *wordtable*. Statement 18 sets the recorder pointer to move from the beginning of the temporary table *wordtable*. That is to say, the words in the tokenised file *TECCTC1.txt* are listed individually in the records of *wordtable*. Statements 19 to statement 31 constitutes a set of loop commands, which means that, as long as the recorder pointer remains in the temporary table, it will ‘scan’ the records till it comes to the end of the table. The words in these records are adjacently combined to constitute bigrams and stored in *twowords*. This process advances one word at a time until it comes to the last word of *wordtable*. Then, all the bigrams previously stored in *twowords* are moved to a temporary text file *temp.txt* as demonstrated in statement 32. In statement 33, the two-field table *bigram.dbf* is accessed in the working area 2, in which all the bigrams in the temporary file *temp.txt* are appended according to statement 34. From statement 35 to 40, these bigrams are computed in terms of frequency and eventually stored alphabetically in a plain text file.

Based on the results obtained using this programme, I copied all of the contents to Microsoft Word or Excel for further editing or adjusting to sort out those whose frequency is equal to or greater than 5 (this is in accordance with the operation definition of collocation in this study). The general bigram information across the two designed corpora is demonstrated in Table 4.2:

Table 4.2 Bigram information in the TECCTC and the NECCD

	NECCD	TECCTC
Number of bigram types	137,285	117,697

It is obvious from this result that not all the bigrams can constitute collocation candidates. Therefore, all the retrieved bigrams will undertake the filtering devices (see 4.6) and the statistical tests (see 4.5.3.2) if they can be identified as a collocation candidate.

Beijing Foreign Studies University Collocator

Apart from *FoxPro*, this study also employs *Beijing Foreign Studies University*

Collocator (Version 1.0, henceforth *BFSU Collocator*) as a tool to retrieve collocation candidates from the statistical perspective. This software is designed and developed by Jiajin Xu and Yunlong Jia (see Xu & Jia, 2009) at The National Research Centre for Foreign Language Education of Beijing Foreign Studies University. In this study, the *BFSU Collocator* is not only used to retrieve discontinuous collocation candidates but also used to compute the statistical values between co-occurring words.

The *BFSU Collocator* offers the Log-likelihood test and the MI score test as needed in this study, and greatly facilitated complicated statistical calculations. Specifically, when examining the collocability of two particular words, I simply needed to load the text files, set the keyword and word span (in this study, the word span is ± 5 based on the criteria in the section 3.5.1), and leave the computation work completely to the software. This convenient operation saves time entering data (e.g. corpus size) and adjusting some advanced settings (e.g. weight case) when compared to other statistical tools, such as SPSS (Statistical Product and Service Solutions). Particularly, the *BFSU Collocator* can compute the statistical values between a particular keyword and all of its collocates simultaneously, and list all these collocates in alphabetical order. In other words, there is no need to calculate the statistical values between a keyword and its collocates individually. In short, the whole process of examining collocability with statistical measures in this study can be best summarised as: corpus loading \rightarrow data setting \rightarrow operating.

As mentioned above, the *BFSU Collocator* cannot retrieve ‘no-keyword’ clusters, such as bigrams, from a corpus. In other words, it cannot meet the criteria of the studies with a corpus-driven approach. This is the reason why it could not be used alone in this study. In this sense, *FoxPro* and the *BFSU Collocator* were complementary with regard to functionality in this study. Nevertheless, researchers using a corpus-driven approach can simply design a model to identify the keywords from the bigrams. In this study, the selection of a keyword was formulated to meet these criteria: a. it must be a content word; b. it must be commerce-related; c. it must occur at least five times in the chart of retrieved bigrams. For instance, Figure 4.3 demonstrates that the word *deficit* is a

commerce-related content word and it also co-occurs with at least 17 words (the word *deficit* still co-occurs with a lot more other words in the examined text file but they are not displayed due to the limit of the window). Therefore, *deficit* shows its strong ability to collocate with other words and can be regarded as a keyword. Based on the bigram lists retrieved from the two corpora, the information of the keywords for statistical tests is shown in the table below:

Table 4.3 Keyword information in the TECCTC and the NECCD

	NECCD	TECCTC
Number of keyword tokens	535,465	689,073
Number of keyword types	1,605	1,285

These keywords can be tested individually with the *BFSU Collocator*, thus obtaining all the collocation candidates across the two corpora. For instance, if the word *deficit* is regarded as a keyword in the NECCD and is entered in the search field of the *BFSU Collocator* operation window, some collocation candidates, such as *budget deficit* and *trade deficit*, can be identified after processing as shown in Figure 4.3 below (not all the collocations are displayed due to the limited size of the window). These collocations show statistical significance (see 4.5.2) and meet the requirements (see 3.5.1) of this study. Therefore, they can be processed further in the following filtering procedure (see 4.6).

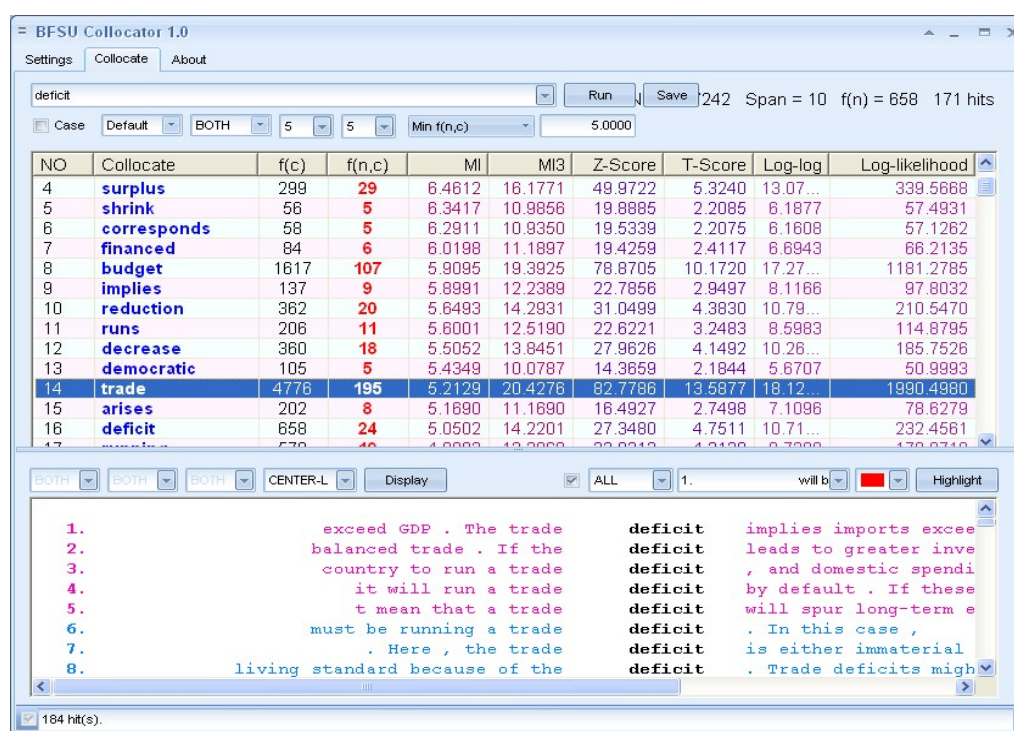


Figure 4.3 Statistical results for *deficit* and its collocates in the NECCD

This operation interface shows the extent to which the word *deficit* collocates with other words statistically, apart from *budget* and *trade*. This process indicates that once a keyword is confirmed, the *BFSU Collocator* can not only compute the statistical results from the observed bigrams but it can also help researchers find more collocates of this keyword. In this respect, this software tool greatly helps reduce some overlapping work when researchers intend to make a collocation list of thousands of entries. In addition, another advantage that the *BFSU Collocator* possesses over other tools is that it can display the contexts in which a particular word pair occurs. Therefore, researchers can trace back to the original texts with ease. This function will also help researchers, to a great extent, look at the actual usage of this word pair, and analyse it further in terms of form, meaning and function. Particularly in this study, this could largely facilitate the data analysis procedure under the proposed theoretical framework. It should be noted, however, different data retrieval software tools may produce different statistical results even from exactly the same operation or calculation (e.g. the Log-likelihood test). This is largely due to the different mathematic formulas or axioms these tools employ. In addition, sometimes different tools may still produce different statistical results even using the same mathematical formula. This may be due to the following reasons:

- a. during the course of lexical processing, different data retrieval software tools may define a lexical item (or even a token) differently, thus resulting in producing different numbers of tokens. For example, the *BFSU Collocator* regards a word, a number and even punctuation as an individual lexical item and takes it into the statistical computation;
- b. different tools may treat hyphenated words (e.g. *million-pound*) differently. For example, the *BFSU Collocator* regards words such as *million-pound* as one lexical unit rather than two.

Factors such as these definitely contribute to the different statistical results obtained when using different data retrieval software tools. This indicates that when researchers are investigating the lexical association between words, the employment of computation tool should be consistent during the whole process of data retrieval. Only in such a way can their quantitative research be more reliable. Therefore, only the *BFSU Collocator* was employed in this study for computing the statistical values between the lexical items to see whether they can constitute a collocation candidate from a statistical perspective.

4.5 Filtering devices

Although this study has attempted to employ robust software tools and statistical measures to ensure the accuracy of data retrieval, the actual outputs of collocation candidates show that not all these word sequences generated can be accepted as suitable collocations according to the initiatives and criteria of the present study. That is to say, although some collocation candidates are ‘cohesive’ and ‘formulaic’ in a statistical sense, they may not make sense at all in business English, or even in a general sense. For instance, the bigram *of the* occurs 60,195 times in the TECCTC, with the LL value being 31,012.44. This bigram can be regarded as a significant collocation in terms of frequency and statistical measures, but it is not a ‘meaningful’ collocation that this study

aims to explore. N-grams as such include those that start or end with function words, such as *of the*, *in the*, *at the*, and *a*. Strictly speaking, these bigrams possess merely a literal sense or appear in a grammatical/syntactical structure even though they may function as a collocation in a real sense, such as *run for (the bus)* (this is in stark contrast with *run for (president)*). Therefore, with these issues in mind, I set three successive filtering devices to manually check the collocation candidates generated by software tools, namely, the frequency filtering device, the form filtering device and the semantic filtering device. These three devices are also in accordance with the theoretical framework of this study. In this perspective, the collocation retrieval procedure can be illustrated as follows:

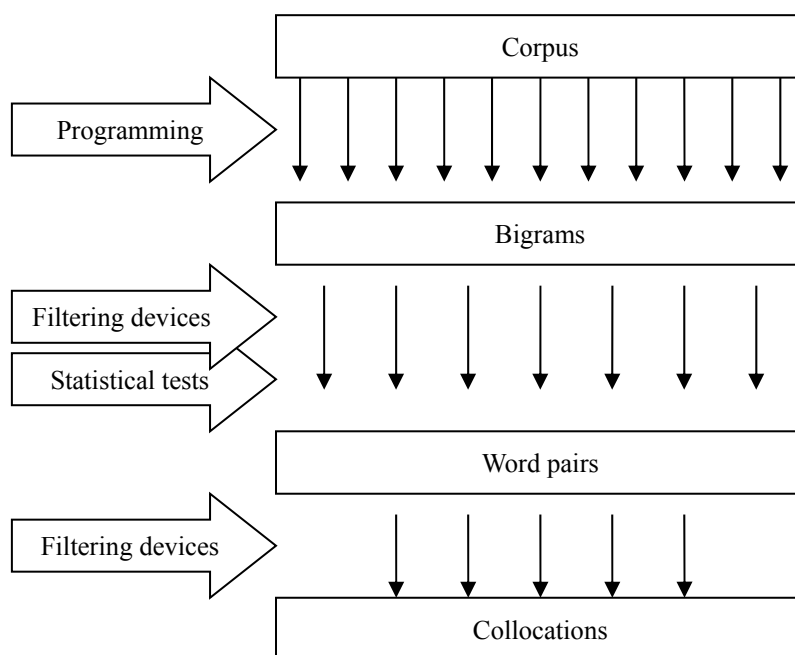


Figure 4.4 Collocation retrieval procedures

4.5.1 Frequency filtering device

This device is required to reduce the instances of overlapping counts of some bigrams and collocation pairs when retrieving collocations. For example, the collocation *boost [the] economy* may occur in the n-grams such as *boosted its economy*, *boosting this*

country's economy and so forth. Therefore, I would double check their grammatical forms and calculate their frequencies appropriately when looking at these apparently 'similar' word strings (especially before they intend to make them a dictionary entry or an item in translation memory from a lexicographical perspective). This device would greatly increase the accuracy of the data retrieved and help draw a more reliable conclusion. In the above example, the word pair *boost* and *economy* in *boosted its economy* and *boosting this country's economy* can be viewed as the same discontinuous collocation, that is, *boost [the] economy*.

4.5.2 Form filtering device

The form filtering device is used to rule out the word pairs that do not qualify as collocations syntactically, or to eliminate those that cannot be seen as multi-word units even though they might be frequently used in language. Such word pairs include: those lexical sequences that are only composed of closed-class words serving as only phrases and clause fragments, such as *in the*, *at the*, *in a* and so forth; those that are partially composed of closed-class words that do not function as part of multi-word units, such as *billion in*, *yuan and*, *yuan in*, *territory of* and so forth; those that are syntactically incomplete, such as *the regional*, *the national*, *the month-long* and so forth; and those that, though syntactically/structurally complete, should be contained and used with other word(s) to constitute a prefabricated unit, such as *five million*, *third quarter* and so forth. In addition, syntactical transformation is also taken into account with the form filtering device. When analysing the collocational relationship between a particular word pair, I would disregard their parts-of-speech. That is to say, the word pair *boost* and *economy* in a nominal phrase as *a boost for the economy* and the one in a verbal phrase as *boost this country's economy* will be regarded as the same collocation for instance. This is also in accordance with the criteria of the corpus-driven approach. In this sense, this device would help investigate a broader range of the actual use of collocations and look at more situations where translators are only able to use collocations as *boost economy* but largely ignore those as *a boost for the economy* in

their L2.

4.5.3 Semantic filtering device

The semantic filter is used to disambiguate the word pairs that share the same lexical form(s) but possess different meanings when they function as collocations. This is also a process of distinguishing the formulaic usage from the non-formulaic usage of word combinations. For instance, when *look into* is used in the context as *Samuel looked into the house but saw nothing*, it is merely used in its literal sense which combines the individual meanings of its constituents, and therefore it should be regarded as a free combination. In contrast, when *look into* is used in the sentence *Samuel looked into the case but found no reliable evidence*, it is then used in its idiomatic sense which differs from the mechanic combination of the individual meanings of the two constituents, and therefore should be regarded as an idiom.

4.6 Collocation retrieval results

After filtering the collocation candidates through these three devices, the statistical results from the two designed corpora are shown in Table 4.4:

Table 4.4 General collocation information across the NECCD and the TECCTC

	NECCD	TECCTC
Number of collocation tokens	101,935	111,450
Number of collocation types	6,366	3,872

Based on these three collocation filters, the obtained statistical results of the top 30 most frequently used collocations from the two corpora can be seen in Appendix D.

4.7 Summary

In response to the third research question, this chapter has outlined a general procedure of how commerce-related collocations from the two designed corpora were retrieved in this study. In terms of research method, this chapter has introduced the notion of Contrastive Interlanguage Analysis and has attempted to construct its own model – a corpus-driven Contrastive Interlanguage Analysis approach. This model indicates that the NL-TL (Native Language-Translational Language) comparison of data in the present study is essentially an NL-IL (Native Language-Interlanguage) comparison. As demonstrated in section 4.2, such a two-way system can provide a clear picture of how L2 deviates from native language through multi-dimensional comparison. Specifically, in this study, this kind of NL-TL comparison can help discover the distinctive features of Chinese translators' use of English collocations and explain the unnaturalness in the translational English they produced.

In addition, this chapter has introduced the corpora employed in this study, namely the NECCD (the reference corpus) and the TECCTC (the target corpus). This chapter also proposed the employment of software tools (*FoxPro* and the *BFSU Collocator*) and two statistical measures (the Log-likelihood test and the MI score test), which not only facilitated the data retrieval task but also ensured the accuracy of the data obtained. Furthermore, I have also elaborated on three filtering devices which proved to be reliable criteria to identify 'meaningful' English collocations in the commercial register. The whole procedure of data retrieval can be specified in this work flow: corpus → bigrams → word pairs → collocations.

More importantly, I have retrieved the collocations from the two designed corpora and have obtained the relevant information needed for data analysis. Therefore, the next chapter will look at the comparison between the two sets of data, attempting to explore findings from the comparison and explain these findings under the theoretical framework constructed in Chapter 3.

Chapter Five Data analysis: Features of Chinese translators' use of English collocations in the commercial register (Part I)

5.1 Introduction

Based on the collocations retrieved using the methodology described in Chapter 4, Chapter 5 and Chapter 6 will report on the results of the empirical investigation into different patterns of collocation use between the NECCD and the TECCTC. As one of the core parts in data analysis, Chapter 5 will look at the distinctive features of Chinese translators' use of English collocations in terms of general collocation density and collocation distribution. This chapter will also investigate these features by comparing the two corpora from the quantitative and statistical aspects.

5.2 Features of collocation density and collocation distribution regarding overall frequency

Collocation density can be defined as the ratio between the number of collocation types and the total number of collocation tokens, that is, the collocation type-token ratio (TTR). Because the present study aims to investigate the difference in using English collocations between texts in the native language and texts in translational language, the concept of collocation density is employed as an important indicator for measuring the significance of collocation in each corpus. The data retrieved with regard to the use of collocation across the NECCD and the TECCTC is generalised as per Table 5.1. Normally, the TTR is calculated with the formula $TTR = 100 \frac{\text{types}}{\text{tokens}}$ (see for instance Biber, 2009). For collocations, the TTR can be formulated as $TTR = 100 \frac{\text{collocaiton types}}{\text{collocation tokens}}$. Thus, the comparison between the two designed corpora can be seen in detail in the following table.

Table 5.1 Comparison of collocation use across the NECCD and the TECCTC

Corpus name	Corpus size	Collocation types	Collocation tokens	TTR
NECCD	5,238,867	6,366	101,935	6.25
TECCTC	5,166,993	3,872	111,450	3.47

Table 5.1 demonstrates that, in terms of overall frequency, native speakers of English in the NECCD produced 101,935 collocations, while Chinese translators produced 111,450 collocations in the TECCTC. That is to say, there are 9,515 more collocations present in the TECCTC than in the NECCD. This indicates that, in the approximate 5-million-token sampling of the present investigation, Chinese translators produced slightly more collocations (with frequency equal or greater than 5) than native speakers. In contrast, when examined in terms of collocation type, these two sets of data show a completely different result. To be specific, there are 6,366 types present in the native-speaker NECCD, with the TTR being 6.25, while there are only 3,872 types occurring in the Chinese translators' TECCTC, with the TTR being 3.47. This result implies that, even though Chinese translators produced slightly more collocation tokens in a similar size of language sampling, they produced 2,494 fewer collocation types when compared with native English speakers. This finding appears to show that Chinese translators may be using the same collocations over and over again while native speakers may be using a diverse range of collocations.

Nevertheless, this discrepancy between tokens and types is not enough to ascertain that there is a difference in terms of collocation use between the two corpora. Therefore, it is necessary to conduct a statistical test to examine whether the difference is significant. Because these two groups of collocation data retrieved from the two designed corpora are independent samples, the T test can be employed here to examine the significance. With the Statistical Package for the Social Sciences (version 16.0, hereafter abbreviated as SPSS), the results of the T test are shown in Table 5.2.

Table 5.2 T-test regarding the comparison of collocation use between the NECCD and the TECCTC

	Levene's Test for		t-test for Equality of Means						
	Equality of Variances								
	F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std. Error difference	95% confidence Interval of the difference	
Equal variances assumed	408.199	.000	14.940	10236	.000	12.77168	.85486	11.09598	14.44738
Equal variances not assumed			13.199	5.438E3	.000	12.77168	.96761	10.87478	14.66858

Note: In the T score test, the critical value at 95% confidence level is 1.96. If the obtained T score is greater than 1.96, the difference between the two groups of data will be significant; if the obtained T score is lower than 1.96, then the difference will not be significant.

As demonstrated in Table 5.2, the absolute value of the T score test is 14.94, which is greater than the critical value 1.96 at 95% confidence interval. This result indicates that there is a statistical significance between Chinese translators and native speakers in terms of the use of English collocations. Furthermore, as Table 5.1 suggests from the aspect of the degree of variation, Chinese translators over-produce collocation tokens by 9.33% ($111450/101935 - 1$) but under-produce collocation types by 39.18% ($1 - 3872/6366$) when compared with native speakers. It should also be noted that the difference of collocation tokens between these two groups of speakers is much less than that of collocation types. From the above comparison, it is clear that Chinese translators' use of collocations, to a great extent, deviates from that of native speakers, all of which might lead to such an assumption that Chinese translators repeatedly produce certain collocations but fail to increase the collocation variety in their L1-to-L2 translation practice. The obviously lower TTR (3.47) in the TECCTC (against 6.25 in the NECCD) also supports this assumption.

Notwithstanding the deviation of Chinese translators' use of English collocations, the exact nature of their production of L2 collocations still remains unclear. Therefore, it is necessary to investigate the issue of deviation in more detail, specifically, how and to

what extent Chinese translators' use of English collocations deviates from that of native speakers. In order to clarify these features and carry out an objective analysis, I designed a 'grouping' approach to compare the two sets of collocations obtained from the NECCD and the TECCTC. First of all, I sorted each set of the collocations in frequency-ascending order. Secondly, I divided each set of collocations into 21 groups from the beginning to the end. These groups are named with English letters from A to U. Each group is made up of collocations of a particular frequency range, which can be demonstrated as follows:

Group A: $5 \leq F \leq 50$ (where F indicates a frequency, and hence forth);

Group B: $51 \leq F \leq 100$;

Group C: $101 \leq F \leq 150$;

Group D: $151 \leq F \leq 200$;

Group E: $201 \leq F \leq 250$;

Group F: $251 \leq F \leq 300$;

Group G: $301 \leq F \leq 350$;

Group H: $351 \leq F \leq 400$;

Group I: $401 \leq F \leq 450$;

Group J: $451 \leq F \leq 500$;

Group K: $501 \leq F \leq 550$;

Group L: $551 \leq F \leq 600$;

Group M: $601 \leq F \leq 650$;

Group N: $651 \leq F \leq 700$;

Group O: $701 \leq F \leq 750$;

Group P: $751 \leq F \leq 800$;

Group Q: $801 \leq F \leq 850$;

Group R: $851 \leq F \leq 900$;

Group S: $901 \leq F \leq 950$;

Group T: $951 \leq F \leq 1000$;

Group U: $F \geq 1001$.

The results are demonstrated in Figure 5.1 and Figure 5.2 respectively:

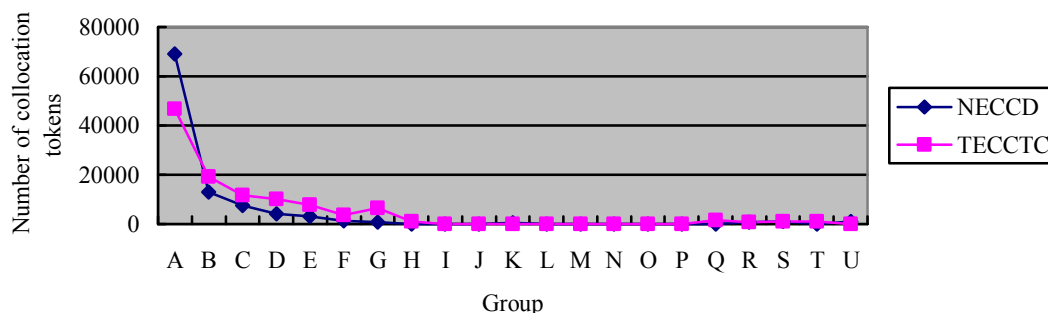


Figure 5.1 Distribution of collocation tokens in the NECCD and the TECCTC

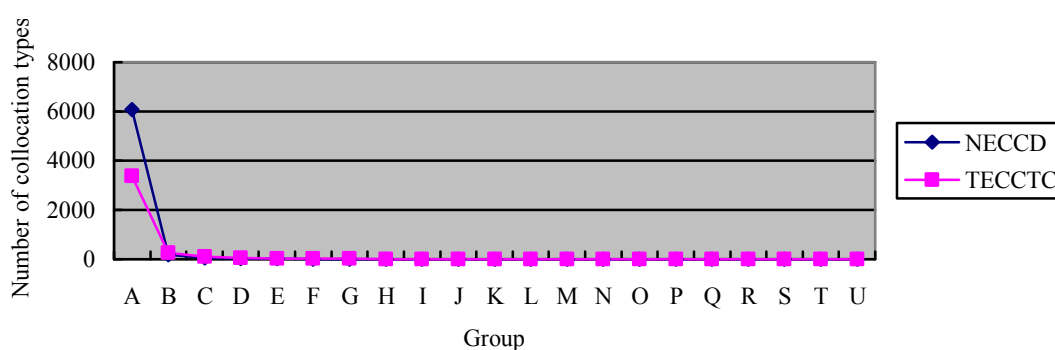


Figure 5.2 Distribution of collocation types in the NECCD and the TECCTC

As illustrated in Figure 5.1, both the curves reveal a declining tendency with the NECCD showing a more striking decline than the TECCTC. This means that with the increase of frequency, the number of collocations tends to decrease. Obviously, from group B onwards (especially, in group B, C, E, F, G, H, Q, S, T), the values in the TECCTC are significantly higher than those in the NECCD with regard to the number of collocation tokens. This is exemplified by the top ten most significantly overused collocations in the TECCTC as compared to the NECCD with the results demonstrated in Table 5.3. Apart from the raw frequencies of these collocations, their respective Log-likelihood ratios (see 4.4.2) are also computed with SPSS to indicate the statistical significance across the two corpora.

Table 5.3 Top 10 overused collocations in the TECCTC as compared with the NECCD

Collocations	TECCTC	NECCD	Log-likelihood
joint venture	974	158	586.391
stock exchange	947	32	999.183
intellectual property	853	350	175.377
financial crisis	839	29	881.395
technological renovation	824	–	–
bilateral relations	364	8	407.956
general manager	357	–	–
custodian fund	352	–	–
application documents	349	–	–
global crisis	348	–	–

As noted in Table 5.3, the blank boxes in the third column mean that either these collocations do not show statistical significance or they did not occur at all in the NECCD. Since the Log-likelihood ratios obtained are far greater than the critical value 3.84, the difference in using these overused collocations appears to be highly significant across the two corpora. These overused collocations occur 6,207 times in the TECCTC, accounting for 5.57% of the complete tokens, while they occur 577 times in the NECCD, accounting for a mere 0.57%. This result indicates that, in the present study, Chinese translators appeared to rely heavily on some favoured collocations (e.g. *joint venture* and *financial crisis*), that is, those whose frequency is greater than 50.

However, Group A in Figure 5.1 shows that the values found in the TECCTC are significantly lower than those found in the NECCD with the result being 46,833 as opposed to 69,038. This result means that low-frequency collocations, specifically those that occur no more than 50 times, make up 42.02% and 67.73% respectively in the TECCTC and the NECCD. Figure 5.2 illustrates that Chinese translators produced 3,368 low-frequency types, accounting for 86.98% of the complete collocation types in the TECCTC, while native speakers produce 6,070 low-frequency types, accounting for 95.35% in the NECCD. In this sense, the major factor differentiating these two groups

of language users regarding the use of English collocations is the production of comparatively low-frequency collocations by both groups. To be specific, in comparison with native norms, Chinese translators produced fewer low-frequency collocations with regard to either collocation tokens or collocation types. This may also be said to be directly implicated in the very dissimilar results regarding the type-token ratios from the comparison of the corpora in the present study.

Taken together, the results obtained from the above statistical tests can simply justify the aforementioned assumption that Chinese translators depend heavily on high-frequency collocations and show repeated use of these favoured collocations in their Chinese-to-English translation practice. Researchers in this area should pay more attention to the low-frequency collocations, especially with frequency being smaller than 50, which are normally produced by native speakers but are not used appropriately in translational English. This will be discussed in Section 7.2.1 in more detail.

5.3 Features of collocation distribution regarding statistical values

In addition to the analysis of overall frequency in translational English, this study also looks at the features of collocation distribution with regard to statistical values to examine whether the production of a smaller repertoire of collocation types by Chinese translators is correlated with the statistical values obtained from data retrieval. As shown in the data retrieval section (see 4.5.2), the MI score test and the Log-likelihood test are employed to make a distinction between word combinations with statistical significance (collocations) and those without statistical significance (non-collocations). The critical values (thresholds) for MI and LL are 3 and 3.84 respectively (see for instance Stubbs, 1995). Notwithstanding the fact that over 100,000 collocations in each corpus are identified to pass the statistical thresholds, these results are not strong enough to present the features of how Chinese translators' production of English collocations deviates from native use. That is, some collocations narrowly pass the statistical threshold with very low statistical values, such as *marketing manager* (MI=3)

and *bear market* (MI=3.02) in the NECCD, whilst some others appear to be more ‘intimately associated’ with high statistical values, such as *real estate* (MI=12.45) and *initial offering* (MI=15.53) in the NECCD. To exemplify this sort of contrast, the top ten collocations with the highest MI scores and lowest MI scores from the two corpora have been listed in Table 5.4 and Table 5.5 respectively:

Table 5.4 Top 10 collocations with highest MI scores across the two corpora

Rank	NECCD	MI	TECCTC	MI
1	initial offering	15.53	ensure accuracy	48
2	local authority	15.47	brake override	13.94
3	keep afloat	15.32	rotary contouring	13.79
4	checklist disclosure	14.93	customary apportionment	13.58
5	software legalization	14.28	audiovisual presswork	13.55
6	wreak havoc	14.19	drilling rigs	13.46
7	conveyor belt	13.61	deficiency repaid	13.27
8	irrational exuberance	13.6	retrieval quotations	13.2
9	pierce veil	13.53	thin capitalisation	13.16
10	uncorrected misstatements	13.48	bird flu	13.14

Table 5.5 Top 10 collocations with lowest MI scores across the two corpora

Rank	NECCD	MI	TECCTC	MI
1	real terms	3	importing country	3
2	cause increase	3	trade pact	3
3	strong relationship	3	fortune securities	3
4	credit facility	3	container inspection	3
5	marketing manager	3	rapid recovery	3
6	records show	3	promote awareness	3
7	acquisition date	3	excess demand	3
8	proposed deal	3	financial crisis	3.01
9	boost revenue	3	panel members	3.01
10	job flexibility	3	chain index	3.01

In order to address the issue of collocational association from the statistical consideration, I developed a similar ‘grouping’ approach as in Section 5.2.1.1 and classified the collocations retrieved in each corpus into five groups based on their MI scores. In this way, all the collocations are graded according to a scale of collocational strength, with the details being as follows:

Group 1: weak associations which have an MI score ranging from 3.00 to 3.99;

Group 2: weak moderate associations which have an MI score ranging from 4.00 to 4.99;

Group 3: moderate associations which have an MI score ranging from 5.00 to 6.99;

Group 4: strong moderate associations which have an MI score ranging from 7.00 to 7.99;

Group 5: strong associations which have an MI score greater than 8.

This method aims to exhibit the comparison of collocation distribution across the two corpora with regard to the degree of association, as well as the proportions of each group of collocations along the scale of strength within each corpus. The coverage of each group regarding collocation tokens and collocation types across the two corpora is illustrated in Figure 5.3 and Figure 5.4 respectively. The horizontal axes in these two figures include all the collocations investigated.

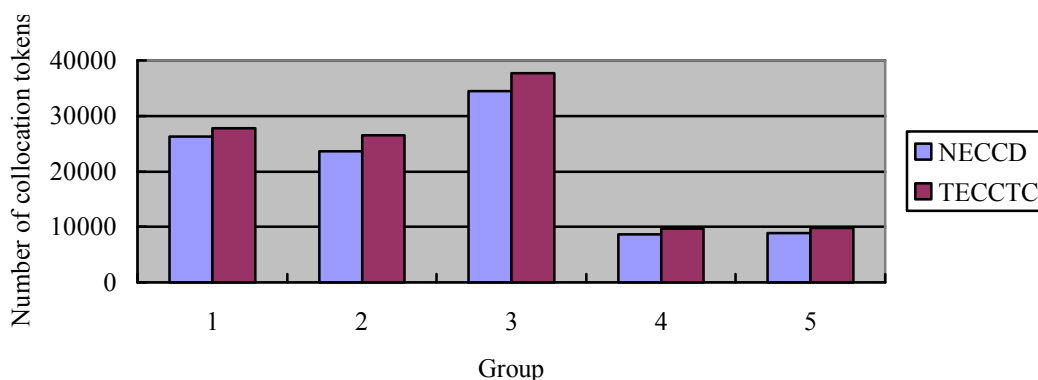


Figure 5.3 Distribution of collocation tokens regarding MI score

The results in Figure 5.3 appear to confirm the aforementioned conclusion (see 5.2) that Chinese translators overuse collocation tokens as a whole, and appear to demonstrate that both the NECCD and the TECCTC display a similar tendency across different levels of strength of collocation association. The independent samples test between these two groups of data also shows that there is no statistical significance, with the t-test result being 0.255 (<1.96) while the significance value is 0.805 (>0.05). This indicates that, when compared with native speakers of English, Chinese translators develop a similar pattern in using collocations of different levels of association in their L2 English. On average, 33.78% of the collocation tokens produced by Chinese translators come from group 3, that is, the band of moderate collocations ($5 \leq MI \leq 6.99$). In comparison with this, a mere 8.82% of collocation tokens comes from group 5, that is, the band of strong collocations ($8 \leq MI$) in the TECCTC. In addition, group 1 and group 2 both appear to indicate that Chinese translators use comparatively weak collocation tokens to a considerable extent in Chinese-to English commercial translations. Therefore, these findings, as a whole, appear to show that Chinese translators rely, to a less extent, on strongly associated collocations. However, they may still use collocation tokens of different levels of strength of collocation association in their L2 English output.

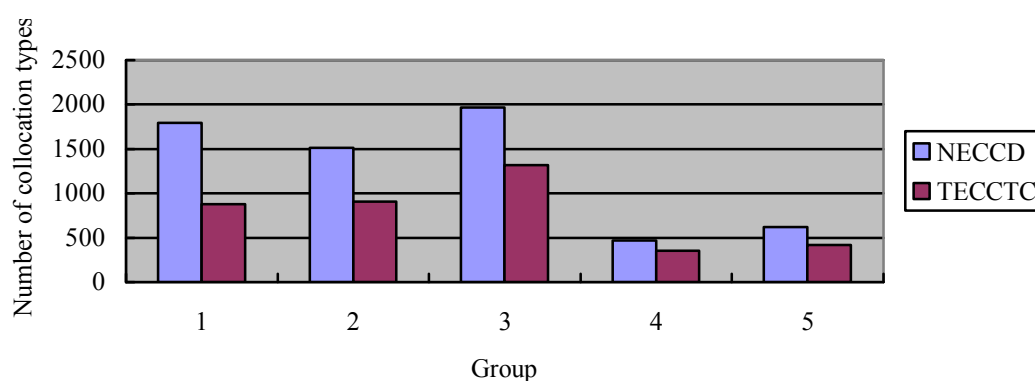


Figure 5.4 Distribution of collocation types regarding MI score

In contrast, the results obtained in Figure 5.4 show a significant difference between the two corpora, which can make a distinction between Chinese translators and native speakers from all the bands of collocational strength. It appears that Chinese translators

under-use collocation types on every scale of MI score groups (group 1: 880-1795; group 2: 906-1513; group 3: 1317-1970; group 4: 350-471; group 5: 419-617), which also corresponds to the previous conclusion (see 5.2) that they produce a smaller repertoire of types as a whole in comparison with native norms. Specifically, the TECCTC/NECCD type ratio (e.g. the ratio in Group 1 is $880/1795 = 49.02\%$) within each group is 49.02%, 59.88%, 66.85%, 74.31% and 67.91%. This tendency is particularly significant in group 1, group 2 and group 3, that is, the collocation types with an MI score smaller than 7. In order to obtain a clearer picture of the extent to which Chinese translators over-produce repeated collocation tokens but under-produce collocation types, it is also necessary to look at the index of the type-token ratio (TTR) based on the scale of collocation strength. This is because TTR can substantially reflect how Chinese translators actually use English collocations in L1-L2 translations. The results of TTRs across the two corpora are shown in Figure 5.5:

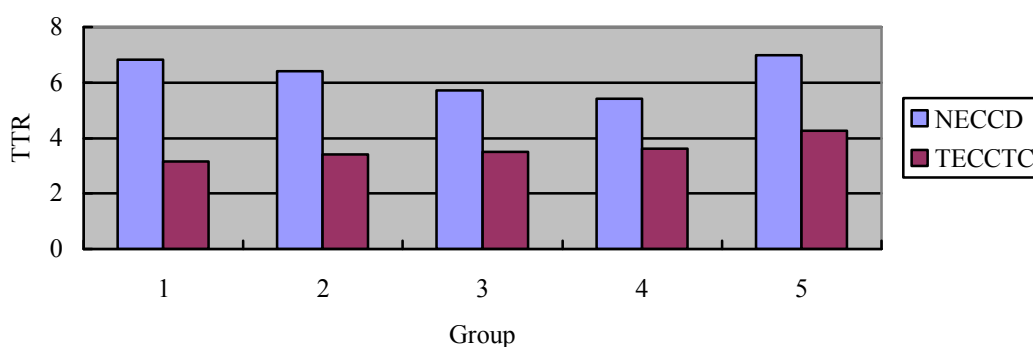


Figure 5.5 Comparison of type token ratios regarding MI score grouping

According to Figure 5.5, the comparison between the two groups of language users is more striking than that of Figure 5.4. Across the five MI score bands, the TTR of the TECCTC is significantly lower than that of the NECCD (group 1: 3.17-8.83; group 2: 3.42-6.41; group 3: 3.5-5.71; group 4: 3.61-5.42; group 5: 4.26-6.98). Even though this result is in stark contrast with the findings shown in Figure 5.3, it largely supports the previously mentioned conclusion that Chinese translators' production of L2 English is characterised by their extensive use of a small number of favoured collocations.

In order to identify the deviation of collocation use in translational English, I designed a model to examine how these repeatedly used collocations are distributed across the MI score bands. Firstly, in Figure 5.4, each value of the TECCTC is divided by its corresponding value of the NECCD, thus, I got 0.49, 0.6, 0.67, 0.74 and 0.68 across the five groups. Secondly, I made the same calculation for the values in Figure 5.5, and similarly, I got 0.46, 0.53, 0.61, 0.67 and 0.61. It should be noted that I used relative value (RV) for each category of the variables by comparing each data pair between the TECCTC and the NECCD to examine the difference between the corpora. It is significant that the NECCD corpus value should serve as the denominator for each data pair comparison because it is the reference corpus in this study. For instance, in terms of type, the RV is $\frac{1317}{1970} = 0.67$ in group 3. The outcome analysis of RV varies from the value 1. More specifically, if the RV is above 1, the greater the value, the bigger the difference between the two corpora; if the RV is below 1, the smaller the value, the bigger the difference; if the RV is equal to 1, there is no difference. The extent to which the two corpora differ can be represented by the absolute value of $RV - 1$. Thus, the two sets of values obtained from step 1 are {0.51, 0.4, 0.33, 0.26, 0.32} and {0.54, 0.47, 0.39, 0.33, 0.39} respectively. Thirdly, I compared the two groups of RVs and obtained difference values, specifically from group 1 to group 5, 0.03, 0.07, 0.06, 0.07 and 0.07. It should be noted that these difference values are also RVs and can only be suitable for linear comparisons. In this case, I propose that the higher the difference value the more overused collocations this group has, because if the overused collocation tokens were distributed evenly across the five groups the differences should be equal or, at least, should not show too much significance. However, group 1 shows a significantly low value when compared with other groups.

Taken together, these results make it very clear that the high-frequency collocations in the TECCTC mainly possess MI scores greater than 4. This result, in return, indicates that Chinese translators rely, to a lesser extent, on weak collocations which, however, show a significantly high proportion in native use. This finding is particularly evident by type and by TTR as shown in Figure 5.4 and Figure 5.5. Furthermore, the

significantly lower TTR of the TECCTC on every scale of MI score groups (see Figure 5.5) also helps provide evidence of Chinese translators' production of more collocation tokens but less collocation types in comparison with native norms.

5.4 Lexical analysis regarding deviation of collocation use in translational English

Based on the foregoing findings, I decided to further analyse the factors contributing to Chinese translators' production of fewer collocation types from the lexical perspective. These factors can also be analysed in terms of lexical coverage and keyword growth. Lexical coverage can be seen as the percentage of a certain size of vocabulary covering investigated texts. Keyword growth, as its name suggests, can be explained as the tendency for keyword types to increase over segmented text(s) of particular lengths in a corpus.

As shown in Figure 4.4, the present investigation includes 101,935 collocations in the NECCD and 111,450 in the TECCTC, made up of 1,783,491 and 1,807,325 word tokens respectively. Further analysis should involve an important notion of lexical coverage (see for instance Nation & Waring, 1997; Laufer & Ravenhorst-Kalovski, 2010), or vocabulary coverage (see Schmitt et al., 2011). Lexical coverage refers to the percentage of a certain size of vocabulary covering one or more texts investigated. To be more specific, for a wordlist W_l , and the vocabulary from a text (or a collection of text) of N word tokens W_t ,

$$W_c = W_l \cap W_t$$

$W_c = \{ W_1, W_2, W_3, W_4, W_5, \dots, W_j \}$, containing j lemmas (a lemma represents a group of words in the cases where their inflectional differences are irrelevant, for instance, *go*, *goes* and *went* are different tokens but the same lemma),

and W_i is any element in W_c , in which has a frequency of F_i ($F_i \geq 0$),

$$W_i \in W_c$$

the lexical coverage C_w can be calculated using the following formula:

$$C_w = \frac{\sum_{i=1}^j F_i}{N}$$

Thus, the concept of lexical coverage is employed in this study for a static analysis of the difference between Chinese translators and native speakers with regard to the use of English collocations, because it is necessary to look at the constituents (words) making up the collocations across the two corpora. With the above lexical coverage formula, the results of lexical coverage for the two corpora are $C_{NECCD} = \frac{1783491}{5238867} = 34.04\%$ and $C_{TECCTC} = \frac{1807325}{5166993} = 34.98\%$ respectively. This result indicates that there is almost no difference between the two groups of speakers regarding word token size. Therefore, it appears that in this study Chinese translators made at least as much use of the units constituting collocations as native speakers, all of which may imply that Chinese translators' deviation may lie in the elements dominating the construction of collocations, that is, the keywords.

As mentioned previously (see Table 4.3), the keyword tokens amount to 535,465 in the NECCD and 689,073 in the TECCTC, and the keyword types amount to 1,605 and 1,285 respectively. In this context, the above lexical coverage formula can also be employed to investigate the coverage of the keywords. Thus, the results of keyword coverage are $C_{NKW} = \frac{535465}{1783491} = 30.02\%$ and $C_{TKW} = \frac{689073}{1807325} = 38.13\%$ respectively. It should be noted that the denominators used in the calculations are the numbers of collocation constituents of the two corpora, because this aims to examine the proportion of the keywords making up collocations. It appears that the keyword coverage of the NECCD is significantly lower than that of the TECCTC. This can be extended to the comparison of the keywords and their collocates, with the results being $Keywords/Collocates \text{ } ratio_{NECCD} = \frac{535465}{1783491 - 535465} = 42.9\%$ and $Keywords/Collocates \text{ } ratio_{TECCTC} = \frac{689073}{1807325 - 689073} = 61.62\%$. In respect to this comparison, it can be concluded that the higher the value the smaller the variety of collocation types, because in the NECCD, on average, one keyword can collocate with

approximately 2.33 words (42.9%) while in the TECCTC one keyword can only collocate with approximately 1.62 words (61.62%). This result indicates that if Chinese translators produce a similar number of collocation tokens when compared with native speakers, their use of collocation types would be lower than that of native speakers. Therefore, Chinese translators' significantly high keywords-collocates ratio may be one of the factors leading to their reliance on more collocation tokens but less types in L2 English output. This finding can also help explain the difference in keyword types between the two corpora.

To test the validity of the above assumption, I cross-examined the keywords across the two corpora. In other words, I checked the NECCD keywords in the TECCTC vocabulary, and I checked the TECCTC keywords in the NECCD vocabulary. The result shows that 99.94% of the NECCD keywords occur in the TECCTC, and 100% of the TECCTC keywords occur in the NECCD. There is basically no difference between the two groups of speakers regarding the number of keywords in their working vocabulary. This indicates that Chinese translators master a keyword range as large as that used by native speakers but they fail to produce a collocation range as large as that used by native speakers. Specifically, Chinese translators in this study under-produced 320 keyword types to make up collocations in comparison with native norms. In other words, it can be assumed that Chinese translators know these 320 keywords but do not know how to combine them with other words to constitute collocations, or at least significant collocations under the criteria of this study (see 3.5.1). In order to further investigate this issue, this study will also use the concept of keyword growth to carry out a dynamic analysis of the keywords regarding Chinese translators' relative lack of collocation knowledge.

As stated above, keyword growth can be explained as the tendency for keyword types to increase over segmented text(s) of particular lengths in a corpus. Keyword growth is a very important notion in the area of English as a Foreign Language (EFL) teaching and learning, and there are a number of studies (e.g. Liu & Nation, 1985; Nation & Waring, 1997; Fan, 2006; Laufer & Ravenhorst-Kalovski, 2010) focusing on this topic. In the

present study, the analysis of keyword growth mainly looks at the dynamic process of how Chinese translators under-produce keyword types in using English collocations when compared with native speakers.

As stated in Section 4.4, the NECCD and the TECCTC are used as both the text base and the vocabulary source in this study, and were collected from the public domain of commercial discourse. In the section of lexical analysis, I used *Perl* (Practical Extraction and Report Language) programming to help carry out the analysis of keyword growth in the two corpora. Perl is another computer programming language, and was created by Larry Wall in the 1980s for natural text processing purposes. Similar to *FoxPro*, *Perl* is easy to obtain, versatile and powerful, featuring short commands and concise programmes. As Fan (2010b) noted, *Perl* is particularly useful for its regular expressions, “which greatly simplify complicated pattern matching in large texts or corpora”, and is powerful for “number crunching, that is, it can be used for math operations with efficiency” (p. 2). Therefore, *Perl* is suitable in this context to carry out lexical analysis in terms of the computation of keyword growth.

With *Perl*, I initially segmented each corpus into a chain of equal-sized texts of 2,000 word chunks respectively (see Appendix E for the *Perl* programme of text chunks segmentation). This is because such a text size can basically reflect the features in regard to stylistics, vocabulary and grammar of the language investigated (Kennedy, 1990). Thus, the NECCD and the TECCTC have been divided into 2,620 and 2,584 text files respectively. At the same time, I lemmatised the two lists of keywords across the two corpora (see Appendix F for the *Perl* programme of lemmatisation). This was not done in the previous analysis because different forms of the same lemma may produce different meanings, rather than mere inflectional distinction, and may produce lexical associates belonging to different categories, such as *bank* - *banking*, and *account* - *accounting*. However, the lexical analysis mainly looks at a certain group of language users’ ability to convey a particular semantic unit (keyword type) and associate it with other units, and for this reason all the keywords are lemmatised in the keyword growth analysis. Therefore, there are 1,501 and 1,219 keyword types across the NECCD and

the TECCTC after lemmatization.

For each corpus, the statistics of keyword types over each text file is computed in equally spaced intervals (see Appendix G for the *Perl* programme of the computation of keyword growth), and advances to include one more text file each time, starting from the top 2,000 words, followed by the top 4,000 words, the top 6,000, and so forth until the final text file is reached and included. Thus, keyword types at these intervals combine together to constitute a keyword growth pattern (normally demonstrated with a non-linear curve) along with the increase of keyword size. Because this task mentioned above involves calculations extremely large for human effort, I employed Köhler and Martináková-Rendeková's (1998) re-parametrized Torquist mathematical model and performed the model fitting with SPSS, which is shown as follows:

$$Y = \frac{a \times X}{b + X} + c$$

where Y : keyword coverage; X : keyword size; a , b and c : parameters. The results for the two corpora are demonstrated in Figure 5.6 and Figure 5.7 respectively:

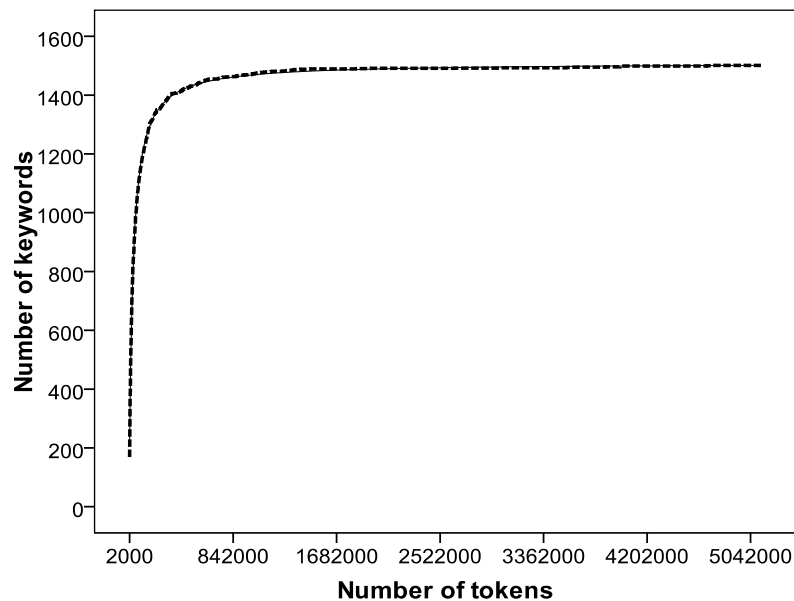


Figure 5.6 Keyword growth in the NECCD

Note: Determination coefficient: $R^2=0.9984$ (99.84%); a :1385.92804; b :30193.4249; c :123.550945. Solid line: model fit; solid squares: observed values.

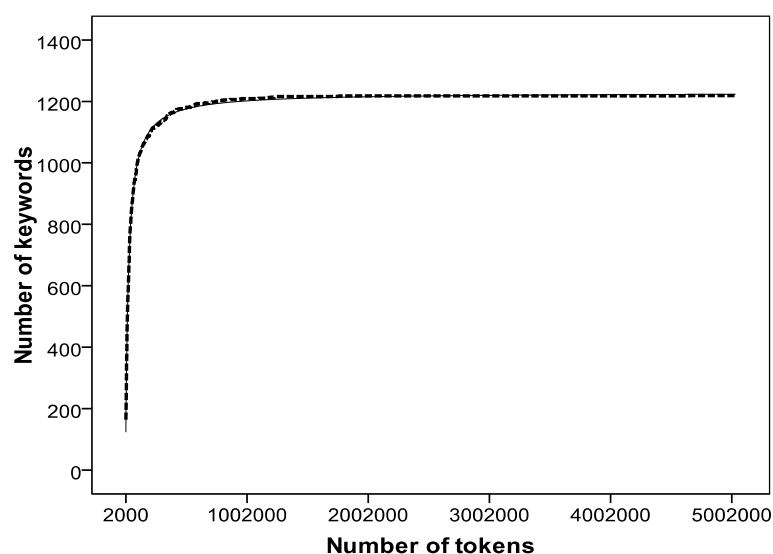


Figure 5.7 Keyword growth in the TECCTC

Note: Determination coefficient: $R^2=0.9959$ (99.59%); a : 1200.84782; b : 23191.8937; c : 28.1615658.
Solid line: model fit; solid squares: observed values.

As shown in Figures 5.6 and 5.7, the results computed with the re-parametrised Torquist model are, to a great extent, in line with the observed values, thus, representing a high degree of congruence. This means that the mathematical model used in this analysis is reliable enough to describe actual keyword growth data. The curves of keyword growth in these two figures both display a non-linear pattern, in which the number of keyword types increases rapidly and then starts levelling off. That is to say, these two curves generally seem to develop a similar tendency regarding keyword type growth. However, a closer comparison between these two sets of data may lead to the conclusion that the TECCTC curve levels off at a much lower point than that of the NECCD. This indicates that Chinese translators under-produced keyword types as a whole when compared with native speakers, and that Chinese translators' actual production of keywords types does not increase as fast as that of native speakers. Furthermore, there is a tendency for the difference between the two groups of speakers regarding keyword growth to become larger and larger as text size increases. Table 5.6 shows the top 20 groups of keyword growth data extracted partially from the NECCD and the TECCTC:

Table 5.6 Partial results of keyword growth data across the two corpora

Number of word tokens	NECCD Keyword types	TECCTC keyword types
2000	169	164
4000	289	191
6000	362	266
8000	405	342
10000	476	388
12000	537	449
14000	578	489
16000	617	513
18000	664	558
20000	698	568
22000	726	590
24000	752	611
26000	777	643
28000	810	668
30000	823	684
32000	842	721
34000	855	753
36000	872	769
38000	897	783
40000	915	791

It can be seen, from this table, that at each point of token size the corresponding value of the NECCD is always larger than that of the TECCTC. In addition, I extracted the data from the point of token size of 500,000, 1,000,000, 2,000,000, 3,000,000, 4,000,000 and 5,000,000. The results from the NECCD and the TECCTC are 1431-1179, 1472-1209, 1491-1218, 1492-1218, 1499-1218 and 1501-1219 respectively. These findings provide reliable evidence for the conclusion that, despite the size of text, Chinese translators' use of keyword types always falls behind that of native speakers, which, as a whole, results in their under-production of keyword types in general (1219

types against 1501 types).

Generally speaking, the slow keyword growth reflected in texts produced by Chinese translators appears closely related to the main factors influencing their production of L2 English collocations. This also, from another angle, supports and explains the foregoing finding (see 5.2.1.1) that Chinese translators appeared to rely heavily on the repetition of favoured collocation tokens but largely fail to produce a more diverse range of collocation types in their L2 English output. In this respect, researchers should also comment that, in translator training, translators who are handling the translation tasks between two, or even more, different languages should not only master a wide range of L2 vocabulary, but also realise how to associate keywords reasonably with other words to form formulaic sequences, such as collocations. This is particularly important for translations in a specific register. Only in such a way, can translators construct smoother texts in their L2 and come closer to native norms.

5.5 Summary

The present chapter has shown that, when compared with native speakers of English, the outputs by Chinese translators show a significant dependence on high-frequency strong collocations which mainly possess a frequency larger than 50 and an MI score greater than 4. This can be a direct factor resulting in the lack of balance of the type-token ratio in the corpus of translational English, which is largely due to the slow keyword growth rate in Chinese translators' production of L2 English collocations. The next chapter will continue to investigate the distinctive features of Chinese translators' use of English collocations in the commercial register, specifically from the three aspects: form, meaning and function.

Chapter Six Data analysis: Features of Chinese translators' use of English collocations in the commercial register (Part II)

6.1 Introduction

This chapter will continue to report on the results of the empirical investigation into different patterns of collocation use between the NECCD and the TECCTC, and will specifically look at the features of Chinese translators' use of English collocations from the perspectives of form, meaning and function. These features will also be examined through the comparison between the two corpora in terms of collocation distribution.

6.2 Formal features of collocation use in the corpus of translational English

As mentioned in Section 2.2.2, since collocation is regarded as a type of word combination (see for instance Cowie, 1981a; 1981b; Howarth, 1998a; Nesselhauf, 2005), one might ask how strong the association is between lexical items constituting collocations, that is, how likely one word tends to predict the co-occurrence of another. A number of researchers have suggested the models of continuum to clarify and classify the flexibility of collocations (see for instance Howarth, 1998; Sinclair, 1991; Wray, 2002), and these models often need to be constrained to one mode of description (Wray, 2002), such as form, meaning or function. In general, this so-called continuum as described in many models is normally recognized as a parameter (e.g. Pawley and Syder, 1983) to distinguish different kinds of collocations or a set of principles (e.g. Sinclair, 1991) in which one switches to another. This implies that the distinction of different types of collocations would, to a great extent, result in different patterns of collocation distribution between different groups of language users, especially between native speakers and L2 speakers. Therefore, this section will look at the formal features of Chinese translators' use of collocations to determine the formulaic pattern of collocation

distribution in their L2 English by comparing the two corpora.

As set out in Section 2.2.1, in the present study I carried out the investigation into the formal features of collocation use based on Sinclair's (1991) 'two principles', i.e. the open-choice principle and the idiom principle (see 2.3), and divided the collocations obtained from the two corpora into three main groups regarding the level of association, that is, free combinations, bound collocations and idioms. In this study, I used a 'bottom-up' approach in the classification procedure, in which I first picked out idioms from the two collocation lists, followed by bound collocations, and free combinations respectively. The results from the NECCD and the TECCTC with regard to the proposed formal classification of collocations are shown in Table 6.1:

Table 6.1 General information of collocations regarding form

	Free combinations		Bound collocations		Idioms	
	tokens	types	tokens	types	tokens	types
NECCD	23,498	993	74,022	5,197	4,415	176
	23.05%	15.6%	72.62%	81.64%	4.33%	2.76%
TECCTC	51,195	1,951	56,848	1,819	3,407	102
	45.94%	50.39%	51.01%	46.98%	3.05%	2.63%

The result shows that the patterns of collocation distribution in terms of formal features, namely, free combinations, bound collocations and idioms, appear to differ greatly between the NECCD and the TECCTC. The comparison of tokens aims to examine the proportions of different kinds of collocations which language users employ to render English texts. In this respect, language users' tendency to use a particular group of collocations can be a reliable index to show how formulaic their use of English is. The comparison of types looks at how Chinese translators diversify their output of collocations in relation to the type they mainly rely on. This also shows the distinctive features of textual construction in language output, and is used as an indicator to make a distinction between different groups of language users.

As demonstrated in Table 6.1, in terms of collocation token, both Chinese translators and native speakers of English employ bound collocations as their main type of collocation production in the commercial register. Bound collocations amount to 74,022 and 56,848 in the NECCD and the TECCTC, and contribute to 72.62% and 51.01% respectively. In contrast, a comparatively low percentage of idioms appears to be significant in both corpora, with the NECCD showing 4.33% (4,415 types) and the TECCTC 3.05% (3,407 types). This indicates that the construction of business English texts requires largely collocations with ‘restricted flexibility’, such as technical terms and the terminologies that have special connotations in the commercial register. Nonetheless, as noted above, there is still a clear difference between the two groups of language users regarding the distribution of collocation tokens, as shown in Figure 6.1:

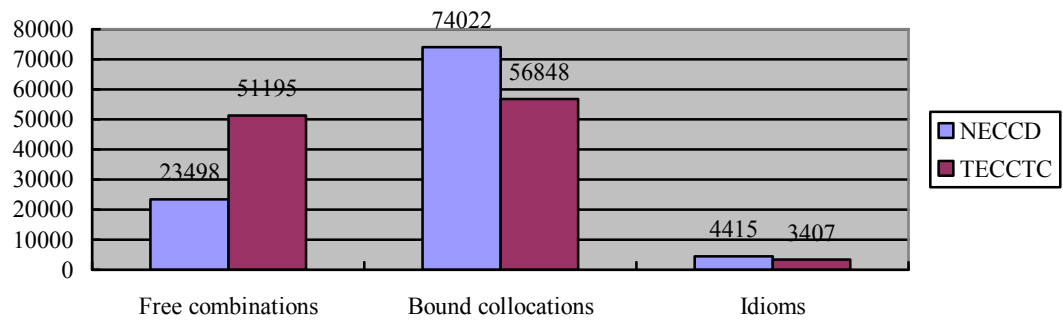


Figure 6.1 Comparison of collocation tokens regarding formal classification

It is clear, from Table 6.1, that Chinese translators used more free combinations but less bound collocations and idioms when compared to native speakers. In particular, there appear to be striking differences regarding free combinations and bound collocations. In order to ensure the reliability of the comparison, I also examined these three groups of data with Chi-square tests, with the results shown in Table 6.2:

Table 6.2 Chi-square tests for comparing tokens of free combinations, bound collocations and idioms between the two corpora

Free combinations

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.225E4 ^a	1	.000	.000	.000
Continuity Correction ^b	1.225E4	1	.000		
Likelihood Ratio	1.249E4	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	1.225E4	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35681.19.

b. Computed only for a 2x2 table

Bound collocations

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.048E4 ^a	1	.000	.000	.000
Continuity Correction ^b	1.048E4	1	.000		
Likelihood Ratio	1.063E4	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	1.048E4	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 39417.80.

b. Computed only for a 2x2 table

Idioms

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.448E2 ^a	1	.000	.000	.000
Continuity Correction ^b	244.427	1	.000		
Likelihood Ratio	244.758	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	244.787	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 3736.61.

b. Computed only for a 2x2 table

At one degree of freedom, the Chi-square value is 1.225e4 for free combinations, 1.048e4 for bound collocations and 244.8 for idioms, with each value being greater than the critical value 3.84 (see for instance Manning and Schütze, 1999), which therefore shows statistical significance in each category. This finding indicates that Chinese translators under-produced bound collocations by 17,174 tokens in comparison with

native speakers. According to the comparison above, bound collocations are obviously the primary constituents in native use of collocations. Native speakers form a great number of language chunks with bound collocations while producing business English discourse. These fixed language chunks can help language users reduce the effort processing mental language information and increase the efficiency of producing smooth and fluent texts. In other words, to achieve native-like rendition of business English, language users' production of bound collocations would have to cover at least 72.62% of the total number of collocations generated. In this respect, Chinese translators' production of bound collocations, in the present investigation, did not match the production of bound collocations that were demonstrated in the corpus of native English. From another aspect, one may speculate that these Chinese translators, to some extent, broke these language chunks into 'small viable units' that might in turn decrease their ability to possess vast knowledge about how to transfer the possibilities of word sequences into their formation of linguistic competence. This can also be the main reason why they over-produced free combinations, but used fewer bound combinations, when compared with native speakers. The use of free combinations may also increase the possibility of mutual lexical choices. In free combinations, the 'small viable units', such as words, are always associated with the context of the situation and become less restricted in mutual selection due to lexical polysemy. Therefore, these units might not combine into language chunks which are supposed to be formulaic in a particular register due to language users' insufficient knowledge of collocation. This would definitely decrease the possibility of language users' production of bound collocations. The formal features exhibited in the TECCTC can be a typical example. In translation practice, translators may be, at times, more inclined to explain the source text by employing the 'word for word' strategy, especially when there is no linguistic equivalence between a particular language pair. In such a situation, translation units, which are supposed to be as formulaic as possible, are constituted largely by free combinations which closely reflect the source text. This may be a major factor resulting in translation universal features (see 5.3) and the foreign sounding nature of the translational language.

I also examined the distribution of types in an attempt to identify more formal features of Chinese translators' use of L2 English collocations. As demonstrated in Table 6.1, bound collocation types amount to 5,197 and 1,819 in the NECCD and the TECCTC, which contribute to 81.64% and 46.98% of the total number of collocation types respectively. In contrast, free combinations amount to 993 and 1,951 in the NECCD and the TECCTC, and make up 15.6% and 50.39% of the total number of collocation types respectively. The percentage of idioms used is still low in both corpora, with the NECCD being 2.76% (176 types) and the TECCTC 2.63% (102 types). The results are illustrated in Figure 6.2:

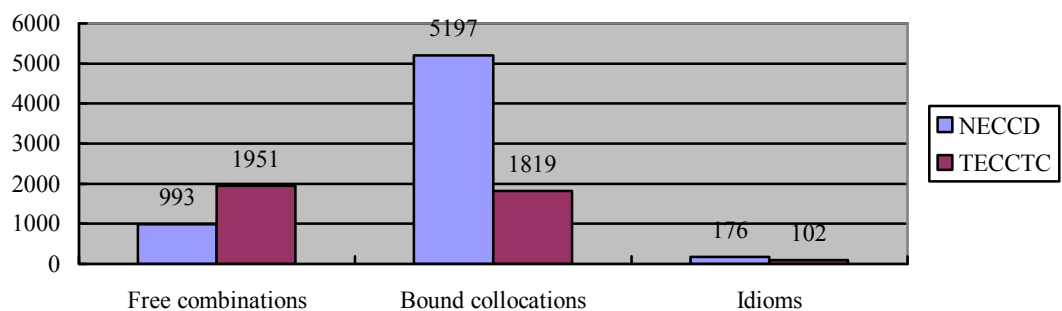


Figure 6.2 Comparison of collocation types regarding formal classification

The difference between the two groups of speakers appears to be more striking in the comparison of classified collocation types than in that of tokens. Nevertheless, to ensure the accuracy of the comparison, I examined these three groups of data individually with Chi-square tests. The results are presented in Table 6.3:

Table 6.3 Chi-square tests for comparing types of free combinations, bound collocations and idioms between the two corpora

Free combinations

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.422E3 ^a	1	.000	.000	.000
Continuity Correction ^b	1.421E3	1	.000		
Likelihood Ratio	1.405E3	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	1.422E3	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1113.42.

b. Computed only for a 2x2 table

Bound collocations

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.341E3 ^a	1	.000	.000	.000
Continuity Correction ^b	1.339E3	1	.000		
Likelihood Ratio	1.328E3	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	1.341E3	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1218.56.

b. Computed only for a 2x2 table

Idioms

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.155 ^a	1	.694	.706	.370
Continuity Correction ^b	.110	1	.741		
Likelihood Ratio	.156	1	.693		
Fisher's Exact Test					
Linear-by-Linear Association	.155	1	.694		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 105.14.

b. Computed only for a 2x2 table

It should be noted that idioms contribute to a very minor proportion of collocation types overall, accounting for a mere 2.76% in the NECCD and 2.63% in the TECCTC respectively. Therefore, even though the test regarding idioms does not show statistical significance, this will not influence the overall comparison. The Chi-square value is

1.422e3 for free combinations and is 1.341e3 for bound collocations, both of which are greater than the critical value 3.84 (see for instance Manning and Schütze, 1999), thus showing statistical significance.

In terms of type, Chinese translators produced more free combinations than bound collocations, which generally means that they rely mainly on free combinations as their strategy for producing collocations. This is in stark contrast with the scenario of native English speakers whose production of bound collocations (5,197 types) is more than five times that of free combinations (993 types). This means that native speakers of English use a large variety of bound collocation types to build up texts in the commercial register. In other words, to achieve native-like output of business English, at least 81.64% of the collocation types produced by language users would have to be formulaic and structurally restricted, such as technical terms. Moreover, bound collocations produced by native speakers achieve a more diverse range of language expressions. This can be simply exemplified with the scenario regarding the production of bound collocation types in this study. Chinese translators produced 1,819 bound collocation types, which is approximately one third (as against 5,197) of what native speakers normally use in the commercial register. This indicates that Chinese translators over-used free combinations to render texts, which will decrease the level of formulaic language in the target language. This may also increase the possibility of translators bringing some translation universal features (or translation universals), particularly simplification (see 7.2.1), into their translational English.

Generally speaking, researchers in this area should realise that, based on the different formal collocation patterns produced by different language users, they can help L2 learners working as translators to develop their collocational English. In this study, the formal pattern of collocation distribution in translational language has shown Chinese translators' weakness in comparison with native norms. In order to achieve native-like rendition of the target language (L2), translators would not only need to recognise a collocation as a unit, rather than further breaking it into smaller viable units, but also need to memorise a large repertoire of collocations and access these collocations in

context, rather than repeatedly using their favoured high-frequency word combinations. Only in such a way can translators keep their L2 language formulaic to a large extent (see also Toury, 1980) and increase language proficiency, which, as a whole, makes their use of collocations come closer to native standards.

6.3 Semantic features of collocation use in the translational corpus

As documented in a number of studies (e.g. Hudson & Francis, 2000; Sinclair, 1991, 2004; Stubbs, 2001, 2005), collocations can be identified as “extended units of meaning” and are “expected to be largely phrasal” (Sinclair, 2004, p. 30). In line with Sinclair’s viewpoint, Hudson and Francis (2000) also propose that most words do not make sense unless they are associated with a particular pattern. It should be noted that pattern in their claim means a syntactic-semantic whole that makes no distinction between form and meaning. In some sense, collocation realises the meanings of individual words. Stubbs (2001) notes similarly that, “it makes little sense to describe the meaning of individual words in isolation, since words are co-selected with other words, and meanings are distributed across larger units” (p. 100). It appears that these researchers try to establish associations between pattern and meaning. According to Stubbs (2001), these insights become the key section of phraseology theory and provide a theoretical foundation for examining the semantic features of collocations for subsequent studies. Since the meanings of words are strongly associated with the patterns they constitute, then the way of formulating the patterns becomes a dominant factor governing the meanings. In this respect, the analysis of semantic features can be a reliable indicator to examine and compare the use of collocations, and help identify a distinction between different groups of language users.

Therefore, this study employs the aforementioned strategies (see 3.5.3), specifically, “from n-grams to content” and “from lexis to co-text” (see Stubbs, 2005), to examine the respective semantic features generated from a native corpus and a translational corpus. The “from n-grams to content” strategy allows me to examine the semantic

features of a particular collocation or formulaic pattern and further explain the distribution of the major senses of this word by employing comparable corpora. The “from lexis to co-text” strategy allows me to investigate how the patterns of meanings are realised through collocations or formulaic patterns in comparable corpora. This will also provide more opportunities to discover what kind of patterns of meanings are realised by Chinese translators through the use of English collocations, as well as whether there is significant difference between translational English and native English from the semantic perspective. Thus, with regard to these two strategies, this study employs the notion of “delexicalization” (see Stubbs, 2001) as a measure to investigate the semantic features of collocations. Delexicalization refers to a process in which the logical core meaning of a word has ceased to function in language formulae or ceased to be the most important, such as *take* in *take a bite* (see Stubbs, 2001). Delexicalization occurs not merely in words but also in larger lexical units, such as collocations. Based on this notion, I established two criteria for distinguishing these two kinds of senses:

- a. either constituent of a collocation retrieved from the two corpora cannot exist independently, then this collocation is used in the delexicalized sense, such as *make* and *decision*;
- b. the literal sense of either constituent in a collocation is changed or even lost in the commercial register, then this collocation is used in the delexicalized sense, such as *zip* and *code*.

The results of the semantic classification from the NECCD and the TECCTC according to the proposed criteria are shown in Table 6.4:

Table 6.4 General information of collocations with regard to meaning

	Literal sense		Delexicalized sense	
	tokens	types	tokens	types
NECCD	84,506	5,647	17,429	719
	82.9%	88.71%	17.1%	11.29%
TECCTC	103,057	3,640	8,393	232
	92.47%	94.01%	7.53%	5.99%

These results show that collocations used in the literal sense make up a very high percentage in both corpora, with the NECCD being 82.9% in token and 88.71% in type, and the TECCTC being 92.47% in token and 94.01% in type. This indicates that collocations with a literal sense contribute to the majority of all the collocations produced in both translational and native-speaker business English. This may be because business English requires language users to produce fewer words mostly used in the delexicalized sense, such as *take* (see for instance Stubbs, 2001), *make* (see for instance May Fan, 1999) and *thing* (see for instance Deng, 2007), while they are producing collocations. For instance, in the NECCD, *take* occurs 14 times in a delexicalized sense and *make* a mere 8 times. This is a very low ratio when compared to other studies (e.g. Stubbs, 2001) of formulaic language in general use. Therefore, delexicalization may be less pervasive when it comes to the use of collocations in the commercial register as compared to general use. In both the NECCD and the TECCTC, the constituents of the collocations with a delexicalized sense can be classified into the following types:

- a. words that cannot form a meaning in isolation (e.g. *out* in *carry out* and *up* in *rack up*);
- b. delexicalized words, (e.g. *keep* in *keep track/updated/organised*, *take* in *take place* and *make* in *make sense*);
- c. words used in figurative meanings, or used to form an idiomatic sequence, (e.g. *chain* in *supply chain*, *giant* in *media giant*, *green* in *green jobs*, *bull* in *bull market*, *ceiling* in *ceiling price* and *zip* in *zip code*);

- d. other words deviate from their literal sense to help realise the meaning of collocations they constitute, (e.g. *real* in *real estate*, *slide* in *slide show* and *run* in *run counter*).

Therefore, it appears that these types of collocations, to a great extent, help maintain the idiomaticity and formulaicity in the native use of language because words can not ‘stand alone’ at times but can only realise their senses through lexical chunking. In this way, the use of delexicalized collocations remains an important indicator to distinguish different groups of language users in relation to language proficiency because meanings are conveyed differently through different word strings across users at different language proficiency levels (see Deng, 2007; May Fan, 1999; Stubbs, 2001). Therefore, delexicalization is more pervasive than expected in language use (May Fan, 1999) and can expand the collocational ranges of words, allowing them to express different meanings in a more flexible way. Thus, through delexicalization the pragmatic meanings of words can become strengthened (Stubbs, 2001). The results in Table 6.4 also show that the collocation distribution in terms of semantic features, namely the literal sense and the delexicalized sense, appears to be different between the NECCD and the TECCTC. In business English, native speakers normally use 82.9% of collocations in the literal sense, which amounts to 84,506 tokens in the NECCD, and they use 11.29% of collocations in the delexicalized sense, which amounts to 17,429 tokens. In comparison with native norms, Chinese translators used 103,057 collocation tokens in the literal sense, thus accounting for 92.47%, and they used 8,393 tokens in the delexicalized sense, accounting for a mere 7.53%. This can be illustrated in Figure 6.3:

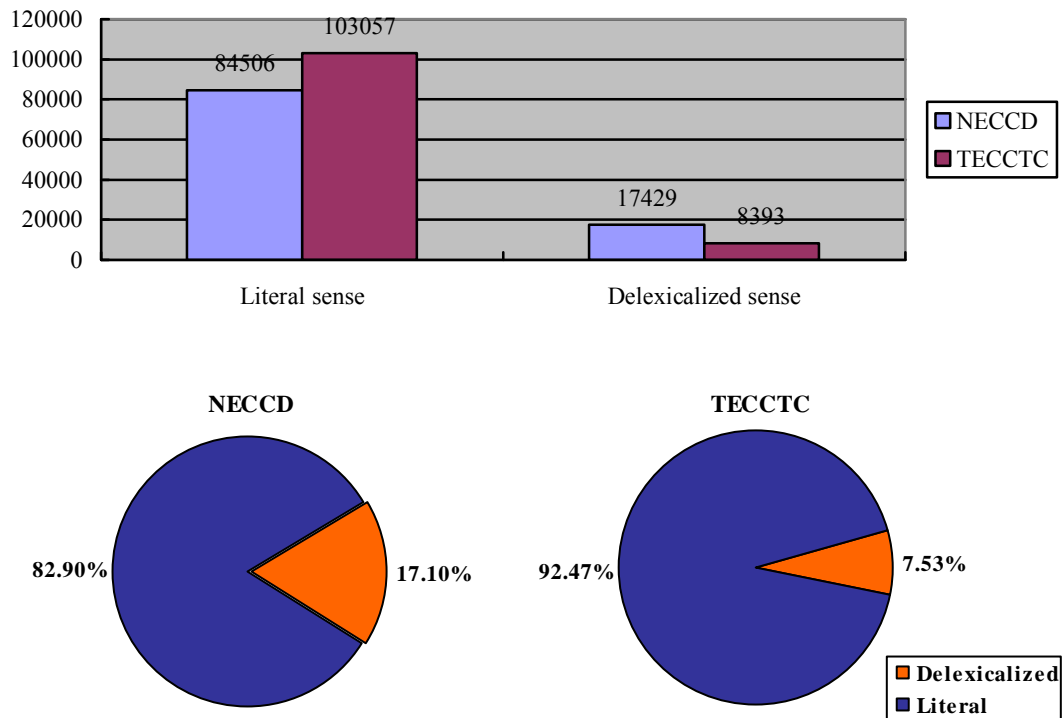


Figure 6.3 Comparison of collocation tokens regarding semantic features

The Chi-square tests for comparing collocation tokens used in their literal sense or the delexicalized sense also demonstrate that the difference between these two groups of data is significant. The statistical results are shown below in Table 6.5:

Table 6.5 Chi-square test results for comparing collocation tokens with regard to semantic features

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.582E3 ^a	1	.000		
Continuity Correction ^b	4.581E3	1	.000		
Likelihood Ratio	4.643E3	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	4.582E3	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12335.29.

b. Computed only for a 2x2 table

The obtained Chi-square value is 4.582e3, which is greater than the critical value 3.84 at 95% percent of confidence (see for instance Manning and Schütze, 1999). It should be noted that the test results for the literal sense and the delexicalized sense are identical because these two senses are complementary in terms of semantic features. Therefore,

there is only one table displayed in Table 6.5.

In terms of collocation type, the difference between the two corpora appears to be more striking. Native speakers of English produced 5,647 collocation types with a literal sense, which account for 88.71%, and produced 719 types with a delexicalized sense, which account for 11.29%. In comparison with native speakers, Chinese translators used 3,640 collocation types used in a literal sense, which make up 94.01% of all types, and used 232 types used in a delexicalized sense, which only make up 5.99%. The comparison between collocation types demonstrates a different scenario because native speakers obviously used more types than Chinese translators in either of these two senses. In addition, the proportion of collocations with a delexicalized sense in native English amounts to nearly twice as large as that in translational language. This result can also be illustrated in Figure 6.4:

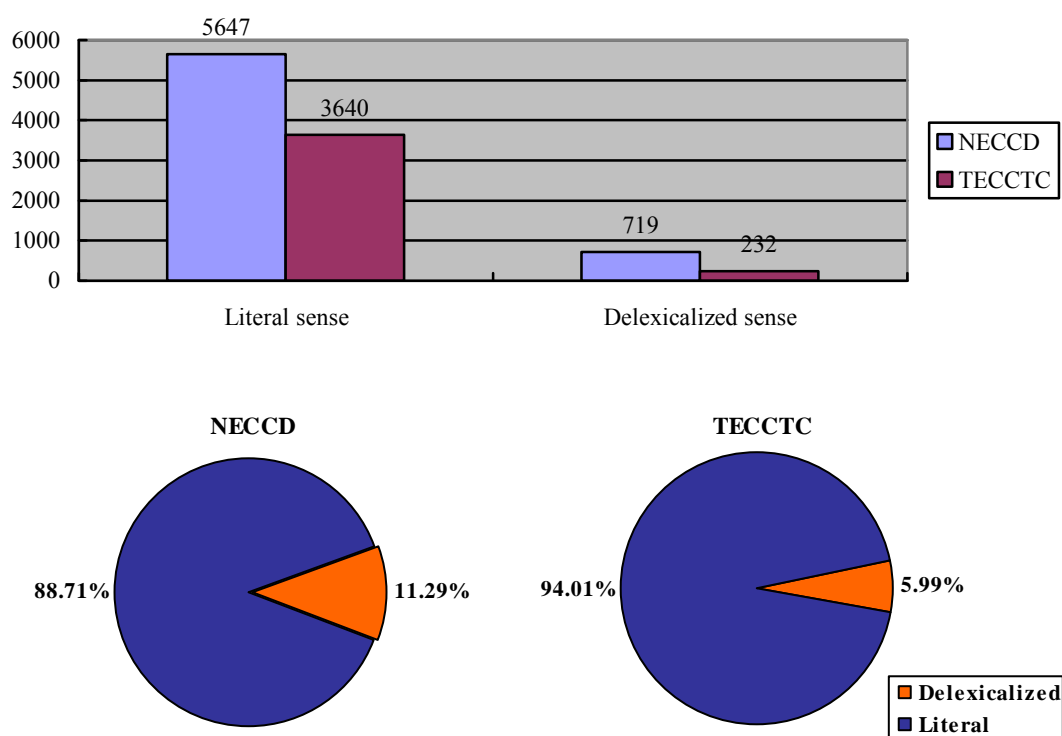


Figure 6.4 Comparison of collocation types regarding semantic features

The Chi-square tests for comparing collocation types of the literal sense and the delexicalized sense also demonstrate that the difference between these two groups of

data is significant. The statistical results are shown in Table 6.6 below:

Table 6.6 Chi-square test results for comparing collocation types with regard to semantic features

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	80.343 ^a	1	.000	.000	.000
Continuity Correction ^b	79.714	1	.000		
Likelihood Ratio	85.095	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	80.335	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 359.67.

b. Computed only for a 2x2 table

The obtained Chi-square value regarding the comparison of types is 80.343, which is larger than the critical value 3.84 at 95% percent of confidence (see for instance Manning and Schütze, 1999).

Therefore, Chinese translators under-used delexical collocations when compared with native speakers of English, which holds true both in terms of token and in terms of type. This finding may indicate that, on the one hand, when translators accumulate their L2 vocabulary knowledge they are more inclined to focus on the literal side of words but ignore largely the ‘depth’ side of exploring the pragmatic meanings of words, which, as a whole, may contribute to them using a narrower range of collocations in comparison with native norms. When some researchers (e.g. Hudson & Francis, 2000; Stubbs, 2001) argue that words make little sense in isolation, they mean that words possess not only literal meaning or logical meanings but also pragmatic meanings which help realise the formation of larger lexical chunks, such as collocations. This is also the reason why these scholars argue at times that words are not independent units of meaning at times. In this respect, it is not enough for L2 language learners to learn words merely from the individual meanings defined in dictionaries because words may be delexicalized at times when they co-occur with other words so they may generate more pragmatic senses in the collocational relationship. When translators are handling translation tasks, they might be confronted with the situations where there is no linguistic equivalence and

particularly the situations where they simply do not realise that certain collocations take on a different pragmatic meaning in certain contexts, and then they may mistranslate these collocations. What they are translating might possess the meanings that they have not yet learnt or that are not defined clearly in general dictionaries. In this case, the employment of the delexicalization strategy might be a valid solution to the occurring problems. This will not only avoid translators segmenting semantic units mistakenly when analysing the source text but also help enable them to look at the collocational relationships between words and render the target text more accurately in accordance with native norms. In this respect, how to realise the importance of delexicalization and apply it to translation practice appears to be a crucial task when translators acquire collocation knowledge.

On the other hand, delexicalization concerns language users' efficiency and proficiency in producing L2 collocations. Since delexicalization is a process which makes the meanings of some particular lexical items implicit to help construct larger semantic units, it will inevitably take account of contextual determination based on a reasonable knowledge system of lexicon and grammar. This kind of knowledge system can be a determining factor in terms of language users' proficiency. As Wray (2002) notes, native speakers "take for granted that certain expressions are so common as to be elementary", whereas L2 language users "cannot know them unless they have actually encountered them before" (p. 182). This indicates that some developing translators may not have sufficient knowledge regarding L2 collocation input, and their employment of delexicalized meanings for collocations might be constrained to some extent. With respect to this, translators might increase the possibility of making explicit the information that is supposed to be implicit according to native norms. This can be exemplified with Farghal and Obiedat's (1995) investigation in which language users with an L1 Arabic background tend to employ some strategies, such as paraphrasing, in their Arabic to English translations to make up for their insufficient knowledge of English collocations.

In the present study, Chinese translators' significant over-production of collocations

with a literal sense appears to reflect their insufficient knowledge of English collocations. Thus, they may have brought some translation universal features (or translation universals), especially explicitation, into the target text they produced. This will be examined and analysed in more detail in Section 7.2.2. As explained in the theoretical framework of this study, translation universals appear to run counter to a native-like production of language. That is to say, the more these features are present in the text the less natural the resulting text. In this sense, insufficient knowledge of delexicalized meanings in forming collocations may have become an obstacle for Chinese translators to reach native speakers' proficiency level. One plausible way of improving translators' use of collocations with delexicalised meanings is, as Wray (2002) argues, to make them pay more attention to the collocational relationships which lead to delexicalization.

Therefore, the design of a feasible model incorporating semantic perspectives for translators in their L2 English input appears to be another significant issue worthy of consideration. This may also be the key to resolving the issue of increasing L2 translators' proficiency in using collocations and producing more native-like target language. Chapter 8 will outline some possible recommendations.

6.4 Functional features of collocation use in the translational corpus

As set out in Section 2.2.4, collocations are found to be strongly associated with communication situations in a great number of studies (e.g. Aijmer, 1996; Coulmas, 1981; Cowie, 1988; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Widdowson, 1989; Wray, 2002; Yorio, 1980), and are therefore defined as form-function composites and assigned functional meanings. In particular, the knowledge of collocations plays an important part in building up language users' communicative competence in social interactions with regard to how language should be used 'correctly' in 'correct' places. In this respect, when second language (L2) learners accumulate their L2 knowledge, it is important to identify the functional features of collocations when they are trying to

learn a large repertoire of collocations. However, this appears to bring difficulties to some L2 learners (see for instance Wray, 2002), as well as some developing translators who deal with translation tasks between two or more languages. Therefore, it is necessary to examine from translators' outputs to see whether there is any deviation in terms of collocation use from the functional perspective.

This study has shown that semantic prosody can serve as an indicator to investigate the functional features of Chinese translators using English collocations in the commercial register in comparison with native norms. In respect to this, I attempted to investigate different semantic prosodies in collocations by comparing the quantitative data from the two corpora. In such a way, I can examine whether unconventional English collocations in native commercial English are still used to perform the desired functions in translational commercial English. I classified collocations obtained from the two corpora into three main categories:

- a. collocations with positive semantic prosodies, and those which contain words with favourable or positive affective meanings (e.g. *achieve success*, *acquire rights*, *collaborative efforts*, *enhance ability* and *grant funds*);
- b. collocations with neutral semantic prosodies, and those which contain words with neutral meanings (e.g. *zip code*, *web domain*, *domestic product*, *mobile phone*, *stock indexes* and *price tag*);
- c. collocations with negative semantic prosodies, and those which contain words with unfavourable or negative affective meanings (e.g. *disciplinary action*, *afford payment*, *risk assessment*, *bail out* and *suffer losses*).

According to the proposed classification of collocations regarding semantic prosody, the results obtained from the NECCD and the TECCTC are shown in Table 6.7:

Table 6.7 General information of collocations regarding semantic prosody

	Positive		Neutral		Negative	
	tokens	types	tokens	types	tokens	types
NECCD	26,560	1,701	56,210	3,300	19,165	1,365
	26.06%	26.72%	55.14%	51.84%	18.8%	21.44%
TECCTC	19,933	831	78,932	2,477	12,585	564
	17.89%	21.46%	70.82%	63.97%	11.29%	14.57%

This result shows that collocations with neutral semantic prosodies make up a very high percentage in both corpora, with the NECCD being 55.14% in token and 51.84% in type, and the TECCTC being 70.82% in token and 63.97% in type. In other words, collocations with neutral semantic prosodies contribute to the majority of all the significant collocations in the commercial register, be they in translational English or native English. In addition, this result also shows different collocation patterns between the two corpora because Chinese translators significantly under-produced collocations with positive and negative semantic prosodies. In terms of token, the result of comparison is illustrated in Figure 6.5:

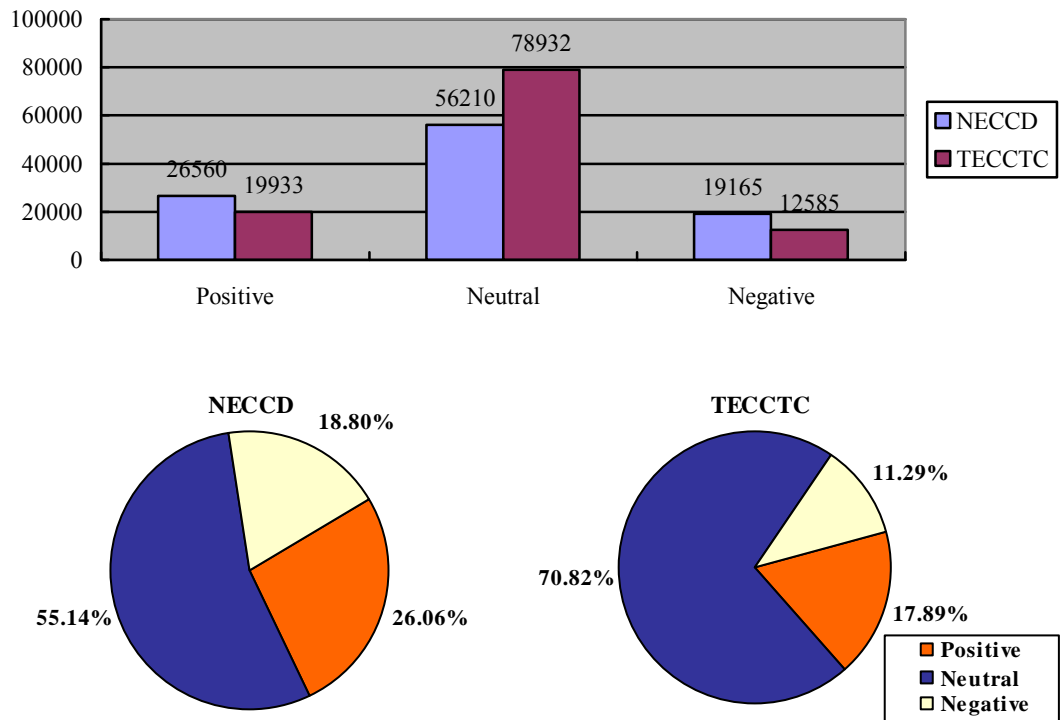


Figure 6.5 Comparison of collocation tokens regarding semantic prosodies

This graph clearly shows that, when compared with native speakers, Chinese translators depend more on collocations with neutral semantic prosodies (78,932 tokens), which account for 70.82%, but under-produce collocations with positive semantic prosodies (19,933 tokens), which account for 17.89%, or collocations with negative semantic prosodies (12,585 tokens), which account for only 11.29%. By contrast, native speakers of English rely, to a greater extent, on collocations with positive semantic prosodies (26,560 tokens), which account for 26.06%, and collocations with negative semantic prosodies (19,165 tokens), which account for 18.8%. These two categories amount to 45,725 tokens, which constitute nearly half of the whole collocations in the NECCD. To carry out a reliable comparison, I conducted Chi-square tests for each category, with the results being in the following table:

Table 6.8 Chi-square test results for comparing collocation tokens with regard to semantic prosodies

Positive semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.086E3 ^a	1	.000	.000	.000
Continuity Correction ^b	2.085E3	1	.000		
Likelihood Ratio	2.087E3	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	2.086E3	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22209.92.

b. Computed only for a 2x2 table

Neutral semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.637E3 ^a	1	.000	.000	.000
Continuity Correction ^b	5.636E3	1	.000		
Likelihood Ratio	5.654E3	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	5.637E3	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 37377.04.

b. Computed only for a 2x2 table

Negative semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.370E3 ^a	1	.000	.000	.000
Continuity Correction ^b	2.370E3	1	.000		
Likelihood Ratio	2.377E3	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	2.370E3	1	.000		
N of Valid Cases ^b	213385				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15167.12.

b. Computed only for a 2x2 table

The obtained values from the Chi-square tests with regard to positive, neutral and negative prosodies are 2,086, 5,637 and 2,370 respectively, all of which are greater than the critical value 3.84 (see for instance Manning and Schütze, 1999). This indicates that the difference between the collocation token patterns of Chinese translators and that of

native speakers of English is statistically significant with regard to semantic prosodies.

In terms of collocation type, the difference appears to be more distinct after comparing the two groups of data, with the results being demonstrated in Figure 6.6:

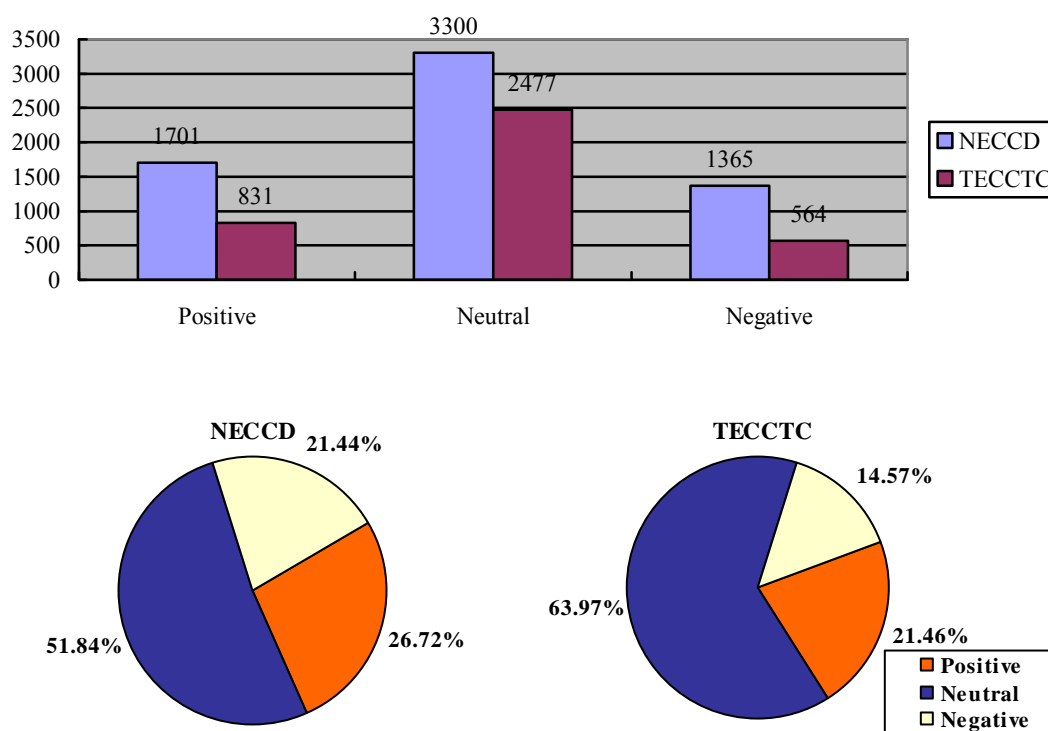


Figure 6.6 Comparison of collocation types regarding semantic prosodies

As shown in Figure 6.6, in each of the three categories investigated, that is, positive, neutral and negative semantic prosodies, the NECCD obviously demonstrates a higher value than the TECCTC. From the aspect of proportion, compared with native speakers Chinese translators still depend more on collocation types with neutral semantic prosodies (2,477 types), which account for 63.97%, yet under-produced collocations with positive semantic prosodies (831 types), which account for 21.46%, or collocations with negative semantic prosodies (564 types), which account for 14.57%. By contrast, native speakers of English rely, to a lesser extent, on collocations with neutral semantic prosodies (3,300 types), which only make up 51.84%. This is nearly equal to the combination of the percentage of positive semantic prosodies (26.72%) and that of negative semantic prosodies (21.44%). Again, to ensure the reliability of the comparison,

I conducted Chi-square tests for each category, with the results being in Table 6.9:

Table 6.9 Chi-square test results for comparing collocation types with regard to semantic prosodies

Positive semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	35.761 ^a	1	.000	.000	.000
Continuity Correction ^b	35.480	1	.000		
Likelihood Ratio	36.242	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	35.758	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 957.60.

b. Computed only for a 2x2 table

Neutral semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.442E2 ^a	1	.000	.000	.000
Continuity Correction ^b	143.687	1	.000		
Likelihood Ratio	145.395	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	144.166	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1687.15.

b. Computed only for a 2x2 table

Negative semantic prosody

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	74.439 ^a	1	.000	.000	.000
Continuity Correction ^b	73.990	1	.000		
Likelihood Ratio	76.578	1	.000		
Fisher's Exact Test					
Linear-by-Linear Association	74.431	1	.000		
N of Valid Cases ^b	10238				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 729.55.

b. Computed only for a 2x2 table

As demonstrated in Table 6.9, the obtained values from the Chi-square tests with regard to positive, neutral and negative prosodies are 35.761, 144.2 and 74.439 respectively, all of which are greater than the critical value 3.84 (see for instance Manning and Schütze, 1999). This indicates that the difference between the collocation type patterns of

Chinese translators and that of native speakers of English is significant from the statistical perspective. In addition to the foregoing statistical analyses, I also conducted a correspondence analysis to capture a clear picture of how the collocation tokens and types from the two corpora are correlated with the three categories of semantic prosodies. Correspondence analysis, as one of the data reduction procedures, is a multivariate statistical technique which applies to categorical data and can describe the relationships between two nominal variables. Correspondence analysis can present the relationships between the categories for each variable by summarising the obtained data in a two-dimensional diagram. In this case, the three categories of semantic prosodies and the corpus groupings can be regarded as two variables for the correspondence analysis. Thus, with SPSS, the correspondence analysis was carried out with the result being demonstrated in Figure 6.7:

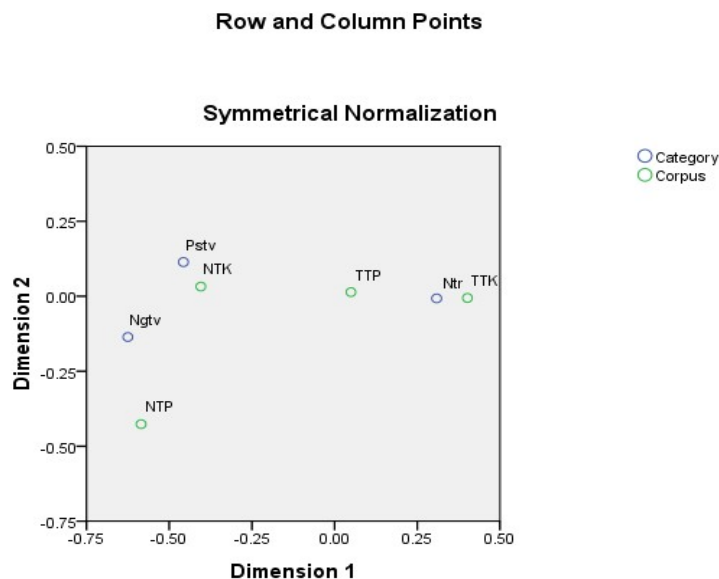


Figure 6.7 Correspondence analysis regarding semantic prosodies between the two corpora

Note: Pstv = Positive semantic prosody; Ntr = Neutral semantic prosody; Ngtr = Negative semantic prosody; NTK = Collocation tokens in the NECCD; NTP = Collocation types in the NECCD; TTK = Collocation tokens in the TECCTC; TTP = Collocation types in the TECCTC.

Figure 6.7 demonstrates a relationship of ‘intimacy’, in which the variables that are strongly correlated come closer to each other while those that are less associated stay further apart. Therefore, it is clear from this correspondence analysis, that native speakers of English rely more on collocations with positive and negative semantic

prosodies in business English (in Figure 6.7, both NTK and NTP are closer to Pstv and Ngvtv than to Ntr). In contrast, Chinese translators appeared to be more inclined to depend on collocations with neutral semantic prosodies in translational business English (in Figure 6.7, both TTK and TTP are closer to Ntr than to Pstv or Ngvtv). This result not only adds to the foregoing analyses, but may also give us some insight as to what might be causing the deviation in Chinese translators' use of L2 English collocations from the functional perspective. The fact that Chinese translators used more neutral collocations is, to a large extent, against native norms in the commercial register. In some sense, Chinese translators did not use 'correct' English collocations in the 'correct' places in their Chinese-to-English translations.

The functional features of Chinese translators' collocational patterns can be formulated from the following two aspects. On the one hand, Chinese translators repeatedly used some particular collocational patterns, which can be a major factor contributing to the distribution imbalance of the three categories of semantic prosodies. On the other hand, Chinese translators' comparatively weak control of the semantic prosodies of certain words can be another factor leading to the imbalance of the three categories of semantic prosodies. As a result, Chinese translators over-conformed to some typical English collocation patterns from the functional perspective, which would make their use of L2 English largely normalised. This will be discussed in more detail in Section 7.2.3.

The important relationship between collocation patterns and semantic prosodies should be well recognised in both language learning and language teaching, particularly translator training as discussed in this study. The use of lexical items with inappropriate semantic prosodies to constitute collocations would make it more possible to produce foreign-sounding language use. To learn collocations in a particular language, or to translate collocations between a particular language pair, is not only to understand what they stand for, but also to learn how they should be used and in what particular situations they should be selected. In this respect, researchers in this area should not only obtain meaningful findings as empirical evidence from comparative and contrastive analyses, but more importantly, they should also construct feasible

theoretical models based on the empirical evidence to help translators build up a sound knowledge of English collocations and essentially enhance their ability of using these collocations more appropriately. Only in such a way can translators render the target texts more faithfully and smoothly with their L2 English and fulfil the expected functions in communicative translation activities. A number of studies have already established useful models (e.g. Harley, 1996; Tognini-Bonelli, 2001; Xiao & McEnery, 2006) with regard to the confident use of semantic prosody in L2 operations. Xiao and McEnery (2006) suggest that it is important to “show learners which synonymous item in an L1 most closely matches which in an L2”, and that one of the best ways to achieve this is “properly sorted KWIC (keyword-in-centre) [key word in context] concordances, as these allow the learner to observe repeated patterns and meanings, and thus help them to become aware of collocation and semantic prosody” (p. 126). Crezee (2013) takes register and context into consideration, and believes that the best way to prepare for translating in a particular area is to actually complete studies in that area. As an example, someone who has studied Business, will be familiar with collocations used in the Business register through their studies, so he or she will become familiar with them in a natural way, much as a native speaker would gain familiarity with such collocations. This would also mean learning them all in context, and not mixing them up. In addition, Crezee (2013) emphasises that no word list can compete with that sort of solid preparation, as they still miss the context, so there is no point in just learning lists of collocations: the best way is to learn collocations in their proper context. On the whole, the common ground in the above viewpoints is that L2 collocations should not be learnt in isolation but associated with the context in which they occur. Therefore, the present study will briefly suggest a model of collocation learning in Chapter Eight and explain why this model is suitable for translators.

6.5 Summary

In response to the fourth research question, this chapter has reported on the results of the empirical investigation into different patterns of collocation use between the

NECCD and the TECCTC, and has specifically looked at the distinctive features of Chinese translators' use of English collocations from the perspectives of form, meaning and function. These features were examined through collocation distribution from the angle of collocability degree (or the level of lexical association), delexicalization and semantic prosody. The results from the quantitative study have shown that, when compared with native speakers of English, Chinese translators tend to produce more free combinations but fewer bound collocations or idioms; more collocations with a literal sense but fewer collocations with a delexicalized sense; and more collocations with a neutral semantic prosody but fewer collocations with a positive or negative semantic prosody in translational English. The deviation in collocation patterns produced by Chinese translators will inevitably result in them introducing some translation universals into translational English. Therefore, the next chapter will respond to the quantitative findings and attempt to demonstrate, with examples, the translation universals which Chinese translators brought into their translations. It will also explore the causes leading to Chinese translators' deviation in producing collocation patterns and carry out an analysis in terms of L1 transfer.

Chapter Seven Translation universals in Chinese translators' use of L2 English collocations

7.1 Introduction

Chapter Seven will outline the role of the control of L2 collocations in translations and attempt to demonstrate, with examples, the translation universals which Chinese translators in this study brought into their translations. This chapter will also respond to both the quantitative and qualitative findings, and discuss possible reasons for the deviations in Chinese translators' use of L2 English collocations focusing on the aspect of L1 transfer.

7.2 A model of the control mechanism between features of collocations and translation universals

This study has shown that inadequate understanding of the features of collocations (e.g. formal features, semantic features and functional features) in L2 may cause some translation universal features (or translation universals) to appear in the target language. This finding provides more empirical evidence to support the theoretical framework constructed in Section 3.5.2. As mentioned in Section 3.4, translation universals can be mainly categorised as explicitation, simplification, normalisation, sanitisation and so forth. This study will primarily focus on the first three categories in relation to Chinese translators' production of English collocations. Since translation universals are referred to as the distinctive features which can be discovered through the comparison between native language and translational language, it is necessary to look at the relationship between the features of collocations and these translation universals. In order to clarify this relationship, I designed a model which is illustrated in Figure 7.1:

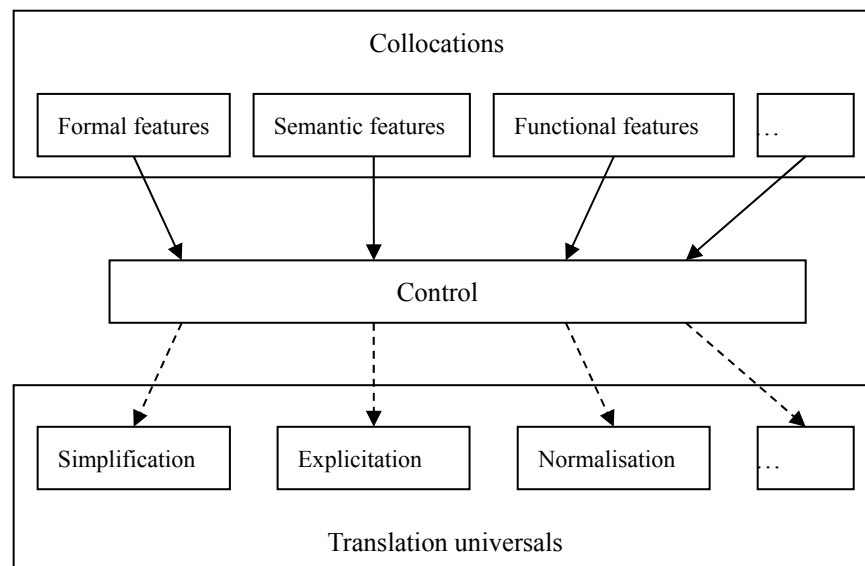


Figure 7.1 A model of the control mechanism between features of collocations and TUs

This graph shows that knowledge of the features of collocations is very important and strongly associated with translators' ability to render native-like target language. This kind of ability is reflected in translators' actual control of the features of collocations, that is to say, the way they understand collocations from multiple aspects (e.g. the formal perspective, semantic perspective and functional perspective) and use them in their translations according to their understandings. More specifically, a good control of these features proves that translators have an in-depth understanding of these collocation features, and this would help reduce translation universals in the target text. In other words, translation universals would be 'under control'. Contrary to this, a poor control of these features proves that translators' understanding of these collocation features is superficial, resulting in the increased possibility of translators introducing translation universals into the target text. In this situation, translation universals would be 'out of control' to some extent. The present study has shown that Chinese translators' collocation patterns in business Chinese-to-English translations significantly deviate from those of native speakers of English from formal, semantic and functional perspectives. This indicates that Chinese translators' ability to control these collocation features still remains weak in comparison with native norms. Therefore, it would be helpful to analyse the translation universals in the translated texts in this study and identify Chinese translators' perceived shortcomings with regard to the control of both

collocation features and translation universals. In addition, with regard to the analyses from Section 6.2 to Section 6.4, it is noteworthy that the relationships between features of collocations and the translation universals appear to be individually directional. This is to say, a poor control of formal features may result in simplification; a poor control of semantic features may lead to explicitation; and a poor control of functional features may result in normalisation in translational language. This is also demonstrated clearly in the model of the control mechanism (see Figure 7.1). Therefore, in respect to this distinction, this study will look at these translation universals individually in the following sections.

7.2.1 Simplification

As noted in Section 2.4.1, simplification is referred to as a process in which translators, while generating the target language, tend to simplify language or the message in translations (see for instance Baker, 1996). The features of simplification in translational language can be demonstrated from many aspects, such as the use of punctuation (see for instance Malmkjær, 1997), mean sentence length (see for instance Laviosa, 1998b) and lexical density (see for instance Xiao, 2010). As the present study investigates the difference in the use of English collocations between texts in the native language and texts in translational language, the concept of collocation density (see 5.2) is employed in an attempt to explore the features of simplification in translational business English. In this respect, the present study has already taken account of two aspects to clarify the difference between native English and translational English, specifically, the overall TTR (the type-token ratio) and the TTR of bound collocations and idioms (non-free-combination collocations). The former measure, that is, the overall TTR is used to uncover the difference with regard to the general features of collocation distribution from comparing the two corpora. The latter measure, that is, the TTR of bound collocations and idioms is used to look at how the collocations that are more formulaic are distributed across the two corpora. The findings from the quantitative perspective (see 5.2 and 6.2) have already shown that the collocation TTR in the

TECCTC is significantly lower than that in the NECCD. Specifically, in terms of overall TTR, the TECCTC is 3.47 while the NECCD is 6.25; in terms of the TTR of non-free-combination collocations, the TECCTC is 3.19 while the NECCD is 6.85.

In addition to TTR, I also examined the ratio of high-to-low frequency collocations in the two corpora. In previous studies, a number of researchers (e.g. Laviosa, 1998b; Xiao, 2010) have defined the threshold for identifying high frequency words, with employing a minimum proportion of 0.1% of the total lexical occurrences. In respect to this, this study employs the proportion of 0.1% as the threshold to identify high frequency collocations.

Table 7.1 Frequency profiles of the NECCD and the TECCTC

	NECCD	TECCTC
Number of types	109	198
Cumulative proportion	19.55%	37.52%
Repetition rate of high frequency collocations	182.83	211.18
Ratio of high-to-low frequency collocations	24.3%	60.05%

Table 7.1 shows the frequency profiles of the two corpora, in which the number of high frequency collocation types in the TECCTC is much higher than in the NECCD, being 198 and 109 respectively. In addition, it is clear that high frequency collocations in the TECCTC make up a significantly greater proportion (37.52%) than those in the NECCD (19.55%). This indicates that, in comparison with native speakers of English, Chinese translators depend more heavily on high frequency collocations and use them repeatedly as a strategy for producing formulaic language. This can also be evidenced by the significantly higher repetition rate of high frequency collocations in translational English (211.18) compared to native English (182.83). These factors, as a whole, result in the ratio of high-to-low frequency collocations in the TECCTC (60.05%) being considerably higher than that in the NECCD (24.3%). This also explains why Chinese translators produce more collocation tokens but fewer collocation types, thus making their L2 English more simplified than is common by native standards.

These results appear to be enough to indicate that the translational English in this study is simplified when compared with native English, and that Chinese translators' translation outputs can be characterised by their repeated use of 'favoured' collocations. Therefore, the translated texts did not show a wide range of collocation types when compared with native-speaker commercial English. To provide more evidence, I exemplified this with the collocations consisting of the word *call*.

The *Collins COBUILD Advanced Learner's English Dictionary* (2006) lists at least 20 meanings of the word *call* and 19 compounds containing *call*, such as *call-in*, *call-up*, *conference call*, *curtain call* and *judgement call*, and 7 phrasal verbs, such as *call back*, *call on*, *call for*, *call off* and *call out*. Generally, the meanings of *call* can be divided into 6 groups with regard to the semantic domains it concerns. This is shown in Table 7.2:

Table 7.2 Meanings realised by *call*

Group	Meanings	Examples
(1) say	to give someone or something a name	<i>call me Sarah</i>
	to describe someone or something as a particular thing	<i>She calls me lazy and selfish</i>
	to say something aloud to attract someone's attention	<i>'Boys!' she calls again</i>
	to shout to someone	<i>He called me over the Tannoy</i>
so-called	to indicate something by the name that you are about to use	<i>the so-called G7</i>
make noise	the characteristic sound that animals make	<i>a wide range of animal noises</i> <i>and bird calls</i>
(2) telephone	to phone someone	<i>call me</i>
	to ask someone to come to you	<i>call an ambulance</i>
	a telephone call	<i>made a phone call</i>
	to telephone for leave due to illness	<i>I called in sick</i>
(3) arrange	to arrange for something to take place at a particular time	<i>call a meeting</i>
summon	to order someone to appear at some place	<i>I was called as an expert witness</i>
(4) stay	to make a short visit	<i>Andrew now came almost weekly</i> <i>to call</i>
	to stop somewhere	<i>The steamer calls at several ports</i> <i>along the way</i>
(5) cancel	to cancel	<i>We called the next game</i>
(6) demand	someone demands that something should happen	<i>calls for a new kind of security</i> <i>arrangement</i>
	something is demanded to be done or provided	<i>there is not too much call for</i> <i>chocolate</i>
	on call; required to work anytime when needed	<i>I'm on call day and night</i>
	need for	<i>no call for him to single you out</i> <i>from all the others</i>
lure	something attracts or interests you strongly	

In the total 908 occurrences of the word *call* in the TECCTC, there are only 9 instances

of *call* constituting one type of significant collocation, that is, *call auction*. This can be exemplified with the following sentence retrieved from the TECCTC:

If most people are optimistic about the stock in question, their bid prices will be higher than [the] ex-right price and the actual opening price formed after call auction will be higher than [the] ex-rights price, and vice versa. (Group 3)

In the rest of the 899 occurrences, the word pairs containing *call* are not identified as collocations largely due to the lack of statistical significance, such as *make calls*, *receive calls*, *call on* and *call for*. In particular, *call on/upon* occurs 165 times and *call for* occurs 298 times in the TECCTC, both of which account for 51.5% of all the instances of *call*. This also provides evidence to hold that Chinese translators' translation outputs can be characterised by their repeated use of 'favoured' collocations.

In contrast, the total 3,123 occurrences of the word *call* reflect a much greater repertoire of collocation types in the NECCD. In these occurrences, 210 instances constitute 11 types of significant collocations, which include *call option*, *call centre*, *call conference*, *covered call*, *naked call*, *duty call*, *bull call*, *margin call*, *call features*, *welcome call* and *desperate call*. In some collocations, such as *call centre* and *welcome call*, the meanings of *call* can be figured out clearly as indicated in Group 2; whilst in some other collocations, such as *bull call*, *call option*, *covered call*, *margin call* and *naked call*, the meanings of *call* are hard to capture unless they are understood through the collocational relationship. In some sense, the latter kind becomes an important part of bound collocations, which can be evidenced with the following sentences from the NECCD:

Using cash as a call option in this case generated an extra 34% of return. (Group 3)

Motley Fool newsletter services have recommended creating a bull call spread position in Microsoft and writing covered calls on GameStop. (Group 3)

The former chief executive of British Land might have carried on with his day's shooting and ignored Alastair Darling's desperate call pleading with him to take the job in mid-October 2008. (Group 6)

Second, any investor who uses broker margin has to manage his or her risk carefully, as there is always the possibility that a decline in value in the underlying security can trigger a margin call and a forced sale. (Group 6)

Collocations such as these, which are widely used by native speakers of English in the commercial register were, however, largely overlooked by Chinese translators, and this may be another important factor contributing to Chinese translators' under-production of bound collocation types. Furthermore, translators' apparent lack of awareness of such bound collocations would also increase the possibility of repeatedly using their favoured word combinations to deal with any complicated text they may face. Alternatively, they might simply resort to the strategy of free lexical combination to 're-interpret' what is already formulaic in native language. All of these aforementioned factors would inevitably bring simplification in the translational English. In respect to these issues, what this brings to translator training is that, when translators are learning L2 collocations, they should not only pay attention to the typical uses as defined in dictionaries but also base their knowledge on the actual occurrences of collocation or authentic texts in the native language. This learning strategy is described by Davies (1998, 2004) as "situated learning" or learning in real life context. A good understanding of how native speakers distribute their use of collocations in terms of formal features would definitely help translators with reducing the situations of simplifying their L2 English in translations. Thus, for researchers in this area it is crucial to construct a valid pedagogical model of situated learning, within which translators can be professionally trained and acquire their L2 English collocations more effectively and efficiently. This will also be discussed in the conclusion section.

7.2.2 Explicitation

In addition to simplification, Chinese translators also exhibited explicitation in their commercial Chinese-into-English translations. As mentioned in Section 3.4.1, explicitation can be described as a phenomenon "which frequently leads to TT [the target text] stating ST [the source text] information in a more explicit form than the

original” (Cowie, 1997, p. 55), or a tendency to “spell things out rather than leave them implicit” (Baker, 1996, p. 180). In the present study, explicitation will be discussed in terms of delexicalization because Chinese translators significantly over-produced collocations with a literal sense but under-produced collocations with a delexicalized meaning, which, on the whole, appears to be a typical case of making explicit the language information which is supposed to be implicit. According to the statistical findings, Chinese translators produced 103,057 collocation tokens in the literal sense, which accounted for 92.47%, and 8,393 tokens in the delexicalized sense, which merely accounted for 7.53%; with regard to type, Chinese translators used 3,640 collocation types with a literal sense, which made up 94.01%, and used 232 types with a delexicalized sense, which made up a mere 5.99%. This differs from the scenario of native speakers of English, who used 82.9% of collocations with a literal sense, which amounted to 84,506 tokens in the NECCD, and they used 11.29% of collocations in the delexicalized sense, which amounted to 17,429 tokens. In terms of type, native speakers produced 5,647 collocation types with a literal sense, which accounted for 88.71%, and produced 719 types with a delexicalized sense, which accounted for 11.29%. Therefore, this section will look at the difference between Chinese translators and native speakers of English in producing delexicalized collocations, and investigate explicitation in translational language with some typical examples.

The phrasal verb *stack up* was examined because the meanings of adverbial particles are mostly realised through collocating with other lexical items. In this sense, collocations with adverbial particles contribute to an important portion of delexicalization in language use. From the delexicalized category across the two corpora, the collocations consisting of adverbial particles amount to 4,718 tokens and 132 types in the NECCD, and those in the TECCTC amount to 4,124 tokens and 73 types. This result indicates that at least 59 types of collocations with a delexicalized meaning were not produced by Chinese translators in the Chinese-to-English translations in the commercial register. In other words, a large number of instances during translation where words are required to be delexicalized in accordance with native norms were substituted with words possessing a literal sense by Chinese translators, thus making the language information

explicit. For instance, *stack up* occurs 17 times in the NECCD, but does not occur in the TECCTC. With AntConc (Version 3.2.1w), *stack up* is retrieved and this concordance result is demonstrated in Table 7.3.

Table 7.3 Concordance lines of *stack up* in the NECCD

Hit	KWIC
1	s they provide and how they stack up to yours. 4. Become a secre
2	o managers and see how they stack up against each other. If man
3	so that you can see how you stack up against other street perfor
4	fixed and variable incomes stack up against their fixed and var
5	r necessary living expenses stack up against your monthly income
6	re so low that they may not stack up as a serious long-term inve
7	rols. How does the facility stack up against the competition?
8	Pinterest and how its users stack up against Facebook's. Boticcs
9	awards and see how schools stack up based on graduates' salarie
10	How 6 Famous Stock Pickers Stack Up.] In other words, while pro
11	How 6 Famous Fund Managers Stack Up.] Not that Roumell's prefer
12	How 6 Famous Stock Pickers Stack Up.] Don Yacktmann (Yacktmann F
13	s, etc. These items tend to stack up , creating useless, dusty c
14	itlement and interest costs stack up). [READ: Dow Closes Above]
15	ffers and see how the plans stack up against each other. Expense
16	ension and healthcare bills stack up . Like a lot of the Govern
17	ow does the investment case stack up from the point of view of s

The corpus-based *Collins COBUILD Advanced Learner's English Dictionary* (COBUILD, 2006) lists three basic meanings of the phrasal verb *stack up*:

1. Phrasal verb, no passive, VP
If you ask how one person or thing stacks up against other people or things, you are asking how the one compares with the others. (INFORMAL)
How does this final presidential debate stack up and compare to the others, do you think? = compare
2. Phrasal verb, no passive, VP
If facts or figures do not stack up, they do not make sense or give the results you expect.
There have been a number of explanations, but none of them stack up.
3. *Stack up* means the same as *stack*
If you stack a number of things, you arrange them in neat piles.
He ordered them to stack up pillows behind his back. (COBUILD, 2006)

Therefore, according to this dictionary, the collocation *stack up* defined in group 1 means ‘to compare’, and is often combined with the word *to* or *against* to constitute a larger formulaic sequence. It is clear, from the concordance results of *stack up* from the

NECCD, that 9 instances out of the whole 17 entries (i.e. concordance line 1, 2, 3, 4, 5, 7, 8, 9 and 15) are assigned the meanings defined in group 1. This can be evidenced with by the concordance line 1 and line 3 as follows:

<S1> *Become a regular “customer” and you’ll quickly see what kind of offers they provide and how they stack up to yours.*

<S3> *There are even busking competitions so that you can see how you stack up against other street performers.*

The definition in group 2 indicates something (e.g. an explanation or a result) to be ‘tenable’, ‘expected’ or ‘anticipated’. There are 5 instances in the above concordance result which have been assigned the meaning defined in group 2, specifically, line 6, 10, 11, 12 and 17. This can be exemplified with line 6 and line 17:

<S6> *However while NS&I’s premium bonds can undoubtedly provide a fun alternative to a savings account for children or grandchildren, the odds of winning are so low that they may not stack up as a serious long-term investment.*

<S17> *How does the investment case stack up from the point of view of an average investor?*

The third meaning of *stack up* corresponds to the literal sense of the word *stack*, which indicates ‘to pile things up (nicely)’. There are three instances (i.e. line 13, 14 and 16), according to the above concordance result, assigned this meaning, which can be evidenced with line 16:

<S16> *Politically popular yes, but it could end up being an economic nightmare as pension and healthcare bills stack up.*

All these three groups indicate that the literal meanings of the two words *stack* and *up* largely vary when they constitute a collocation in English. Particularly, the meaning of *up* is completely delexicalized in this collocation. In this sense, the literal meanings of these two words are, to some extent, made implicit when native speakers of English use them in the commercial register. The implicit meanings can be directly conveyed to addressees (i.e. group 3), or can be indirectly transferred through figurative strategies

(i.e. group 1 and group 2) into a larger semantic unit which differ from these implicit meanings. In addition, this semantic unit cannot be substituted by employing the strategy of synonymy even though it can be explained through synonymy in dictionaries. To be more specific, in meaning group 1, *stack up* is regarded as equal to ‘*compare*’, but it cannot be replaced with ‘*compare*’. Otherwise, it would make no sense in the example attached (*How does this final presidential debate stack up and compare to the others, do you think?*) to use two identical phrases to make clear of the intention of the addressor. In this respect, native speakers of English who use the collocation *stack up* essentially intend to keep the implicit meanings in their discourse, which can help them achieve their aim that is not available alternatively in language communication. However, this may cause problems for those EFL translators who cannot fully master the appropriate understanding of implicit language information due to a lack of collocation knowledge and who are more inclined to make explicit the implicit language information. The absence of the collocation *stack up* in the five-million-token TECCTC may be a typical example. To further develop this argument, I examined 11 instances where the word *compare* (active) is used in the TECCTC, and found some situations where the implicit meanings may be made explicit with the use of *compare*. This can be evidenced with the following sentence:

When answering questions of reporters about how to compare China-US MOU with China-EU MOU, Mr. Bo Xilai said that the two agreements were balanced.

The source text of this translation is as follows:

在回答记者有关如何比较中美、中欧纺织品谅解备忘录时，薄熙来表示，两个协议是平衡的。

According to the source text, this sentence states the former Minister of Commerce Bo Xilai’s attitude and stance towards two versions of MOU (Memorandum of Understanding). In this translation, the word 比较 (*bi3jiao4*) (the numbers in the phonetic transcriptions indicate the values of intonations in Mandarin Chinese, ranging from 1 to 4) in the source text is literally translated into *compare* (the English equivalent

of 比较). Even though it is seemingly reasonable for translators to seek the linguistic equivalent, a further thought in this case would make them realise that this kind of ‘comparison’ between China-US MOU with China-EU MOU, as indicated in the source text, appears to imply more than the mere static description of difference or similarity. To be more specific, the word *bi3jiao4* in the source text also implies the expected outlook of these two versions of MOU, which appears to be clear from the context:

中欧的协议在今年6月11日就已达成，欧盟表示了诚意，对于营造中欧纺织品正常稳定的出口环境起到了重要作用。这次中美之间又在平等务实的气氛下达成协议，和中欧协议互为补充，成为通过平等磋商来解决贸易争端的两个成功范例。(the source text)

Agreement between China and EU has been reached on Nov. 6, and EU expressed the sincerity, which played an important role to stable [stabilise] the export environment for China-EU textile trade. Besides, [the] China-US agreement was concluded in an atmosphere of equality and practice, and the two agreements were complementary to each other and became successful typical examples of trade dispute settlement through negotiation on the basis of equality. (the English translation in the TECCTC)

In this context, the word *bi3jiao4* in the source text does not merely require the addressor to show his judgement regarding ‘which is better’ as defined by the connotation of the word *compare* in English. Rather, it is more concerned with what kind of outlook that the China-US MOU would achieve, and whether it would gain the similar significance as the China-EU MOU in the textile trade. In this respect, *bi3jiao4* has some implicit information in the source text. If this is simply understood as *compare* literally in English, then the meaning of *bi3jiao4* is obviously made explicit and shows translators’ lack of awareness of the other meaning of this word. The language information implied by *bi3jiao4* should also be taken into consideration by translators and should be included in the target text. In some sense, this kind of ‘comparison’ between the two versions of MOU not only corresponds to the first meaning of *stack up* but also indicates the second meaning of this collocation. Therefore, *stack up* is perfectly acceptable in this situation and can better fulfil the potential of the discourse than the word *compare*. Native speakers of English would bear these perspectives in mind and make the word choice in a more implicit way. In other words, the use of *stack*

up can imply, to some extent, what would happen subsequently in the text. Thus, the translation can be revised as follows:

When answering questions of reporters about how the China-US MOU stacks up against the China-EU MOU, Mr. Bo Xilai said that the two agreements were balanced.

The revised version not only avoids the null subject *he* in the lexical sequence ‘*how (he) to...*’, but, more importantly, connects this sentence closer with its context by making some language information implicit. In this respect, this revised version essentially achieves a better result in terms of textual cohesion and language information conveyance. Dimitrova (2005) proposed that the quality of the target text can be a measure of translators’ task performance results in handling translations based on the “assumption that the amount of experience correlates to the level of quality of the TT [target text]” (p. 33). In this respect, Low (2003) further noted that “good translators need real understanding of the ST [source text]” and that “[w]hen a translator decides to elucidate a text, the specific function of the TT [target text] can help to determine which choice to make out of several available options” (p. 102). Therefore, it appears that translators would need to optimise their translation outcomes when generating the TT.

However, in the above example, Chinese translators’ use of the word *compare* does not correspond to the content of the source text and is obviously not the optimal selection of the target text, and their use of *compare* largely overlooks the functions of the words with implicit meanings, such as *stack* and *up*, thus over-clarifying the precise semantic message in more detail in the target text. In this sense, their commercial translations, to some extent, did not correlate with the quality of the target language by making the language information largely explicit in their business translations, which may be due to the constraints of the target language (English), such as the lack of appropriate recognition of delexicalization in English.

Therefore, poor control of collocations with delexicalized meanings, such as *stack up*, appears to be one of the major factors contributing to the explicitation in translational

English, and one of the root causes of the Chinese translators' production of a smaller repertoire of collocation types in comparison with native speakers. One of the plausible solutions to this problem may be to increase translators' awareness of using adverbial particles with delexicalized meanings, such as *up*, *on* and *with*. Because the meanings of the functional words are hard to capture at times, particularly when these words are combined with other words to constitute collocations, the valid capture of the 'invisible' language information behind these words appears to be an important task for translators to express the meanings of words. In return, the recognition of delexicalization and the appropriate use of delexicalized collocations would also help translators reduce explicitation in their L1-L2 translations and come closer to a native-like rendition in the L2 target language. It should be noted, however, that delexicalized adverbial particles are not the only factor resulting in delexicalization, and delexicalization can be studied from other aspects, such as delexicalized verbs (e.g. *take* and *make*). Nor is delexicalization the only factor resulting in explicitation, because explicitation can also be studied from other angles, such as connectives and conjunctions (see for instance Xiao, 2010). This section is simply intended for exemplifying translation universals in translational business English, and therefore it will not allow for other factors which would inevitably require more research.

7.2.3 Normalisation

A third kind of translation universal feature that Chinese translators brought to their commercial Chinese-to-English translations is normalisation. As discussed in Section 3.4.3, normalisation refers to the "tendency to exaggerate features of the target language and to conform to its typical patterns" (Baker, 1996, p. 183), whereby translators' use of their L2 appears to be more 'formal', 'conventional' and 'normalised' than the target language. Normalisation is typically manifested when translators are using "typical grammatical structures", "punctuation" and "collocational patterns or clichés" (1996, p. 183) in the target language. In this respect, this study has briefly analysed normalisation from the aspect of semantic prosody in Section 6.4 with demonstrating the functional

features of the collocation use in the TECCTC.

As shown in the statistical analysis of data, the functional features of Chinese translators' collocational patterns can be basically formulated from two angles. On the one hand, Chinese translators repeatedly used some particular collocational patterns, which can be a major factor contributing to the distribution imbalance of the three categories of semantic prosodies (positive, neutral and negative). To explain this kind of imbalance, I compared the top 20 most frequent collocations with neutral semantic prosodies between the NECCD and the TECCTC, and found that they amount to 5,748 and 9,012 tokens respectively in the two corpora, as seen in the table below.

Table 7.4 Top 20 most frequently used collocations with neutral semantic prosodies in the two corpora

Rank	NECCD	Frequency	TECCTC	Frequency
1	board directors	925	joint venture	974
2	chief executive	853	stock exchange	947
3	real estate	529	intellectual property	853
4	supply chain	273	technological renovation	824
5	credit cards	251	bilateral relations	364
6	take place	221	general manager	357
7	business entity	219	custodian fund	352
8	prime minister	216	documents application	349
9	third party	210	export volume	347
10	fiscal year	206	commercial administrations	345
11	customer service	203	competent authorities	344
12	European Union	199	natural gas	339
13	at stage	192	trading partner	338
14	provisions law	186	push forward	338
15	cash flows	185	application materials	329
16	keep mind	181	comply with	327
17	management review	179	monetary policy	327
18	board member	177	press conference	325
19	fleet management	174	prime minister	322
20	global economy	169	implementation measures	311
Total		5,748		9,012

According to this rank list, it is obvious that some particular collocations are repeatedly used in the TECCTC, such as *push forward*, *application materials*, *comply with*, *export volume* and *implementation [of] measures*. This result is largely in accordance with the aforementioned statistical findings (see 6.4). It is also a major factor contributing to the distribution imbalance of the three categories of semantic prosodies and the over-production of collocations with neutral semantic prosodies in translational commercial English. As a result, translational business English shows a smaller

repertoire of collocation types and appears to be more ‘neutralised’ and ‘normalised’ than native-speaker commercial English in terms of collocation use. In this respect, some collocations that are widely used by native speakers of English in business discourse were not produced or used appropriately in correct places in translational commercial English, particularly those with positive or negative semantic prosodies.

On the one hand, normalisation results from Chinese translators’ lack of awareness of increasing the variety and diversity of collocations in their L2 English output. Translators appeared to rely heavily on their favoured collocations to fulfil their communicative aims. For instance, in the total six occurrences of collocation types containing *action* in the TECCTC, there are four instances used with neutral semantic prosodies, such as *action framework* and *take action*, and two with negative semantic prosodies, such as *infringement action* and *rectification action*. It is clear that collocations with neutral semantic prosodies were the main outputs where the word *action* is concerned. In contrast, in the total 21 occurrences of collocation types containing *action* in the NECCD, there are only six instances that contribute to neutral semantic prosodies, such as *take action* and *action plan*. The others either go into the positive category, such as *prompt action*, or go in the negative category, *disciplinary action* and *enforcement action*. In this respect, it would appear that the use of collocations in the TECCTC tended to ‘exaggerate some features’ of English, and therefore makes translational business English more ‘conventional’ and ‘normalised’ than native-speaker business English.

On the other hand, Chinese translators in the current research showed their comparatively weak control of the semantic prosodies of certain words and over-conformed to some typical V+N patterns while using English collocations. This is another factor contributing to normalisation in the translational commercial English, and may result in misuse of some free combinations. As an example, I examined the word *face (verb)* in the collocations retrieved from the two corpora. In the collocations retrieved from the NECCD, the word *face* occurs 177 times and collocates with other words to form 14 different collocations types. All of these 14 collocation types indicate

negative affective meanings, such as *fine*, *obstacle*, *opposition*, *penalty*, *pressure*, *problem*, *prosecution*, *challenge* and *competition*. The three sentences from the NECCD below are provided as examples:

The European Commission said Friday eight Chinese and two Indian airlines face fines totalling 2.4 million euros (\$3.1 million) for not paying for their greenhouse gas emissions on flights within the bloc.

Farha and other green entrepreneurs will face plenty of obstacles, though, not the least of which is that VC investing in the area seems to be on a downward trend.

With more and more anger and frustration in the shareholder base, HP may face pressure for yet another CEO change.

Therefore, it can be inferred that the word *face* normally possesses a negative semantic prosody in native English used in the commercial register.

In contrast with this result, the collocations involving the word *face* from the TECCTC amount to seven types. Among these seven types, six types indicate negative affective meanings, which specifically include *challenge*, *competition*, *difficulty*, *pressure*, *problem* and *task*; and one indicates a positive affective meaning, that is, *opportunity*.

The two sentences from the TECCTC below are given as examples:

While acknowledging that the new regulation is a "good thing" for homebuyers, Tu Zhibin, with a Shenzhen-based project supervising company, said it posed a challenge to the abilities of supervisors, who will face a major task inspecting all the new apartments.

Gao Hucheng said that investment and cooperation between China and ASEAN will face new opportunities with the in-depth implementation of Investment Agreement for China-ASEAN Free Trade Area.

The word *face* possesses three different semantic prosodies in the corpus of translational English, which is to some extent in opposition to the result from the corpus of native English. However, *face* and *opportunity* do not form a significant collocation because they do not co-occur in the five-million-token NEECD. This indicates that word pairs

such as *face_opportunity* are not used often in business English according to the native norms. The two sentences from the NECCD below are given as examples:

Substantial export opportunities are available to U.S. companies, and to increase U.S. business participation, the Department of Commerce maintains liaison offices at the MDBs.

There are plenty of job opportunities available to citizens brave enough to take them.

It appears, from the two examples, that Chinese translators in the current research over-conformed to the *face+N* pattern and overlooked the semantic prosody indicated by this pattern in English. Therefore, they might have directly transferred the phrase in relation to *opportunity* using the *face+N* pattern from their L1 Chinese, because 面对机遇 (*mian4dui4_ji1yu4*, which literally means *face_opportunity*) is acceptable in Chinese. The use of free combinations such as *face_opportunity* in translations obviously shows a deviation from native norms in regard to functional features and may result in normalisation in translational language. In addition, the example of *face_opportunity* can also be analysed from other angles, such as explicitation and simplification, because this word pair appears to make the meaning of *face* explicit and simplifies the some use of formal/structural patterns in translational language. This corresponds to the assumption made in Section 3.4.4, that translation universals are essentially associated with each other, and that the lines between them are blurred.

On the whole, Chinese translators' translation outputs reflect their unawareness of the functional features of L2 English collocations. This finding indicates that translators did not take context into consideration when learning L2 collocations, and that they might have used or even 'created' English collocations based on their knowledge of their own native language. As a result, they would at times suffer from L1 interference and appear unable to use correct collocations in the correct places when they render the TT using their L2. In other words, to conquer the barriers of L1-L2 differences and essentially tell 'right' from 'wrong', or 'marked' from 'unmarked', appears to be an important task for both translators and EFL researchers. If translators, as Xiao and McEnery (2006) argue,

are made aware of L1-L2 differences and are able to compare L1 collocation patterns with their L2 translation equivalents, “this should considerably reduce the number of errors from L1-L2 semantic prosody differences” (p. 126). Therefore, the next section will discuss the factors that may be responsible for the deviation in Chinese translators’ production of L2 English collocations in terms of L1 transfer.

7.3 Factors that may be responsible for the deviation in Chinese translators’ production of L2 English collocations

The present study has revealed a number of types of deviation, relating to both quantitative and qualitative perspectives, which occur in the commercial English collocations used by senior Chinese-speaking translators. Furthermore, this study has also generalised a variety of distinctive features of variation from the aspects of form, meaning and function under the proposed theoretical framework. In this sense, the findings have provided some answers to the question as to what distinguishes Chinese translators from native speakers in terms of the use of English collocations in the commercial register. In addition to these findings, the next section will also explore the reasons for the deviations in Chinese translators’ use of L2 English collocations. It should be noted here that the proposals to improve the teaching of translational skills in this section will be further expanded upon in a future publication.

With some typical examples in the section of data analysis (see 7.2), this study has shown that Chinese translators’ production of English collocations is, to some extent, influenced by their mother tongue, which is, in language studies, referred to as L1 transfer. L1 transfer, also termed as cross-linguistic influence and L1 interference, can be defined as a phenomenon in which language users carry over their language knowledge or language patterns from their native language (L1) to their second language (L2). In the present translation-oriented study, L1 transfer can be viewed as a tendency, in which translators, particularly those who have not had a native-level command of their L2 English, transfer collocation patterns from the source language

(L1) to the target language (L2) due to the lack of the knowledge of L1-L2 difference. In addition, this kind of linguistic interference can also increase the possibility of bringing translation universals to the target language when translators are handling L1-to-L2 translations.

L1 transfer can be discussed from both positive and negative influences on the L2. To be more specific, L1 transfer can help language users enhance their L2 acquisition and make their L2 production correspond to native speakers' norms of speech acceptability (positive transfer), particularly when the language units (e.g. collocations) are available in both L1 and L2; at the same time, it can also interfere with language users' L2 production and make their L2 production deviate from or even oppose native norms (negative transfer), especially when the transferred language units are not the same in L1 and L2. It should be noted, nevertheless, that theories upholding negative L1 transfer, such as Contrastive Analysis, suggest that L1 has more negative than positive influences on L2 acquisition (see for instance James, 1980). Therefore, studies of L1 transfer are mostly carried out from the perspective of the negative impact on L2 production when researchers discuss it through the perspectives of Contrastive Analysis. In particular, James (1980) clarified two main points in his hypothesis of Contrastive Analysis: a. transfer is definite and is always negative from L1 to L2; b. difficulties in L2 learning can be predicted by L1-L2 differences. As discussed in Section 4.3.3, the present study employs the Contrastive Interlanguage Analysis (CIA) approach to provide a benchmark regarding how translational English is different from native English in terms of collocation use. In this sense, L1 transfer to L2 in this study can be considered as the influence resulting from the differences between the source language and the target language. Therefore, this study will primarily focus on the negative view of L1 transfer with regard to how Chinese translators' L1 (Chinese) influenced their commercial English translations and what relationship of the source-target elements is particularly susceptible to negative L1 transfer. Hereafter, L1 transfer will be used to stand for negative transfer in this study.

The evidence from the section of quantitative research has demonstrated that L1 transfer

may be an important factor in relation to the underuse and overuse of English collocation tokens and types by Chinese translators in the commercial register. On the one hand, most English collocations underused by Chinese translators have no idiomatic equivalents or have partial translation equivalents in the source language (Chinese), such as, *bull call*, *Crown Court*, *direct debit*, *dim view* and *health coverage*. Therefore, when translating Chinese into English Chinese translators might have chosen the avoidance strategy to substitute these collocations with other word strings and left a ‘trace’ of their L1 in the target text. On the other hand, most English collocations overused by Chinese translators have direct translation equivalents in the source language, such as *financial crisis*, *bilateral relations*, *enhance cooperation* and *mutual benefit*. This may make translators rely heavily on these ‘familiar’ and ‘favoured’ collocations, which would increase the probability of using them repeatedly in translations but decrease the possibility of enlarging their L2 English collocation variety. The overuse of these ‘favoured’ collocations will definitely reduce the chances of using those collocations with no direct translation equivalents in the source language, which, overall, conforms more to the norms of the source language rather than the target language, thus also leaving a ‘trace’ of L1 in translations. Next, this study will further explore Chinese translators’ use of L2 English collocations from these two aspects.

Based on the two aforementioned assumptions, this section will examine L1 transfer with the word pair **deepen_reform* used by Chinese translators in the TECCTC. The word pair **deepen_reform* was identified as a significant collocation and occurs 55 times in the TECCTC. The word pair **deepen_reform* means ‘to push through, accelerate or promote reform’ and indicates ‘to build upon what has been achieved in the process of reform (mostly refers to economic reform)’. It is translated from a frequently used phrase in business Chinese 深化改革 (*shen1hua4gai3ge2*). In this pair of translation equivalents, *deepen* literally corresponds to 深化 (*shen1hua4*), and *reform* literally corresponds to 改革 (*gai3ge2*). The word pair **deepen_reform* is widely accepted and used in the public domain whenever *shen1hua4gai3ge2* is required to be translated into English. This can be exemplified with the following sentences from the TECCTC:

It will deepen the reform of energy prices to introduce a pricing mechanism favorable for energy conservation.

There is [a] need to firmly deepen [the] reform to the IPO and exit system.

In the next step, the CSRC will continue to deepen the reform and devote itself to building [an] optimized market structure, improved market restraint and operation system, in order to protect the lawful interests of investors.

However, the word pair **deepen_reform* did not occur at all in the NECCD, which indicates that *deepen* does not normally collocate with *reform* in native English, or native speakers of English do not understand the connotation of *shen1hua4gai3ge2* the same way as Chinese translators. Therefore, **deepen_reform* does not sound like native-speaker English according to the norms of the English language. Instead, the verbs or phrasal verbs which most commonly collocate with *reform* in English are *adopt*, *bring about*, *introduce*, *push through*, *carry out/through*, *implement*, *promote*, *reinforce*, *undertake*, *accelerate* and so forth (see for instance *Oxford Collocations Dictionary for Students of English*). In addition, the collocates of *reform* can also be seen with a number of examples in the NECCD:

That is being blocked the moment because the Conservatives do not want to have that debate and that's why we can't move forward with the wider reforms to our welfare system.

If we're going to do further welfare reform - you need to start having a debate about how we ask people at the very top to change the benefits that they receive.

The OECD's report says that the crisis has accelerated the pace of pension reform in OECD countries.

The legal environment for secured lending can be strengthened through collateral widening measures that codify land rights, promote legal reform for institutions, cooperatives and NGOs, and expand borrowing laws to increase the participation of poor.

To make sure we're in a better position to create the industries and jobs of the future, we need comprehensive reform of our business tax system.

The IMF and the World Bank can help to accelerate the process of financial

sector reform in several ways.

The difference between L1 and L2 English has shown clearly that language users with different L1 backgrounds may view the same thing differently. In this case, *gai3ge2* has a direct translation equivalent, that is, *reform* in English, which means that *gai3ge2* and *reform* make the same sense in different languages. However, in the knowledge map of L1 Chinese, native speakers, such as translators in this study, would regard *gai3ge2* as a product which indicates a static semantic property, either specific or abstract, such as 井 (*jing3*, literally ‘well’), 水 (*shui3*, literally ‘water’), 渠道 (*qu2dao4*, literally ‘channel’), 呼吸 (*hu1xi1*, literally ‘breathing’), 思想 (*si1xiang3*, literally ‘thoughts’), 知识 (*zhi1shi0*, literally ‘knowledge’) and 友谊 (*you3yi2*, literally ‘relationship’). Therefore, Chinese words relating to this category can always be associated with the meaning 深 (*shen1*, literally ‘deep’). In this sense, *gai3ge2* can be described to be ‘the deeper the better’, which is the reason why native speakers of Chinese use the word *shen1hua4* (literally ‘deepen’) to modify *gai3ge2* in Chinese. This is widely accepted in Chinese-speaking speech communities. Contrary to this, in the knowledge map of L1 English, native speakers would mostly regard *reform* as a process or procedure which indicates a dynamic semantic property, either specific or abstract, such as *growth*, *income*, *trend*, *progress*, *innovation*, *manufacturing* and *development*. Therefore, English words relating to this category normally collocate with lexical items which indicate description of a process, such as *accelerate*, *do further*, *promote*, *push through* and *speed up*. In this sense, the word *reform* can collocate with some verbs or phrasal verbs, such as *accelerate*, *do further* and *promote* to indicate the connotation of *shen1hua4gai3ge2*. In addition, transformation of part-of-speech can also be a valid strategy to indicate the meaning of *shen1hua4* in translations. For instance, *shen1hua4* (‘deepen’) can be transformed into adjectives, such as *wider* and *comprehensive*, to modify the word *reform*, which can be evidenced from the first and the fifth examples obtained from the NECCD. In sum, it appears that *gai3ge2* and *reform*, even though signifying the same thing, are viewed differently by different language users in different dimensions of connotation. In this respect, L1-L2 difference makes it clear that using *deepen* to modify *reform* in English may achieve a similar result of ‘weighing a thing

with a tape measure'. Therefore, the translation of *shen1hua4ga3ige2* into **deepen_reform* virtually conforms to the conventions of the Chinese language but largely clashes with the norms of English, thus leaving a 'trace' of L1 (Chinese) in the translator's output of L2 (English).

It also appears that L1 transfer makes L2 language users 'construct' direct translation equivalents between L1 and L2 which, however, do not exist between the two languages. In the above example, when modifying *gai3ge2*, *shen1hua4* does not have a direct and definite translation equivalent in English. The word *shen1hua4* can be translated as *accelerate*, *do further*, *promote*, or even *wider* and *comprehensive*. However, if *shen1hua4* is transferred from Chinese to English, then it will have a direct translation equivalent in English, that is *deepen*. The word pair **deepen_reform* will be regarded as a 'prototype' translation of *shen1hua4gai3ge2* in English, with which Chinese translators would increase the chance of using **deepen_reform* repeatedly but largely decrease the possibility of producing more appropriate English collocations, such as *accelerate_reform*, *promote_reform*, *do further_reform* and *(with) wider reform*. This can be evidenced by the fact that *deepen_reform* occurs 55 times, but *accelerate_reform*, *promote_reform* and *wider reform* only occur 12 times, 40 times and twice respectively in the TECCTC.

To sum up, L1 transfer is an important factor influencing translators' collocation learning and production in their L2. It will not only result in translators deviating from appropriate understanding in their L2 but also decrease the accuracy of conveying language information in translation tasks. More importantly, L1 transfer can also make translators depend on particular collocation patterns, which would definitely result in their overuse or underuse of some particular collocations when compared with native speakers. This, from another angle, indicates the importance of collocation control in translations. Good control of collocation use in L2 will not only reduce the interference from the native language but also help them essentially clarify the L1-L2 differences. Therefore, researchers in this area should also take account of the factor of L1 transfer when they are trying to lay out a model incorporating the role of collocation in L2 input.

This will be discussed further in Chapter Eight.

7.4 Summary

The present chapter discussed the importance of recognising the features of collocation patterns in using L2 English, and proposed that translators should have appropriate control of these features because these features are strongly associated with translation universals (which are in turn associated with non-native translational language). This chapter also outlined the role of the control of L2 collocations in translations, and exemplified a number of instances where Chinese translators' poor control resulted in a decreased use of formulaic language and decreased accuracy in L2 production and introduced translation universals into translational English due to the translators' lack of adequate understanding of the features of collocations. Translation universals in this study were examined from three main aspects, specifically, explicitation, simplification and normalization through the comparison between the NECCD and the TECCTC. The findings regarding these three aspects can be generalised as follows:

- a. translators' poor control of collocations with delexicalized meanings is one of major factors contributing to the explicitation in translational English;
- b. translators' unawareness of bound collocations may increase the possibility of repeatedly using their favoured word combinations, thus simplifying their L2 English in translations;
- c. translators' weakness in distinguishing different types of semantic prosodies in English may result in the target text being normalised through the overuse of collocations with neutral semantic prosodies.

All the quantitative and qualitative findings indicate that Chinese translators' lack of knowledge with regard to L1-L2 differences is the key factor leading to the transfer of their native language (L1 transfer) into translational English, which essentially causes the deviation in Chinese translators' use of English collocations. These findings also

echo those in some previous studies, where L2 learners' collocational use is adversely affected by their native language and the confusion with their L2, and therefore, their collocational use in their L2 is hampered by the deficiency in both lexical and grammatical words (e.g. Fan, 2009). In some examples provided in this chapter, Chinese translators' insufficient L1-L2 knowledge caused them to transfer some collocation patterns directly from their native language into their L2 English. This will not only constrain translators' collocation use in their L2 English, such as Chinese translators' production of a smaller repertoire of collocation types, but also increase the possibility of introducing translation universals in the target text.

Therefore, researchers in this area should take note of any differences between the collocation patterns produced in native English and translational English, and take advantage of the findings from their own research to suggest a valid pedagogical model of situated learning which clarifies and exemplifies the L1-L2 differences and points out some 'false friends' such as *deepen reform*. As Low (2003) noted, "it is essential that the skill of translators be deployed, both for the sake of the performers (so that they can render the words well) and for the sake of the listeners (to give them at least some idea of the verbal dimension of the performance)" (p. 94). Only in such a way can translators be advised more effectively as to how to avoid L1 transfer in translations and how to achieve the most natural target texts using appropriate L2 collocations. This is also a crucial issue regarding how to increase translators' language proficiency through quality L2 input. In respect to this, Chapter Eight will discuss the implications of these findings in translator training.

Chapter Eight Implications of findings

8.1 Introduction

This section will summarise the implications of the findings from theoretical, practical and pedagogical perspectives. Section 8.2 will summarise the theoretical implications of the findings based on the theoretical framework established. This section will look at the role of collocation in translation and the relationship between collocation and translation universals, and will also examine some previous theoretical models in relation to collocation learning. Section 8.3 will outline the practical implications of the findings and show translators how they could apply the knowledge about the collocations retrieved to their future commercial translations. Section 8.4 will generalise the pedagogical implications of this study and attempt to offer a number of suggestions in regard to translator training.

8.2 Theoretical implications

The present study described the gaps in the literature where the role of collocation has not been identified in translation. It also argued that the actual collocation use in translational language has not been discussed systematically in previous relevant studies. In respect to these issues, this study attempted to outline a theoretical framework (see 3.4.2) and address these gaps in the following three ways. First of all, this theoretical framework clarified the role of collocation in translation and emphasised the importance of using appropriate collocations in L1-to-L2 translations. Secondly, it elaborated on the different strategies of learning collocations between L1 and L2 learners, thus serving as a method of re-assessing the previous models in relation to collocation learning. Thirdly, this theoretical framework attempted to show the differences between native-speaker language and translational language, thus offering an opportunity to investigate translation universals. These points also underpin the rationale of the present study. In particular, Chapter Five to Chapter Seven has provided a lot of empirical data and

suitable examples, all of which appear to be ‘strong evidence’ to support this theoretical framework. Therefore, this framework appears to be valid and may have some theoretical implications.

8.2.1 Clarifying the role of collocation in translation

There are two reasons why learning L2 collocations is beneficial to learners of an L2. One is that it can facilitate L2 learners’ language acquisition and development (e.g. Wray & Perkins, 2000). The other is that it can help L2 learners achieve native-like selection and fluency (e.g. Pawley & Syder, 1983). The translators referred to in this study are also learners of English as an L2, and may be said to be somewhere on the continuum between somewhat advanced and very advanced learners of English as an L2. Even so they are distinct from L2 learners who are not translators in a number of ways as explained in Section 3.3.

Based upon Wray and Perkins’s (2000) model (see Figure 2.2) and Ellis’s (2001) model (see Figure 3.1) relating to L1 learning, I will describe L2 users’ production of collocations from the angle of facilitating L2 language acquisition and development. I have borrowed two notions from Wray and Perkins’s (2000) model, that is, holistic involvement and analytical involvement, in an attempt to clarify and depict the role of collocation in the different phases of L2 learning. It should be noted, however, that the holistic and analytical proportions differ from those described in Wray and Perkins’s proposal. This difference is illustrated in Figure 8.1 (see Figure 2.2 for comparison):

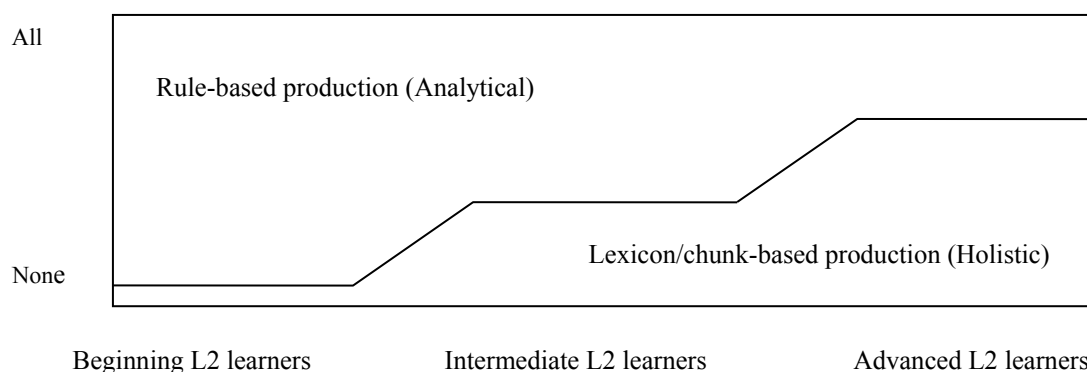


Figure 8.1 Relative proportions of L2 learners' holistic and analytical involvement in language processing from beginning level to advanced level

This model indicates that L2 collocation learning is substantially different from that of L1. L1 learning strategies are dominated by the memory-based (holistic) mechanisms in their early childhood and adult phase. Therefore, L1 learners do not tend to show increased grammatical awareness until they have accumulated moderate knowledge of collocation in language operations. It appears that L1 learners construct their grammar knowledge (or language rules) on the basis of pre-existing knowledge of collocation, and develop their language use to a standard where they can meet the needs of daily communication, even though they may not be aware of the collocational relationship between lexical items. In this sense, adult L1 learners rely mainly on memory-based mechanisms.

Figure 8.1 shows that L2 learners, in contrast, are normally exposed to both vocabulary and grammatical rules before they can access a large variety of collocations, and that rule-based (analytical) production accounts for the majority of their L2 output strategy in the early phases. In this respect, individual words, instead of lexical chunks, are more likely to be absorbed into L2 learners' long-term working memory system (see Figure 3.2), and L2 grammatical rules operate when learners produce their L2. It appears that L2 learners construct their collocation knowledge on the basis of pre-existing knowledge of L2 grammar (or language rules). This means that when L2 users have not accumulated enough knowledge about collocation they would have to resort to other strategies to produce their L2, which might lead to the deviation in collocation use, such as L1 transfer/interference (see 7.3 for an example). L2 users' inadequate knowledge of

collocation might also become a barrier if they plateau in their L2 development (see Ellis, 1994). Figure 8.1 also shows that L2 learners tend to enhance their language ability by gradually decreasing the analytical involvement and gradually increasing the holistic involvement in their L2 operations. In other words, L2 learners' language development can be described as moving towards a gradual decrease in the reliance on rules and towards an increase in the reliance of collocations in their L2 operations (Paradis, 2004). This process will not cease until their L2 use has reached the standard where the L2 learners can meet needs of daily communication according to native norms (see also Paradis's, 2004 for the distinction between implicit and explicit knowledge).

For translators, the appropriate use of L2 collocations is an important factor determining their rendition of native-like target texts in L1-to-L2 translations. Translators can be regarded as advanced L2 users (see 3.3) who employ collocations as one of the 'tools' in their 'toolkit' to transform the decoded information into linguistic representations (e.g. translation units). Both quantitative and qualitative data in this study showed that Chinese translators' actual use of L2 English collocations involves more rule-based production but less lexicon/chunk-based production when compared with native speakers. This can be evidenced by their production of more free combinations but fewer bound collocations and idioms, more collocations with literal meanings but fewer collocations with delexicalized meanings, and more collocations with neutral semantic prosodies but fewer collocations with positive or negative semantic prosodies. These findings indicate that the holistic and analytical involvement in translators' language processing is essentially different from that of native speakers of English. These findings also imply that their collocation knowledge still remains somewhere at the explicit stage (the reliance on the analytical mechanism) and has not yet reached the implicit stage (the reliance on the holistic mechanism).

Implicit knowledge refers to the knowledge which comes from language users' systematic verbal performance and is used without awareness or effort (Paradis, 2004). In this sense, lexicon/chunk-based language production is based on implicit knowledge.

In contrast, explicit knowledge refers to the procedural knowledge language users are aware of in language use, which requires their analysis in verbalizing (Paradis, 2004). In this respect, rule-based language production is based on explicit knowledge. When translators are exposed to L2 collocations and become aware of these they will accumulate their knowledge in their 'database' system, implicitly or explicitly. Translators' implicit and explicit knowledge systems of the L2 jointly interact with their control of collocation use when they produce a target text. In other words, the extent to which a translator's knowledge system includes implicit and explicit knowledge of collocations might have an impact on the naturalness of the L2 collocations they produce. To be more specific, if implicit knowledge exceeds explicit knowledge in their L2 collocation knowledge system, translators will be more inclined to produce such collocations without awareness. This would consolidate their implicit knowledge of the target language, which will make them 'closer' to native speakers in terms of collocation use. In contrast, if explicit knowledge is greater than implicit knowledge, translators will be more inclined to produce their L2 collocations consciously and think about what they are doing while handling translation tasks, which may make them deviate from the use of native-like collocations.

However, it should be noted that when translators accumulate collocation knowledge, mere exposure might not be guaranteed to lead to awareness of the nature of collocations and transformation of explicit knowledge into implicit knowledge. In respect to this, I propose that collocations are best learnt in the 'real life' context where situated learning is effective (Gonzalez Davies, 2004; Kiraly, 2000). In addition, situated learning needs to be mediated. To be more specific, teachers in translator training need to check whether trainee translators are aware of collocations in authentic texts and able to paraphrase the meanings of collocations, and more importantly, whether trainee translators are able to use them correctly in correct situations in their L2. Only in such a way can translators achieve implicit collocation knowledge effectively and turn their explicit knowledge into implicit knowledge. This will be discussed in more detail in Section 8.4.

8.2.2 Re-evaluation of theoretical models

Based on the theoretical framework of this study, the quantitative analysis section involved the re-evaluation of the previous theoretical models of learning and using L2 collocations. As mentioned in Section 3.2, there are basically two kinds of viewpoints regarding the learning and use of collocations. One standpoint holds that L2 learners, like L1 learners, can learn collocations using the chunking or priming mechanism through an associative process and they can retain complete or partial collocational information after exposure to collocations, which can be described as a formulaic approach (e.g. Ellis, 2003; Durrant, 2008). The other emphasises that L2 learners tend to ignore the collocations they see or hear due to their inadequate knowledge of formulaic language. Rather, they are more inclined to ‘notice’ individual words than recognise formulaic sequences or memorise them as wholes, so they cannot retain collocational information after the exposure to collocations, which can be described as a non-formulaic approach (e.g. Wray, 2002). Obviously, the latter viewpoint supports Kjellmer’s hypothesis that EFL learners normally produce their L2 from individual words rather than collocating words, and, as Kjellmer (1991) noted, learners “building material is individual bricks rather than prefabricated sections” (p. 124). Durrant (2008) found that L2 learners are able to establish association between words and retain collocational information under particular conditions. Crezee and Grant (2013) showed that advanced learners can mostly correctly interpret collocations when exposed to them in context.

In respect to the discrepancy in modelling the learning and use of L2 collocations in previous studies, the present study used native-speaker texts which largely avoid the so-called ‘artificial elements’. The results showed that even though Chinese translators are able to use the chunking mechanism to produce a certain number of collocations in translations, their knowledge about formulaic language still remains weak and is essentially distinguished from the native level. This can be seen from their under-production of collocation types from a number of aspects (see 5.2, 5.3, 6.2, 6.3 and 6.4). These results provide empirical evidence to support Wray’s (2002) model

regarding the use of L2 collocations. These results also indicate that translators' pre-existing knowledge associated with their L1 Chinese may at times interfere with the associative process of chunking in their L2 English production. Therefore, Chinese translators' use of English collocations in the current study presents different collocation distribution patterns to native English norms.

Furthermore, the section of lexical analysis (see 5.4) has shown that the lexical coverage in both the NECCD and the TECCTC corpora is quite similar, with the results being 34.40% and 34.98%. This means that Chinese translators appear to acquire an English vocabulary about the same size as that of native speakers in the commercial register. One may also speculate that Chinese translators' deviations may result from the elements dominating the construction of collocations, that is, the keywords. The keyword coverage across the NECCD and the TECCTC is 30.02% and 38.13% respectively. This indicates that the higher the value the smaller the variety of collocation types, because in the NECCD, on average, one keyword can collocate with approximately 2.33 words (42.9%) while in the TECCTC one keyword can only collocate with approximately 1.62 words (61.62%). This finding is also echoed in the subsequent examination of keyword growth. The results of keyword growth analysis have shown that Chinese translators under-produced keyword types on the whole when compared with native speakers, and that Chinese translators' actual production of keyword types does not increase as fast as that of native speakers. It also appears that there is a tendency for any difference between the two groups of speakers regarding keyword growth to rise proportionally with increases in text size.

In this respect, it is evident that, even though Chinese translators appear to be able to master a large vocabulary, their ability of pairing up words into larger lexical chunks still remains weak. It appears, from another angle, that translators do not acquire their L2 English completely from the input to which they are exposed. In other words, Chinese translators in the present study appear able to retain some collocational information, but it also appears that they have not yet reached the native-like level of using L2 English collocations. This finding to some extent supports Wray's (2002)

model where L2 users tend to break chunks into individual words and fail to identify the collocational relationship. This finding also indicates that Durrant and Schmitt's (2010) statement regarding L2 learners being able to establish association between words and retaining collocational relationships might not apply to actual instances, and that their proposal may only be valid in 'laboratory-based' situations. This is because L2 learners who participate in a particular test and are trained under particular conditions (e.g. single exposure and verbatim exposure) may find it easy to keep collocational information in their short-term memory system. In addition, they know for what purpose they are participating in the test. This would greatly raise their attention to particular co-occurring words in the test. However, these co-occurring words may not be effectively recognised and learnt if they do not keep the collocational information in mind or associate the collocational information with linguistic situations (e.g. what it refers to and where it is used).

In actual language learning, L2 learners may sometimes find it very hard to store co-occurring words in their long-term working memory system (see Figure 3.2) unless the collocational information has been strongly associated with their situated cognition and has become their implicit knowledge. In this respect, the mere increase of exposure to L2 collocations might not be an ideal method for translators to achieve native-like collocation selection and fluency. This is in stark contrast with Durrant and Schmitt's (2010) proposal that the shortfall in L2 learners' collocation knowledge is "more likely to be the result of insufficient exposure to the language than of a fundamentally different approach to learning" (p. 182). Instead, as Nation (1990) suggests, "[t]he network of associations between words in a native speaker's brain may be set as a goal for second language learners, but this does not mean that directly teaching these associations is the best way to achieve this goal" (p. 190). In line with Nation (1990), Crezee and Grant (2013) further propose that the knowledge of idiomatic collocations can be more effectively acquired by L2 learners and translators when they are exposed to collocations in context and learn them as authentic language materials. Therefore, the associative process of chunking in L2 learning and teaching should also involve a scientific method or pedagogy which essentially takes account of more important

factors, such as L1-L2 difference, knowledge map planning and situated learning (see 8.3.2). Only in such a way can L1-to-L2 translators, as well as ordinary L2 learners, construct a solid implicit knowledge system of L2 collocations and come closer to the native-like level in collocation use.

8.2.3 Providing evidence for the hypothesis of translation universals

Based on the theoretical framework of this study, the quantitative analysis section has provided empirical evidence to support the hypothesis of translation universals (TUs) and may encourage relevant future studies. As an important research area in Descriptive Translation Studies (DTS), the study of TUs has been controversial and problematic because the existence of translation universals still remains debatable. Some researchers (e.g. House, 2008; Malmkjær, 2007; Tymoczko, 1998) are sceptical about the hypothesis of TUs whilst some other researchers (e.g. Baker, 1996; Blum-Kulka, 1986; Chesterman, 2004; Mauraanen, 2007) accept the hypothesis because TUs are generalised to a high standard to distinguish translational language from native-speaker language. The present study looked at TUs from simplification, explicitation and normalisation by comparing translational English and native English, and examined them in terms of linguistic indicators, namely collocability, delexicalization and semantic prosody.

The quantitative analysis (Chapter Five and Chapter Six) revealed Chinese translators' weaknesses in using L2 English collocations in respect to the aforementioned linguistic indicators when compared with that of native speakers. To be more exact, the collocation distribution patterns produced by Chinese translators indicated an imbalance in that they clearly reveal translators' inclination to use a particular strategy in language production, that is, the over-use of free combinations and collocations with a literal sense or a neutral semantic prosody. The qualitative analysis (Chapter 7) also provided some typical examples in terms of simplification, explicitation and normalisation in an attempt to prove the existence of TUs in the corpus of translational English. These examples appear to be appropriate evidence to demonstrate how Chinese translators

were influenced by L1 interference and directly transferred collocations from their mother tongue, such as *深化改革-deepen reform*, due to their lack of awareness in L1-L2 differences.

The findings in both the quantitative and qualitative sections greatly support the proposed theoretical framework of this study on the one hand. For translators, the accurate use of L2 collocations can help them combine words to constitute accurate high-frequency translation units, which are strong enough to break the constraints of TUs and show naturalness in the target text. Contrary to this, the inappropriate use of L2 collocations will prevent translators from producing appropriate translation units and make them more likely to produce a target text which involves many of the aforesaid TUs. On the other hand, these findings also imply that in order to achieve the native-like rendition of target text translators would not only need to master a large repertoire of L2 collocations but also need to identify the features of these collocations, such as formal features, semantic features and functional features. Only in such a way can they know how to use correct collocations in correct places. This indicates that translation teachers or tutors would need to take the aforementioned points into consideration and design an effective pedagogical curriculum in translator training. This will also be discussed in more detail in Section 8.4.

8.3 Practical implications

Chapter 4 outlined the method used to retrieve collocations from corpora. This procedure provided a rationale to retrieve collocations using a corpus-driven approach. This method is not restricted to the commercial register only, but can also be used for investigating collocations in other registers, such as medical or legal English.

Another possible practical merit of this study is that translators whose L1 background is not English can use the collocations retrieved from the NECCD and associate them with their L1 to construct collocation pairs in commercial translation. In this sense, the

collection of these collocations can be used as a kind of database which contains not only the ‘prototype’ of how native speakers of English currently use collocations in the commercial register but also the equivalent linguistic representations of the translators’ mother tongue. Thus, when translators come across the occurrences of any of these linguistic representations in their L1-to-L2 translations, they may easily find the equivalent English collocation representing the sense. Furthermore, this approach is particularly important for those translators who are now able to use modern pre-designed software tools, such as SDL Trados and MemoQ, to facilitate their translation tasks, rather than relying on paper dictionaries. Translators can import collocation pairs into the database of the terminological management, such as Trados MultiTerm, and align this database with the translation memory system (a database which stores translated language pairs for future reference in translations) of the translation software tool. Here are two examples with storing significant collocations identified from the NECCD, specifically, *stimulate [the] economy* and *boost [the] economy* into Trados MultiTerm. These two can be made into two collocation equivalents 刺激经济-*stimulate [the] economy* and 促进经济-*boost [the] economy* in plain text or Microsoft Excel format, and the file containing these two collocation pairs can be incorporated in Trados MultiTerm, with the result being demonstrated in Figure 8.2:

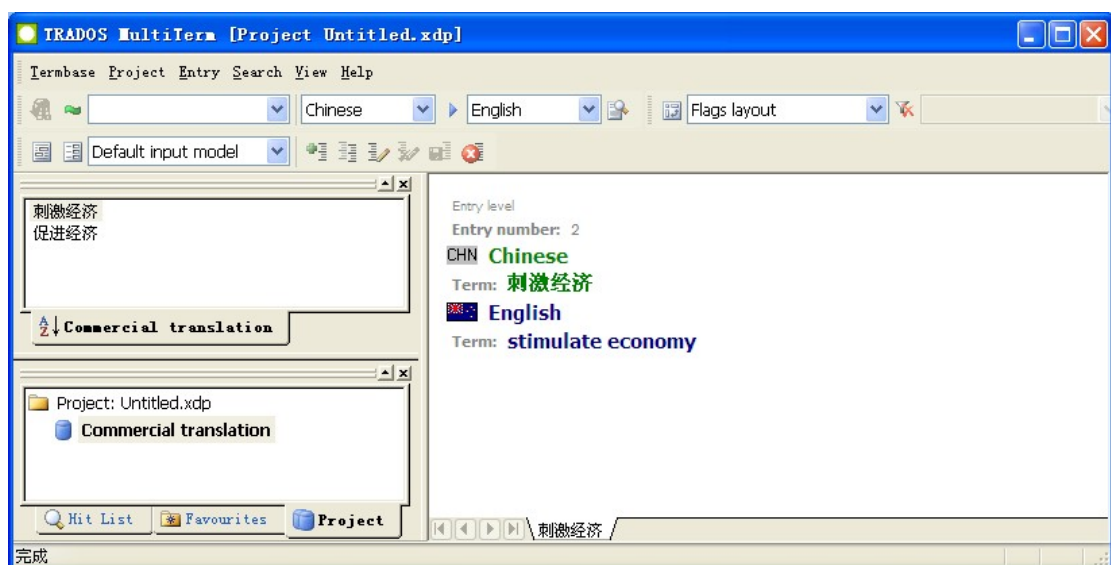


Figure 8.2 Examples of collocation pairs in MultiTerm

Thus, these two collocation equivalents are ‘memorised’ by the translation software tool. In other words, whenever translators come across 刺激经济, for example, in the source text and try to seek a linguistic representation in the target text for their Chinese-to-English translations, *stimulate [the] economy* will show up as a useful reminder to help them facilitate this wording process. This example is simply intended as a brief demonstration of the functionality of aligning existing collocation pairs with translation software tools. In practice, all the collocations identified in the NECCD can be made into collocation equivalents and stored in a terminological management system (e.g. Trados MultiTerm, MemoQ) of commercial English. In this way, translators working with translation software tools can not only build on what they already know, but also explore more possibilities of being exposed to a wider range of L2 collocations. In return, this approach will also help translators overcome their weaknesses in using L2 English collocations and ensure accuracy, fluency and complexity in their translation tasks. In this sense, the utilisation of theoretical findings has the potential to increase translators’ work efficiency and effectiveness.

8.4 Pedagogical implications

This section will now move to the pedagogical merits of this study and attempt to provide recommendations regarding how the findings of this study can be used in translator training. As explained in Section 3.3.3, translators training courses or programmes are now offered at a great number of China’s higher education institutions, where trainee translators can improve their translation skills and obtain their professional qualifications through systematic study on translation theory and translation practice. Translation educators’ training strategies and pedagogical approaches appear to be very important because they directly determine whether trainee translators can reach the expected goals in training and develop the skills that will help them achieve professional translation competence in their future career. In this respect, the findings of this study could provide useful strategies for translation trainers and educators to improve their curriculum design to help trainee translators overcome their

difficulties in translation practice. Therefore, I will summarise the implications of these findings from both practical and pedagogical aspects.

This study suggests that it might be useful to take advantage of the theoretical and empirical findings to enrich translators' knowledge regarding the use of L2 collocations in translator training. Translation programmes aimed at students working in this language pair are offered at various universities around the world, such as in Mainland China, the United Kingdom, the United States, Canada, South Africa, Australia and New Zealand. This means that many more translators who are working between Chinese and English may be confronted with similar difficulties and may have shown similar weaknesses in their commercial translations as to identifying the features of English collocations as described in Chapter 6. In this respect, the findings of this study could be applied to the curriculum design of such translation courses where teachers would consider how to show trainee translators the formal, semantic and functional features of English collocations in the commercial register. In particular, the two self-built corpora, specifically the NECCD and the TECCTC, could be potentially useful resources for teachers to encourage trainee translators to identify their weaknesses through contrastive approaches. Thus, trainee translators would pay more attention to the collocations of low frequencies, those used in delexicalized senses and those with positive or negative prosodies. Accordingly, they would also learn to use these English collocations appropriately when dealing with commercial Chinese-to-English translations. In addition, this method would enable trainee translators to evaluate what they have learnt and explore what they need to improve through in-class exercise and interaction, thus making them become aware, resourceful and reflective practitioners.

However, as mentioned previously, the approach of exposing translators to as many collocation types as possible, or the 'massive exposure' model, might not be pedagogically effective because learning L2 collocations is not a simple linear process but rather a complicated, repeated and life-long one. More importantly, acquiring the knowledge of L2 collocations for trainee translators should not stay on the theoretical

grounds only; rather it should practically involve translators' awareness of idiomatic language they have seen or heard to avoid 'missing the plot' (i.e. their unfamiliarity with such idiomatic expressions) in translation practice (see Crezee and Grant, 2013). In this sense, trainee translators need to not only add declarative knowledge to what they know, but also integrate procedural knowledge into the L2 collocation knowledge system they have already constructed. Therefore, it is important to associate the findings of theoretical studies with translators' situated cognition, which is what Davies (1998, 2004) has termed 'situated learning' or learning in a real life context. One of the most effective methods to offer situated learning for educators is to use authentic natural texts (see Crezee & Grant, 2013).

For teachers and translation trainers, teaching L2 collocations with authentic language materials essentially aims to improve trainee translators' implicit, rather than explicit knowledge of their L2, which requires trainee translators to master not only 'what to learn' but also 'how to learn'. In order to clarify the process of L2 collocation learning and acquisition, I have provided a number of suggestions and outlined a possible pedagogical method in an attempt to contribute to curriculum design for teaching collocations in an L2 context. These suggestions specifically involve exposure, exercise, evaluation, exploration and feedback and are illustrated in Figure 8.3. It should be noted, however, that the validity of these recommendations would inevitably call for relevant future research to provide more empirical evidences.

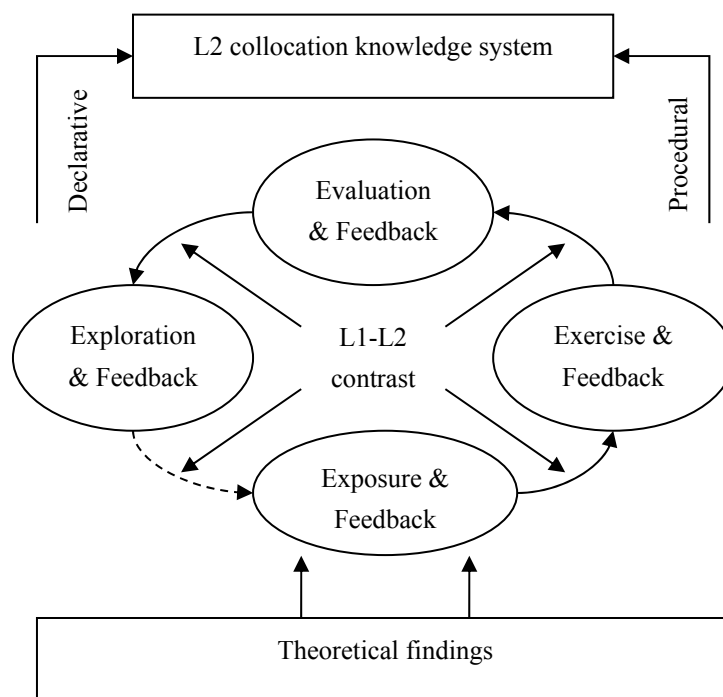


Figure 8.3 ‘Four E’s’ strategies in L2 collocation learning and acquisition

Figure 8.3 shows that theoretical findings can be used to help increase trainee translators’ exposure to L2 collocations in the learning and use of collocations, so they can observe the features of collocation distribution patterns in a variety of ways from the authentic native language materials. That is to say, the more trainee translators see or hear L2 collocations the more likely they will be able to memorise a large repertoire of collocations and consolidate their L2 collocation knowledge (‘exposure’ in Figure 8.3). This is in line with the criteria of the ‘massive exposure’ model. This would also enhance the probability of trainee translators noticing some keywords and combining them with other words to form collocations in their L2. Thus, the lexical coverage and the keyword growth (see 5.4) would both be increased in their use of L2 collocations. Nevertheless, this is only the prerequisite. In this step, teachers or translation trainers would need to employ a number of strategies, such as the frequency of exposure and the control of repetition, to ensure the effectiveness of translators’ exposure to L2 collocations, as well as individualised feedback on draft translations produced by translators. Such strategies could enable trainee translators to successfully recognise a collocation as a unit, rather than further breaking it into smaller viable units. This might prove to be a valid method of reducing free combinations and increasing bound

collocations or idioms when translators tend to produce formulaic language. For instance, the collocation list from the NECCD in this study can be used for training those translators who specialise in commercial translation. Based on the observed collocations, translators can practise using them and exchange opinions with their peers interactively ('exercise' in Figure 8.3) whilst receiving feedback from their translation teachers and tutors. This is a reciprocal step, in which teachers could build on what translators already know and extend their collocational competence and increase collocational variety in trainee translators' L2 by motivating them to refer to native-speaker texts in some resources, such as a corpus. A number of corpora are available, such as the International Corpus of English (ICE) and the Brigham Young University corpora which include free access to the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), to set good examples for trainee translators to follow. As Trebits (2009) pointed out, data-driven activities could assist second language study in a number of ways, among which the most salient one is that "[d]ata-driven activities have the advantage of allowing students [trainee translators] to access the real-life language use of their particular context of interest" (p. 477). In this sense, trainee translators can possibly identify the properties and features of collocations (e.g. what to use, how to use, when to use and where to use) when they look at the context in which collocations have occurred. Thus, this method would help translators use 'correct' collocations in the 'correct' situations.

In addition, trainee translators' use of L2 collocations could undergo an evaluation step, which includes both self-evaluation and teachers' evaluation ('evaluation' in Figure 8.3) or evaluative feedback. In self-evaluation or self-reflection (see Bernardini, 2004), trainee translators may retrieve collocations from their translations and compare their use of L2 collocations with that of native speakers. This is to examine whether they have used L2 collocations to which they have been exposed and whether these uses are appropriate in comparison with native norms. The retrieval of collocations can be carried out with the method introduced in this study. Furthermore, teachers could look at trainee translators' production of L2 collocations from two aspects, specifically accuracy and complexity (see for instance Lewis, 2000). In other words, teachers could

carry out an error analysis to demonstrate trainee translators' weaknesses in controlling formal, semantic and functional features in their use of L2 collocations. This step is to reveal translators' inappropriate use of L2 collocations when compared with native norms and give them a better understanding about what they already know and how they can improve. Based on trainee translators' inappropriate use of L2 collocations, teachers could motivate them to 'trace back' the errors in their translations and encourage them to check these errors against the uses in native-speaker texts ('exploration' in Figure 8.3). Trainee translators can use a number of search methods, such as concordance, wordlists, collocates and clusters. In this way, students can generalise rules between collocating words and have a clear idea as to how to identify and recognise these collocating words. This would help them to a great extent use the L2 collocations appropriately and transfer their explicit knowledge to implicit knowledge in the use of L2 collocations. In addition, teachers could motivate trainee translators to take account of context-related aspects including socio-pragmatics and register (Crezee and Grant, 2013) so that they can identify different collocation distribution patterns across different registers and use L2 collocations appropriately. For instance, through exploring the word *economy* in commerce, translators can not only identify the words that *economy* normally collocates with in commercial English but also know how often *economy* collocates with these words based on the statistical values, such as the MI scores. This can be illustrated in Figure 8.4:

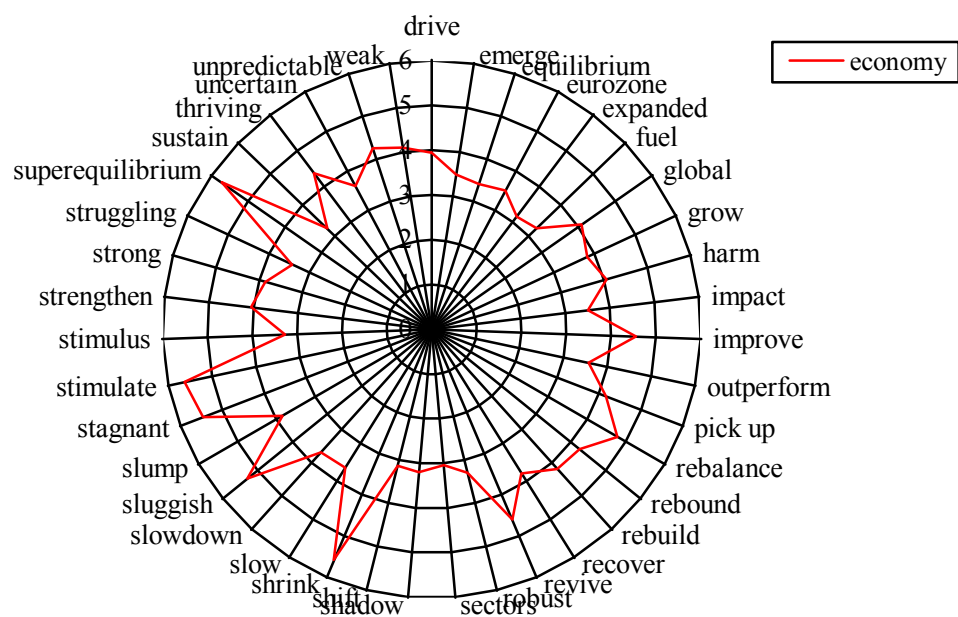


Figure 8.4 Knowledge map of *economy* and its collocates in the NECCD

After the four steps of learning collocations, that is, exposure, exercise, evaluation/evaluative feedback and exploration, some collocations will become integrated into trainee translators' knowledge system, becoming either declarative knowledge or procedural knowledge. Trainee translators are able to paraphrase some L2 collocations they have seen correctly used in native-speaker texts. However, teachers may also find that some L2 collocations are not yet fully mastered by translators after giving feedback to translators. In this case, teachers could encourage trainee translators to repeat the 'four E's' process in class and especially take account of their feedback, till these collocations are fully understood and acquired by trainee translators.

More importantly, it should be noted in these suggestions that L1-L2 contrastive analysis plays a very important role in facilitating the running of the whole 'four E's' process and the understanding teachers' feedback. The valid recognition of L1-L2 differences would not only help translators increase the effectiveness in learning L2 collocations with their teachers' feedback and memorise more collocations in their knowledge system, but also enable them to turn receptive knowledge into productive knowledge. Contrary to this, the insufficient recognition of L1-L2 differences will

become an obstacle which prevents translators from moving forward smoothly in the process. In this respect, both teachers and translators could pay more attention to L1-L2 differences and carry out contrastive analysis appropriately in translator training. This is where the input of translation teachers on students' translation work is essential. Tutors can comment on L1 interference and differences between the L1 and L2 when providing feedback on students' (draft) translations.

8.5 Summary

In response to the fifth research question, this chapter mainly discussed how theoretical findings of this study would be useful for enhancing translators' skills in L1-to-L2 translations. This chapter briefly summarised the theoretical implications of the findings by clarifying the role of collocation in translation based on the proposed theoretical framework. More importantly, it provided suggestions from the practical and pedagogical aspects as to how translators as well as translation teachers would effectively take collocation into account in translator training. For trainee translators, it would be crucially important to realise the important role of collocation in translation, and recognise L2 collocations as wholes in L2 learning and using them as wholes as well in translating. For translation teachers, to help translators achieve these goals, they would need to design an effective pedagogical curriculum, with which they can motivate trainee translators to practise and explore the use of L2 collocations with authentic language materials. In addition, translation teachers would need to provide trainee translators timely and useful feedback regarding the recurring problems in translation practice. Only in such a way can trainee translators transfer their explicit knowledge into implicit knowledge more effectively and efficiently in regard to learning and using L2 collocations.

Chapter Nine Conclusion

9.1 Introduction

This chapter will summarise the major findings of this study and look at the limitations of this study with regard to research design. It will also suggest from what aspects relevant future research in this area might be carried out.

9.2 Summary of major findings

According to the theoretical framework proposed in Chapter 3, I have provided both a quantitative and qualitative analysis in an attempt to explore the distinctive features of the collocation distribution patterns produced by Chinese translators. In the quantitative section, I clarified the relationship between collocations and the indicators of translation universals and proposed that the general features of collocation distribution be examined in terms of degree of collocability, delexicalization and semantic prosody. The analyses of these three aspects correspond to the formal, semantic and functional properties of collocations, thus reflecting the nature of collocations. Based on the data obtained from the quantitative section, the qualitative analysis section examined Chinese translators' use of English collocations from the formal, semantic and functional perspectives respectively. The qualitative analysis section demonstrated the existence of translation universals in the translators' translation outputs. Therefore, it could be concluded that the presence of such translation universals in the target text may have resulted from the translators' insufficient understanding of the features of English collocations and the resulting tendency to 'transfer' collocations from their mother tongue (L1). I will briefly summarise the findings addressing the gaps in the literature as follows.

9.2.1 Overall frequencies

The quantitative analysis section showed that the collocation distribution patterns in the two corpora are very different from the t-score test result. In order to discover the factors leading to such divergences, this section compared the type-token ratios (TTR) in the two corpora and found that the TTR in the corpus of translational English (3.47) was significantly lower than that of native English (6.25). This was also reflected in Chinese translators' over-production of collocation tokens (111,450 – 101,935) compared to a considerably smaller number of collocation types (3,872 – 6,366) in comparison with native speakers. It would appear that Chinese translators produce fewer types of collocations and more of the same collocation tokens

9.2.2 Frequency and statistical values

The quantitative analysis section also used a 'grouping' method with regard to frequency and statistical value, in an attempt to explore the features of collocation distribution in the translational English. It showed that, when compared with native speakers, Chinese translators produced collocations with high frequencies, specifically, those whose occurrences in the TECCTC are greater than 50. Such selection of frequently used collocations resulted in an imbalance between collocation tokens and types in the translational English. The overuse of strong collocations (i.e. frequency \geq 50) also prevented them from increasing collocation variety, thus showing a strong tendency to produce a narrower range of collocation types.

9.2.3 MI score of high-frequency collocations

From the aspect of statistical value, the quantitative analysis section showed that the high-frequency collocations repeatedly used by Chinese translators in the TECCTC mainly possess an MI score greater than 4 (see 5.3). This finding indicates that Chinese translators tended to under-produce weak collocations which, however, contributed to a

large proportion of all the collocation types produced in native English. All these findings indicate that Chinese translators' L1-to-L2 translation outputs contained high numbers of a small selection of collocations, with only a limited range of variety. Furthermore, in order to explore the reasons for the low type-token ratio (TTR) in the TECCTC, the quantitative analysis section also examined the keyword growth rate (see 5.4) by comparing the two corpora, where the keyword growth rate is defined as the tendency for keyword types to increase over segmented text(s) of particular lengths in a corpus. This showed that the slow keyword growth is essentially a main factor influencing Chinese translators' production of L2 English collocations. This finding indicates that for Chinese translators to produce native-like outputs in their L2 English translations, they not only need to master a large L2 vocabulary but also recognise various collocational relationships between words and use collocating words (collocations) appropriately in their L2 English.

9.2.4 Formal analysis

In order to further explore the features of collocation distribution patterns in translational English, the quantitative analysis section also looked at the comparison between the two corpora from the formal, semantic and functional perspectives. The formal analysis section presented a model of collocational continuum regarding the degree of collocability or the level of association. I found that Chinese translators' outputs showed a large proportion of free combinations, rather than bound collocations or idioms in comparison with those in similar texts produced by native speakers of English. In particular, bound collocation types (1,819 types) produced by Chinese translators account for a mere one third of the number of types normally produced by native speakers of English (5,197 types). This indicates that the formulaic nature of collocation in Chinese translators' L1-to-L2 translation outputs still remains at a comparatively low level because bound collocations and idioms showed a stronger lexical association than free combinations. This finding, to a great extent, supports Wray's (2002) proposal that L2 language learners basically employ a non-formulaic

approach in language learning and production, and tend to divide lexical chunks into individual words and memorise those words separately when they are exposed to their L2 input. This finding also implies that over-production of free combinations may bring some translation universal features (or translation universals), particularly simplification, in the target text.

9.2.5 Semantic analysis

The semantic analysis section looked at the distinction between the collocation distribution patterns of the two corpora from the angle of delexicalization. Collocations were categorised into two groups, specifically, those with a literal sense and those with a delexicalized sense. The empirical results showed that Chinese translators under-used delexical collocations when compared with native speakers of English, which holds true both in token and in type. This finding might lead one to speculate that, on the one hand, when translators accumulate their L2 vocabulary knowledge they may be more inclined to focus on the literal side of words but largely fail to notice the ‘depth’ side of exploring the pragmatic meanings of words. This, on the whole, reduces their use of collocations to a narrower range in comparison with native norms. On the other hand, this finding also indicates that inadequate knowledge of delexicalization in the L2 may result in the presence of translation universals in target texts, particularly explicitation.

9.2.6 Functional analysis

The functional analysis section used semantic prosody to examine whether the functions are performed appropriately in the translational English produced by Chinese translators. Semantic prosodies were divided into three categories, namely positive semantic prosody, neutral semantic prosody and negative semantic prosody. The results from quantitative and statistical analyses showed that Chinese translators used more collocations with neutral semantic prosodies, which account for 70.82% by token and 63.97% by type, whereas native speakers of English produced collocations with neutral

semantic prosodies to a lesser extent and nearly half of the collocations in L1 English are assigned with either positive or negative semantic prosodies. This finding indicates that, in comparison with native speakers of English, Chinese translators over-conformed to neutralised collocations, which would decrease the effectiveness to achieve naturalness in commercial English. More importantly, this kind of deviation in controlling semantic prosodies in translational English would also increase the possibility of introducing translation universals, particularly normalisation, into the target text.

9.2.7 Contrastive Interlanguage Analysis

Based on the features of collocation distribution patterns found in the TECCTC, the data analysis sections (Chapter Five and Chapter Six) employed the Contrastive Interlanguage Analysis approach (CIA) by comparing the two corpora. In particular, the explanatory section (Chapter Seven) exemplified a few instances as to how L2 English was made explicit, simplified and normalised by Chinese translators in the target text. The findings from these sections, as a whole, have provided empirical evidence to support the hypothesis of translation universals. These examples attempted to clearly demonstrate the importance of recognising the features of collocation distribution patterns used in the native language, that is, the appropriate control of collocability, delexicalization and semantic prosodies in using collocations. The control of these factors is strongly connected with translators' ability to cope with translation universals and produce native-like target texts. An appropriate control of these features may indicate that translators are resourceful (able to find right collocations) and reflective (able to reflect on their own practice) (Bernardini, 2004) when they understand these collocation features, and this would in turn help reduce translation universals in the target language. In contrast, an inappropriate control of these features may indicate that translators' understanding of these collocation features is superficial, and would probably increase the possibility of translators introducing translation universals into the target language. In addition, the explanatory section exhibited, with examples, that L1

transfer from Chinese is the key factor influencing Chinese translators' control of the features of English collocations, which may be largely due to their inadequate knowledge of L1-L2 differences.

Generally speaking, the findings from the quantitative and qualitative analyses have provided evidence to demonstrate that the use of collocations plays a very important part in language production. In some sense, the use of collocations can be a benchmark to measure how well an L2 user is accepted by a particular speech community where that L2 is used. In the present study, Chinese translators' deviation in using English collocations implies that their L2 skills, to some extent, have not reached the native level and there are still some important issues left to be addressed in translator training. In this respect, it is important for researchers working in this area to take advantage of their theoretical findings and provide effective solutions to the problems in translations.

9.3 Limitations of the present study

The limitations of the present study can be demonstrated in a number of ways, specifically, the conceptual framework of collocation, the collection of corpus materials, the data retrieval procedure and the research methodology.

The notion 'collocation' is defined based on the Bigram Model, in which word pairs are required to occur at least five times in the corpus and show the statistical significance with the MI test and the Log-likelihood test if they are to qualify as successful candidates of collocation. However, this approach rules out some 'meaningful' collocations whose occurrences are less than five in the collocation retrieval procedure, such as *fiscal stance* and *fiscal shortfall* in the NECCD, even though this may increase the reliability of the data retrieved. In addition, the data retrieval criteria applied to 2-word collocations only but overlooked some more complicated situations, such as 3-word collocations and 4-word collocations, based on the design of this study. Therefore, more in-depth considerations based on the N-gram Model should be given in

this type of research.

In addition, this study attempted to employ a corpus-driven approach as the research methodology, with the comparable corpora built up being approximately 5 million English running words in size. However, the corpus-driven approach, when compared with the corpus-based approach, normally requires very large corpora and attempts to filter the data through apparently random sampling (McEnery, Xiao & Tono, 2006). Furthermore, it would be hard to exploit data fully and maintain the integrity of data if a small-sized corpus is employed (McEnery, Xiao & Tono, 2006). In this sense, in corpus-driven studies, the larger the corpora chosen the better the results will be because larger corpora will allow more precise and accurate observations to be made. Because it is very difficult to find an authoritative corpus of commercial English, the NECCD and the TECCTC are self-built and basically meet the requirements of the current study. Nevertheless, the size of these two corpora is still relatively small compared to some other existing corpora, such as the British National Corpus (BNC, approximately 100 million words) and the Bank of English (approximately 650 million words as of 2012 and still increasing in size). Therefore, larger corpora of commercial English could be built in future studies to do justice to the corpus-driven approach.

Furthermore, the data retrieval section used both FoxPro programming and the BFSU (Beijing Foreign Studies University) Collocator to fulfil the task of extracting collocations. The former provided the method of retrieving all the bigrams from the corpora while the latter offered statistical tests to measure the significance for word pairs. Nonetheless, this procedure appears to be quite time-consuming because word pairs are examined individually. Therefore, it would be less time-intensive if researchers were to incorporate the MI computation formula and the Log-likelihood formula into the FoxPro programme of retrieving bigrams, and extract collocation candidates automatically from the corpus without examining these word pairs one by one. However, this would require a lot more sophisticated technical support, and even if there was such a programme, it would involve large-scale mathematical computation.

A final limitation of the study is that I only had the translation outputs at my disposal, without any information about the conditions Chinese translators worked under: what the deadlines were, whether translators were able to do first and second drafts for these commercial translations, and whether they had any proofreader input (from either native English proofreaders or L2 English proofreaders) or no proofreader input at all. In other words, the information about the deadlines, the first and second drafts and the proofreader input was not presented in the current study. For example, if translators have to do a rush job, they may not produce as many of the correct collocations. However, if they are given the opportunity to leave a translation aside for a few hours or even a day, and then read only the target text (i.e. without being exposed to a second round of L1 interference), they may come up with the correct collocations (Crezee, 2014).

9.4 Directions for future research

The present study has looked at Chinese translators' use of L2 English collocations in commercial Chinese-to-English translation, and describes collocation distribution patterns in the translational English. With regard to the research design, this study has elaborated on a theoretical framework and examined the use of L2 English collocations by Chinese translators from four aspects, that is, quantity, form, meaning and function, through comparing two designed corpora. Both quantitative and qualitative findings have proved the validity of the proposed theoretical framework in this study. Based on the findings, this study also proposes a pedagogical model in an attempt to help enhance translators' proficiency in the use of L2 collocations. Nevertheless, future research is still needed to confirm the conclusions in this study and such research can be conducted from the 'horizontal' and 'vertical' perspectives. Future research would gain additional dimensions if the researcher had information about the translators' backgrounds, training, working conditions and whether or not (native English) proof-reader input was available to them.

Another interesting direction for future research might be to use an Action Research approach by testing pedagogical interventions of the type described in Section 8.3 above. One group of students could be given feedback on their use of collocations over the course of a semester, while the other group of students not given such feedback, and results of an end-of-semester translation task could then be used to compare collocation use between groups.

From the ‘horizontal’ perspective, this study has provided a research design which can be replicated by other researchers to investigate the use of English collocations taken from larger corpora in other registers of Chinese-to-English translations, such as academic translation, literary translation, legal translation, medical translation, political translation and general translation. In addition, while researchers attempt to examine whether the findings of this study hold true for other registers, I believe this research design can also be used to investigate translations which involve other language pairs. Thus, researchers can examine whether the theoretical framework proposed in this study can be universally accepted in Translation Studies. In this respect, researchers can further explore whether translators whose L1 background is not Chinese produce similar L2 collocation distribution patterns as found in this study, and rely on the strategy of L1 transfer at times and produce translation universals in the target text. In other words, research from the ‘horizontal’ perspective can be used to examine the reliability of the theoretical framework of this study with a wider range of empirical evidence.

From the ‘vertical’ perspective, researchers can focus on commercial Chinese-to-English translations and conduct research in three directions. First of all, researchers may investigate L2 English collocations based on larger comparable corpora (e.g. 100-million-token corpora) and attempt to explore more complicated collocational relationships, such as 3 or 4 word collocations or chunks. In this way, researchers can better examine the reliability and validity of the theoretical framework proposed in this study, and further explore whether the 2-word collocation distribution patterns produced by Chinese translators in this study still hold true for more complex situations.

Secondly, because this study has proposed a comparatively complicated procedure of data retrieval, other researchers may attempt to simplify and facilitate this procedure. As discussed in Section 8.4, researchers can attempt to incorporate the MI and Log-likelihood formulae into the *FoxPro* programmes of retrieving bigrams, and extract collocation pairs automatically from corpora without examining these candidates one by one. Alternatively, researchers with special knowledge of computer programming skills may wish to develop another set of programmes with another computer programming language (e.g. *Perl*), which performs the same function as described above.

Finally, this study presents a list of collocations used in native commercial English and provides some suggestions for teaching and learning these collocations. In this respect, researchers can launch more projects to investigate the validity of these suggestions and examine whether trainee translators, as well as developing EFL learners, can essentially enhance their L2 proficiency by adopting these suggestions. In addition, future research should always incorporate information about translators' background and working conditions, so as to include both translation outputs and the conditions under which such outputs were achieved.

To sum up, the present study has demonstrated the importance of identifying the role of collocation and the distribution patterns in both translated and native-speaker texts. This will not only enable developing translators to familiarise themselves with the collocational relationship between lexical items but also help them overcome the shortcomings they may have in using L2 collocations in L1-to-L2 translations. This may help L2 users have a clear idea as to how they can enhance their competence in handling L2 collocations and substantially increase their L2 proficiency based on the 'real-life' language-use strategy suggested in this study. It is hoped that the findings of this study will make a meaningful contribution to the curriculum design focusing on the use of L2 English collocations and encourage other researchers in this research area to further explore the use of corpora to benefit Translation Studies.

References

- Aijmer, K. (1996). *Conversational routines in English*. London & New York: Longman.
- Aitchison, J. (1987). Reproductive furniture and extinguished professors. In Ross Steele & Terry Threadgold (Eds.), *Language Topics. Essays in Honour of Michael Halliday*, 2 (pp. 3-14). Amsterdam, Netherlands: John Benjamins.
- Bahns, J. & Eldaw, M. (1993). Should we teach EFL students collocations?. *System*, 21 (1), 101-114.
- Bai, C. (2014). Personal communication: Email to Ineke Crezee.
- Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis and E. Tognini-Bonelli (Eds.), *Text and Technology: in Honour of John Sinclair* (pp. 233-250). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation Studies in Language Engineering: In Honor of Juan C. Sager* (pp. 175-186). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Baker, M. (2001). *Routledge Encyclopaedia of Translation Studies*. London, England: Routledge.
- Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2), 167-193.
- Barfield, A., & Gyllstad, H. (Eds.) (2009). *Research collocations in another language: Multiple interpretations*. Hampshire, England & New York, USA: Palgrave Macmillan.
- Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage-based models of language*. Stanford, CA: CSLI Publications.
- Barnbrook, G. (2007). Sinclair on collocation. *International Journal of Corpus Linguistics*, 12(2), pp. 183-199.
- Battus, H. (1983). *Rekenen op taal*. Amsterdam, Netherlands: Querido.
- Bell, R.T. (1991). *Translation and translating*. London, England: Longman.
- Ben-Shahar, R. (1994). Translating literary dialogue: a problem and its implications for translation into Hebrew. *Target* 6, 195-221.
- Bernardini, S. (2004). Corpora in classrooms: An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15-38). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3), 275-311.
- Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House, & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17-35). Tübingen, Germany: Gunter Narr.

- Blum-Kulka, S. & Levenston, E. (1983). Universals of lexical simplification. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (119-139). London, England: Longman.
- Bowker, L. & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London, England: Routledge.
- Butler, C. S. (1985). *Statistics in linguistics*. Oxford, England: Basil Blackwell.
- Butler, C. S. (1997). Repeated word combinations in spoken and written text: some implications for Functional Grammar. In C. S. Butler, J. H. Connolly, R. A. Gatward & R. M. Vismans (Eds.) *A fund of ideas: recent developments in Functional Grammar*. Amsterdam, Netherlands: IFOTT, 60 - 77.
- Campbell, S. (1991). Towards a model of translation competence. *Meta* 36(2-3), 329-43.
- Campbell, S. (1998). *Translation into the second language*. London, England: Longman.
- Carter, R. & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, 16(2), pp. 141-158.
- CATTI. (2014). *Interim Regulations on Senior Translator/Interpreter and Translator/Interpreter Professional Qualification/Level I Assessment, 2011*. Retrieved from http://www.catti.net.cn/2011-05/03/content_354389.htm
- Chen, W. (2006). *Explication through the use of connectives in translated Chinese: A corpus-based study*. Unpublished doctoral thesis, University of Manchester, Manchester, England.
- Chesterman, A. (2004). Beyond the particular. In A. Mauranen and P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 33-49). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), 22-29.
- Clear, J. (1993). From Firth principles: Computational Tools for the Study of Collocation. In M. Baker, G. Francis and E. Tognini-Bonelli (Eds.), *Text and Technology: in Honour of John Sinclair* (pp. 271-292). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Chomsky, N. (1965). *Aspects of a theory of syntax*. Cambridge, MA: MIT Press.
- Cohen, A. (2011). *Strategies in Learning and Using a Second Language (2nd edition)*. Harlow, England: Longman.
- Coulmas, F. (1981). Introduction: conversational routine. In F. Coulmas (ed.) *Conversational routine* (pp. 1-17). The Hague, Netherlands: Mouton.
- Cowie, A. P. (1981a). Lexicography and its pedagogic applications: An introduction. *Applied linguistics*, 2(3), pp. 203-206.
- Cowie, A. P. (1981b). The treatment of collocations and idioms in learners' dictionaries. *Applied linguistics*, 2(3), 223-235.
- Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy (Eds.) *Vocabulary and language teaching* (pp. 126-139). London & New York: Longman.
- Cowie, P. (1992). Multiword lexical units and communicative language teaching. In Pierre J. L. Arnaud & Henri Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 1-12). Houndsmills, England: Macmillan.

- Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopaedia of language and linguistics* (pp. 3168-3171). Oxford, England: Pergamon.
- Cowie, A. P. (1997). Phraseology in formal academic prose. In Jan Aarts et al. (Eds.), *Studies in English and Language Teaching*, 43-56.
- Crezee, I. H. M. (2013). *Introduction to healthcare for interpreters and translators*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Crezee, I., & Grant, L. (2013). Missing the plot? Idiomatic language in interpreter education. *International Journal of Interpreter Education*, (5) 1, 17-33.
- De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London, England: Longman.
- De Haan, Pieter & van Hout, Roeland. (1986). A loglinear analysis of syntactic constraints on postmodifying clauses. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora* (pp. 79-97). Amsterdam, Netherlands: Rodopi.
- De Sutter, G., Delaere, I., & Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modelling. In M. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies* (pp. 325-346), Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Deng, Y. (2007). *An Investigation into Formulaic Sequences in Chinese EFL Learners' Spoken English*. Unpublished doctoral thesis, Shanghai Jiao Tong University, Shanghai, China.
- Dimitrova, B. E. (2005). *Expertise and explicitation in the translation process*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Durrant, P. (2008). *High frequent collocations and second language learning*. Unpublished doctoral thesis, University of Nottingham, Nottingham, England.
- Durrant, P. & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research* 26(2), 163-188.
- Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge, England: Cambridge University Press.
- Ellis, N. C. (2003). Constructions, chunking, and connectionism: the emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Oxford, England: Blackwell.
- Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27(2), 305-352.
- Ellis, N. C., & Larsen-Freeman, D. (2006). Language Emergence: Implications for Applied Linguistics-Introduction to the Special Issue (Vol. 27, pp. 558-589).
- Evert, S. (2004). Computational approaches to collocations. Retrieved 14 December, 2007, from www.collocations.de
- Ellis, R. (1994). *The study of second language acquisition*. Oxford, England: Oxford University Press.
- Fan, F (2006). A corpus-based empirical study on inter-textual vocabulary growth. *Journal of quantitative linguistics*, 1, 111-127.

- Fan, F. (2010a). *Data processing and management for quantitative linguistics with Foxpro*. Lüdenscheid, Germany: RAM-Verlag.
- Fan, F. (2010b). *Quantitative linguistic computing with Perl*. Lüdenscheid, Germany: RAM-Verlag.
- Fan, M. (2009). An exploratory study of collocational use by ESL students-A task based approach. *System*, 37, 110-123.
- Farghal, M., & Obeidat, H. (1995). Collocations: a neglected variable in EFL. *International review of applied linguistics in language teaching*, 33(4), 315-331.
- Feng, H. (2010). *A corpus-based study on multi-word verbs used by Chinese translators in political document translations*. Unpublished masteral dissertation, University of Auckland, Auckland, New Zealand.
- Fillmore, C. J. (1979). On fluency. In Charles J. Fillmore, Daniel Kempler, & William S.-Y. Wang (Eds.), *Individual Differences in Language Ability and Language Behaviour* (pp. 85-101). New York: Academic Press.
- Firth, J. R. (1957). *Papers in Linguistics 1934-1951*, 190-215. London, England: Oxford University Press.
- Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected papers of J.R. Firth 1952-1959* (pp. 168-205). Harlow, England: Longman.
- 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 & M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 75-94). London, England: Longman.
- Gellerstam, M. (1986). Translationese in Swedish Novels Translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation Studies in Scandinavia* (pp. 88-95). Lund, Sweden: CWK Gleerup.
- Gellerstam, M. (2005). Fingerprints in Translation. In G. Anderman & M. Rogers (Eds.), *In and Out of English: For Better, For Worse?* (pp. 201-213). Clavedon, England: Multilingual Matters.
- Gilquin, G. (2000/2001). The integrated contrastive model: Spicing up your data. *Languages in contrast* 3(1), 95-124.
- Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In Gilquin, G., Papp, S., & Díez-Bedmar, M. B. (Eds.), *Linking up contrastive and learner corpus research*. Amsterdam, Netherlands: Rodopi.
- González Davies, M. (1998). Student assessment by medical specialists. In Henry Fischbach (Ed.), *Translation and machine*, 93-104. Amsterdam, Netherlands: John Benjamins.
- González Davies, M. (2004). *Multiple voices in the translation classroom*. Amsterdam, Netherlands: John Benjamins.
- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast, Papers from a Symposium on text-based cross-linguistic studies, Lund 4-5 March 1994 [Lund Studies in English 88]* (pp. 37-51). Lund, Sweden: Lund University Press.

- Grant, L. & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics* 25(1), 38-61.
- Grant, L. E. (2005). Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics* 10(4), 429-451.
- Gries, S. Th. (2010). Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics* 15(3), 327-343.
- Hansen, G. (1997). Success in translation. *Perspectives: Studies in Translatology* 5(2), 201-10.
- Halliday, M. A. K. (1961). Categories of the Theory of Grammar. *Word* 17(3), 241-92.
- Halliday, M. A. K. (1978). *Language as social semiotic*. London, England: Edward Arnold.
- Halliday, M. A. K., McIntosh, M. & Stevens, P. (1964). *The linguistic sciences and language teaching*. London, England: Longman.
- Hasselgård, H. & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- 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, 237-259.
- Hatim, B. & Mason, I. (1997). *The translator as communicator*. London & New York: Routledge.
- Herbst, T. (1996). What are collocations: sandy beaches or false teeth? *English Studies*, 4, 379-393.
- House, J. (2008). Beyond intervention: Universals in translation? *Trans-kom*, 1(1), 6-19.
- Howarth, P. (1998a). Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1), 24-44.
- Howarth, P. (1998b). The Phraseology of Learners' Academic Writing. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications* (pp. 161-86). Oxford, England: Oxford University Press.
- Hunston, S. & Francis, G. (2000). *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Irujo, S. (1986). A piece of cake: learning and teaching idioms. *ELT Journal* 40(3), 236-242.
- James, C. (1980). *Contrastive Analysis*. London & New York: Longman.
- Jones, S. & Sinclair, J. (1974). English lexical collocations. *Cahiers de Lexicologie*, 24, 15-61.
- Jurafsky, D., & Martin, J. (2000). *Speech and Language Processing. An Introduction to Natural Language Process, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice-Hall.
- Kennedy, G. D. (1990). Collocations: Where grammar and vocabulary teaching meet. In Sarinee A. (Ed.), *Language Teaching Methodology for the Nineties. Anthology Series 24* (pp. 215-229). Singapore: SEAMEO Regional Language Centre.
- Kenny, D. (1998). Creatures of habit? What translators usually do with words. *Meta*,

43(4), 515-523.

- Kenny, D. (2000). Translation at play: Exploitations of collocational norms in German-English translation. In B. Dodd (Ed.), *Working with German corpora* (pp. 143-160). Birmingham, England: University of Birmingham Press.
- Kenny, D. (2001). *Lexis and Creativity in Translation: A Corpus-based Study*. Manchester, England: St. Jerome Publishing.
- Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111-127). London, New York: Longman.
- Köhler, R. & Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In G. Altmann, W.A. Koch (Eds.), *Systems. New paradigms for the human sciences* (pp.514-546). Berlin, Germany: de Gruyter.
- Lakoff, G. (1987). *Women, fire and dangerous things*. Chicago, Illinois: University of Chicago Press.
- Laufer, B. & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22, 15-30.
- Laviosa, S. (1998a). The corpus-based approach: A new paradigm in translation studies. *Meta*, 43(4), 474-479.
- Laviosa, S. (1998b). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4), 557-570.
- Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, application*. Amsterdam, Netherlands: Rodopi.
- Laviosa-Braithwaite, S. (1997). Investigating simplification in an English comparable corpus of newspaper articles. In K. Klaudy & J. Kohn (Eds.), *Transfere necesse est. Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting* (pp. 531-540). Budapest, Hungary: Scholastica.
- Lewis, M. (2000). *Teaching Collocation: Further Developments in the Lexical Approach*. London, England: Language Teaching Publications.
- Liu, N and Nation, P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16, 3-42.
- Louw, B. (1993). 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies' in M. Baker, G. Francis, and E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 157-76). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Louw, B. (2000). 'Contextual prosodic theory: Bringing semantic prosodies to life' in C. Heffer, H. Sauntson, and G. Fox (Eds.), *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham, England: University of Birmingham.
- Low, P. A. (2003) Translating poetic songs: an attempt at a functional account of strategies. *Target* 15(1), 91-110.
- Malmkjær, K. (1997). Punctuation in Hans Christian Andersen's stories and their translations into English. In F. Poyatos (Ed.), *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media* (pp. 151-162). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.

- Malmkjær, K. (1998). Love thy neighbour: Will parallel corpora endear linguists to translators?. *Meta* 43(4), 534-541.
- Malmkjær, K. (2007). Norms and nature in translation studies. In M. Rogers & G. Anderman (Eds.), *Incorporating Corpora. The Linguist and the Translator* (pp. 49–59). Clevedon, England: Multilingual Matters.
- Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Mauranen, A. (2007). Norms and nature in translation studies. In M. Rogers and G. Anderman (Eds.), *Incorporating corpora: The linguistics and the translator* (pp. 32-48). Clevedon, England: Multilingual Matters.
- Mauranen, A. & Kujamäki, P. (2004). *Translation universals: Do they exist?*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- May Fan, C. (1999). An Investigation into the Pervasiveness of Delexical Chunks in Authentic Language Use and the Problem they Present to L2 Learners. In Berry et al. (Eds.), *Language Analysis, Description, and Pedagogy*. Hong Kong, China: Language Center, The Hong Kong University of Science and Technology.
- McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh, England: Edinburgh University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An advanced resource book*. London & New York: Routledge.
- McEnery, T. & Xiao, R. (2007). Parallel and comparable corpora: What is happening? In M. Rogers and G. Anderman (Eds.), *Incorporating corpora: The linguistics and the translator* (pp. 18-31). Clevedon, England: Multilingual Matters.
- Meunier, F. & Granger, S. (Eds.) (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge, England: Cambridge University Press.
- Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge, England: Cambridge University Press.
- Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy*. (pp. 6-19). Cambridge, England: Cambridge University Press.
- Nattinger, J. R. & DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford, England: Oxford University Press.
- Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam, Netherlands: John Benjamins.
- Neubert, A. (2000). Competence in language, in languages, and in translation. In C. Schäffner, & B. Adabs (Eds.) *Developing translation competence* (pp. 3-18). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Nitschke, S., Kidd, E. & Serratrice, L. (2010). First language transfer and long-term structural priming in comprehension. *Language and Cognitive Processes* 25(1), 94-114.

- Nord, C. (1992). Text analysis in translator training. In C. Dollerup, & A. Lindegaard (Eds.) *Teaching translation and interpreting 1* (pp. 39-48). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Olohan, M. (2004). Introducing corpora in translation studies. London & New York: Routledge.
- Olohan, M. & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious process of explicitation?. *Across Languages and Cultures*, 1(2), 141-158.
- Oxford Collocations Dictionary for Students of English* (2002). Oxford, England: Oxford University Press.
- Oxford University (2005). *British National Corpus*. Available at: <http://www.natcorp.ox.ac.uk/>.
- Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta*, 43(4), 557-570.
- PACTE. (2003). Building a Translation Competence Model. In Fabio Alves (Ed.), *Triangulating Translation: Perspectives in Process Oriented Research* (pp. 43-66). Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Paradis, M. (2004). A neurolinguistic theory of bilingualism. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Partington, A. (1996). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Partington, A. (2004). Utterly content in each other's company: Semantic prosody and semantic preference, *International Journal of Corpus Linguistics*, 9(1), 131-156.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.
- Pawley, A. & Syder, F. H. (2000). The One-Clause-at-a-Time Hypothesis. In Heidi Riggensbach (Ed.), *Perspectives on Fluency*. Ann Arbor, Michigan: University of Michigan Press.
- Peters, A. M. (1977). Language-learning strategies: does the whole equal the sum of the parts? *Language*, 53(3), 560-573.
- Pym, A. (2003). Redefining Translation Competence in an Electronic Age: In Defence of a Minimalist Approach. *Meta* 48(4), pp. 481-97.
- Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). London, England: Longman.
- Robinson, D. (2003). *Becoming a translator: an introduction to the theory and practice of translation*. London & New York: Routledge.
- Schmitt, N. (ed.) (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Schmitt, N, Jiang, X and Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal* 11, 26-43.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, 10, pp. 209-241.

- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, England: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Lexis, corpus, discourse*. London & New York: Routledge.
- Sinclair, J., & Renouf, A. (1998). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160). London & New York: Longman.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.
- Snedecor, G. W., & Cochran, W. G. 1989. *Statistical methods*. Ames, Iowa: Iowa State University Press.
- Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 23-55.
- Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, England: Blackwell Publishers.
- Stubbs, M. (2005). The most natural thing in the world: quantitative data on multi-word sequences in English. Conference presentation at *Phraseology 2005*. Louvain-la-Neuve, Belgium.
- Sugiura, M. (2002). Collocational knowledge of L2 learners of English: A case study of Japanese Learners. In T. Saito, J. Nakasura & S. Yamazaki (Eds.), *English Corpus Linguistics in Japan* (pp. 303-323). Amsterdam, Netherlands: Rodopi.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work: Studies in corpus linguistics*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Toury, G. (1980). *In Search of a Theory of Translation*. Tel Aviv, Israel: The Porter Institute for Poetics and Semiotics.
- Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam, Netherlands & Philadelphia, Pennsylvania: John Benjamins.
- Trebits, A. (2009). The most frequent phrasal verbs in English language EU documents-A corpus-based analysis and its implications. *System*, 37, 470-481.
- Tymoczko, M. (1998). Computerized corpora and the future of translation studies. *Meta*, 43(4), 652-660.
- Vanderauwera, R. (1985). *Dutch Novels Translated into English: The Transformation of a "Minority" Literature*. Amsterdam, Netherlands: Rodopi.
- Wray, A. (2000). Formulaic sequences in second language teaching: principles and Practice. *Applied Linguistics*, 21(4), 463-489.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press.
- Wray, A. & Perkins, M. (2000). The functions of formulaic language: all integrated Model. *Language and Communication*, 20, 1-28.
- Xiao, R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1), 5-35.
- Xiao, R. & McNery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1), 103-129.
- Xiao, Z. & Xu, J. (2008). Corpus and language education. *Foreign language education in China*, 1(2), pp. 48-58.

- Xu, Jiajin & Jia, Y. (2009). Collocator 1.0: A collocation extraction tool. Beijing, China: The National Research Centre for Foreign Language Education, Beijing Foreign Studies University.
- Yorio, C. A. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly* 14(4), 433-442.
- Zanettin, F. (2013). Corpus methods for Descriptive Translation Studies. *Procedia Social and Behavioral Sciences* 95, 20-32.

Glossary

collocability	the probability of words co-occurring with each other
collocation	a prefabricated, structurally coherent and semantically complete lexical combination consisting of at least two words, whose occurrence is more frequent than by chance in the commercial discourse and can show statistical significance quantitatively
corpus	a collection of machine-readable texts, compiled for linguistic research purposes
delexicalization	the linguistic tendency a reduction of the distinctive contribution made by that word to the meaning
explicitation	the tendency of translators making implicit information in the source language explicit in their translations where such implicit information does not need to become explicit in the target text
normalisation	the tendency to exaggerate features of the target language and to conform to its typical patterns
semantic prosody	a consistent aura of meaning with which a form is imbued by its collocates or a form of meaning which is established through the proximity of a consistent series of collocates; semantic prosodies constituted by collocations are normally categorised as positive prosody, neutral prosody and negative prosody
simplification	the tendency of the language in translated texts being simpler than that in the same target language
translation universals	the inherent features revealed in the translated texts, independent of source language, which can essentially distinguish translational language from native-speaker language, such as explicitation, simplification, normalisation, sanitisation, levelling out/convergence, under-representation and so forth

List of appendices

Appendix A:	Assessment of Translators in China	224
Appendix B:	Sample Texts from the Corpora	227
Appendix C:	The FoxPro Programme for Retrieving Bigrams (exemplified with the TECCTC)	230
Appendix D:	Top 30 Most Frequently Used Collocations from the Two Corpora	231
Appendix E:	Source codes for the Perl programme of text chunk segmentation and computation of vocabulary growth	233
Appendix F:	Source codes for the Perl programme of lemmatisation	238
Appendix G:	Source codes for the Perl programme of computing keyword growth	240

Appendix A: Assessment of Translators in China

In reference to the eligibility for the professional title as senior translator, as was first stipulated by the Ministry of Foreign Affairs in 1982, translators should have built up rich experience in translation, proofreading or finalising translated texts with broad scientific and cultural knowledge. They should also have published some translation works with the versions consistent with the original and have enjoyed a good reputation in the translation industry.

In 1986, in line with Interim Regulations on Assessment of Translator/Interpreter Professional Titles, adopted by the Professional Titles Reform Leader Group of the Central Government, there were four professional levels for the translation sector: assistant translator, translator, associate senior translator (proofreader), and senior translator (proofreader). This is summarised as follows:

Assessment of translator professional titles in China in 1986

Title	Requirements
Assistant translator	Undergraduates majoring in a foreign language; undergraduates with two bachelors; non-language postgraduates; junior college graduates with three years of translation experience.
Translator	Systematic basic foreign language specific knowledge; certain scientific and cultural knowledge; considerable achievements; a second foreign language; a doctoral degree; a master with two years of translation experience; a holder of the assistant translator for 4 years.
Assistant senior translator	Considerable experience; considerably broad scientific and cultural knowledge; some translation studies; considerable ability of comprehension and expression; a second foreign language; a postgraduate or undergraduate who has held 2-3 to 5 years.
Senior translator	Long engagement in translating, proofreading and finalising; broad scientific and cultural knowledge; rich experience; some publications, in the consistent style with the original; a holder of the assistant senior translator (proofreader) professional title for 5 years or longer.

(Quoted from the official website of CATTI, translated by Chongshun Bai in 2014)

In 1994, a scoring system was established in consideration of three indicators: quality, difficulty and weighting co-efficient. This assessment approach was found to be more efficient, easier for statistics, and fairer for the assessment outcome with little interference from subjective factors. This approach is illustrated below:

Scoring system of translator professionalism assessment

Translating item	Quality	Difficulty	Weighting coefficient	Holistic scores
Foreign to Chinese translating	A1	B1	D1	$C1=1/2(A1+B1)D1$
Chinese to Foreign translating	A2	B2	D2	$C2=1/2(A2+B2)D2$
Chinese to Foreign proofreading	A3	B3	D3	$C3=1/2(A3+B3)D3$
Foreign to Chinese proofreading	A4	B4	D4	$C4=1/2(A4+B4)D4$
Chinese and Foreign compiling	A5	B5	D5	$C5=1/2(A5+B5)D5$
Chinese and Foreign papers and books	A6	B6	D6	$C6=1/2(A6+B6)D6$
Second foreign language achievement	A7	B7	D7	$C7=1/2(A7+B7)D7$
Interpreting achievement	A8	B8	D8	$C8=1/2(A8+B8)D8$
The translator's score				$Y = C \cdot 65\%$

(Quoted from the official website of CATI, translated by Chongshun Bai in 2014)

As of 2012, annual assessment for the professional title was carried out at 34 branches of Foreign Affairs Office nationwide. More quantitative and scientific approaches to the assessment are being explored. Up to date, there are at least three main translation qualification assessment tests offered in mainland China, namely China Accreditation Test for Translators and Interpreters (CATI), National Accreditation Examinations for Translators and Interpreters (NAETI) and Shanghai Interpretation Accreditation Test. Among these tests, CATI appears to be the most influential and acceptable one in the translation industry of China. Tens of thousands of trainee translators sit CATI tests in China every year to assess their translation competence and skills. There are three levels in CATI, that is, Level III, Level II and Level I with Level I being the most difficult one. The tests for Level III and Level II are offered twice a year, in May and November respectively. These two tests are suitable for undergraduates and postgraduates in universities, or trainee translators with less than five years of experience in translation and interpreting. The test for Level I is administered only once a year, and is suitable for specialist translators with over 10 years of experience in the profession. Examinees who passed CATI Level I are the equivalent of assistant senior translators or senior translators depending on their translation/interpreting performance. In addition, the successful candidate could be promoted to senior translators through additional assessment.

According to *Interim Regulations on Senior Translator/Interpreter and Translator/Interpreter Professional Qualification/Level Assessment*, the eligibility for the essential qualities and professional competence as senior translator and interpreter includes the following:

1. Rich translating and/or interpreting experience and extraordinary professional ethics;
2. Broad and in-depth knowledge, familiarity with Chinese and relevant country's cultural background, and high bilingual proficiency;
3. Qualification for demanding translation tasks, ability to deal with tough

problems, to proofread and finalise important translation texts, or to interpret for key negotiations and international conferences;

4. Rigorous translation attitude, in the same style of the source text;
5. In-depth knowledge about Translation Studies, showing professionalism in accomplishment of translation tasks, excellence in translator training;
6. Brilliant achievement in translation, satisfactory or outstanding usual performance and annual assessment (quoted from the official website of CATI, translated by Chongshun Bai in 2014).

Appendix B: Sample Texts from the Corpora

Sample text from the NECCD

The degree of economic integration among countries has important implications for the exchange rate regime they choose. Countries that are highly integrated with each other with respect to trade and other economic and political relations and have high labour mobility, symmetric shocks, and high income correlation are likely to constitute an optimum currency area (OCA). It is beneficial for these countries to establish regional cooperation on exchange rate policy. Because integration substantially reduces the benefits of their own monetary policy, small countries are better off pegging their currencies to a large neighbour's or adopt a neighbour's currency as their own. These arrangements would reduce transaction costs and interest rates, eliminate exchange risks, and encourage further integration and growth. In countries satisfying OCA conditions, but where a regional common currency is not politically feasible, for example in East Asia, McKinnon (1999) advises establishing efficient common monetary rules to stabilize their exchange rates to avoid competitive devaluation under a common dollar peg.

There are three main approaches to regional exchange rate cooperation. One approach is Mutual exchange rate pegging arrangement. In this arrangement, members of the group agree to limit fluctuations of their exchange rates to within agreed bands around prescribed central parities. They also agree to coordinate economic policies to react collectively when the exchange rates near the edges of the bands. The Exchange Rate Mechanism (ERM) of the European Monetary System (EMS) is a good example. The ERM was established in 1979 by 11 of the 12 member countries to eliminate intra-European exchange rate volatility along the lines of the Breton Woods System. As the effective capital market integration increased in Europe, the ERM became increasingly vulnerable to speculative attack in 1992-93, after which the bands were widened. In 1999, the system evolved into Europe's Economic and Monetary Union (EMU) with its current single currency Euro.

The second approach is to create a regional currency union. This is a more ambitious approach because it may involve giving up national currencies and building regional monetary institutions and macroeconomic coordination. The largest currency union is EMU. Other examples include CFA franc zone, the East Caribbean dollar area, and the Common Monetary Area. The CFA franc zone consists of two separate monetary unions of sub-Saharan African countries and the Comoros. The first union includes eight members and the second group consists of six members. Both groups have their own central banks to conduct the common monetary policy for the groups. Each group maintains a separate currency, but these currencies are pegged at the same fixed rate against the French franc (and the euro) with financial support from the French Treasury. The East Caribbean dollar area includes eight members. The East Caribbean Central Bank conducts the common monetary policy. The common currency, the Eastern Caribbean dollar, has been pegged to the US dollar since 1976. The Common Monetary Area includes four southern African Countries: South Africa, Lesotho, Namibia, and Swaziland. The South African rand circulates freely in Lesotho,

Namibia, and Swaziland along with their own currencies.

A third approach is common links to an outside currency or a basket of currencies as the monetary standard for the regional group. This approach avoids the need to create complex intra-regional institutions such as a central bank, but requires very close policy coordination among the members of the group. This may be an option in the longer term for ASEAN and Mercosur¹³. For these groups a currency union does not seem to be feasible at this time because intra-regional trade links, while important, are significantly less than in Europe, and countries in these groups seem to be subject to much greater asymmetry of shocks.

Sample text from the TECCTC

Measures on Further Promoting Standardized Operations and Deepening the Reform in Overseas-listed Companies

Overseas-listed companies (referred hereinafter to Company/ies), a form of modern corporate system, which raise capital from overseas, should meet higher requirements in Companies' operations and higher degree of transparency in information disclosure. Currently, most of the Companies have made headway in adopting new systems transforming operational mechanism. A proportion of such Companies, however, has not yet completed the transformation of operational mechanism, leaving a number of problems in standardizing their operations and internal management. In order to further promote the strict compliance of relevant domestic and overseas laws and regulations on the part of the Companies and the fulfillment of consistent obligations to be undertaken by the Companies to investors, and establish a favorable image of the Companies in the international capital markets, the following measures are now raised regarding standardizing operations and deepening the reform of the Companies:

The Companies should improve corporate management in line with the requirements for a modern enterprise system. The Companies and their holding institutions (referring hereinafter to companies, entities and institutions as the major shareholders of the Companies with legal person qualification) must conduct independent accounting and independently assume responsibilities and risks. The holding institutions shall primarily exercise their authority as shareholders according to legal procedures and by way of general meetings of shareholders. Divisions under the Companies, especially the board, the management, the accounting and marketing sections shall be independent of the holding units. Those that are not independent so far must be separated from the holding institutions by the end of 1999. There is no superior-subordinate relationship between the internal offices of the holding institutions and their counterparts in the Companies, and therefore the former is prohibited from issuing documents or in any other forms to influence the independence of divisions under the Companies.

Holding institutions appoint their representatives as board members by law. Executives from holding institutions that serve concurrently in the Companies as chairman, vice-chairman of the

board of directors or executive director should not exceed two in number, and should serve each of the posts with clearly defined job descriptions, assume all legal responsibilities and rights alike associated with that concurrent post, and ensure sufficient time and necessary professional knowledge to perform duties in the Companies. Executives from holding institutions are not allowed to serve concurrently as general/deputy general manager, chief accountant, marketing director and secretary to the board of directors of the Companies.

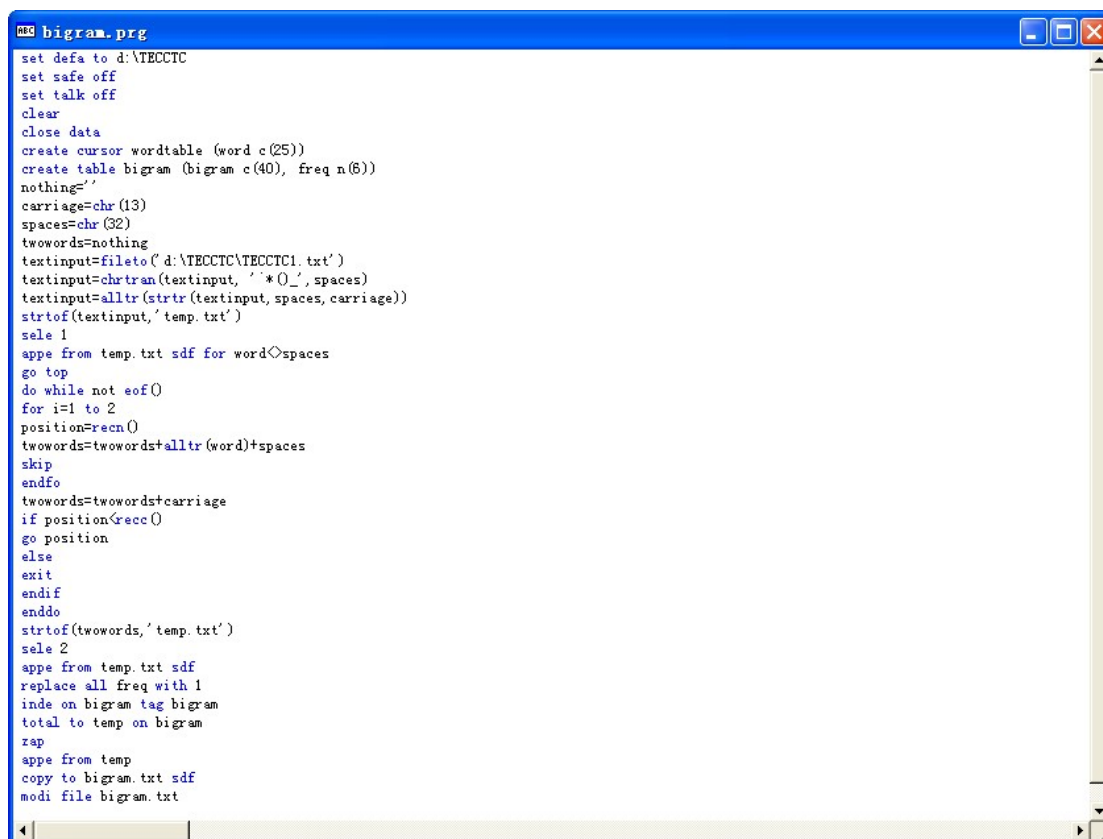
In case major businesses and assets of a state-owned holding institution have already been transferred to the Company, the divisions and corresponding functions of the holding institutions should be assigned and transferred gradually to or merged into another legal-person entity. A holding institution, which possess other assets and businesses rather than the main operations of listed companies should reduce the engagement of connected transactions with the Company and avoid competition in the same trade.

The social functions and non-operating assets of a holding institution should be gradually separated and socialized operation can be fulfilled by way of auction, merger, transfer to local governments, bringing to local insurance system or other means. In case a thorough separation is difficult to be achieved at present, strict management measures should be adopted to ensure a separation in accounting and personnel from the Companies.

In separating social functions and non-operating assets from the Companies, all relevant parties should strictly observe the agreement signed by the Company and its holding institution; where the separation is incomplete, continuous efforts should be made to complete the separation with a time limit. The newly listed Companies should work out specific plans to for the separation from their social functions and non-operating assets. Further solutions and responsibilities in connection with the relevant remaining issues should be clearly defined, otherwise the approval for listing shall not be granted. Local governments and relevant authorities at various levels should take active measures to support the restructuring of the Companies and their holding units.

Appendix C: The *FoxPro* Programme for Retrieving Bigrams

(exemplified with the TECCTC)



```
bigram.prg
set defa to d:\TECCTC
set safe off
set talk off
clear
close data
create cursor wordtable (word c(25))
create table bigram (bigram c(40), freq n(6))
nothing=
carriage=chr(13)
spaces=chr(32)
twowords=nothing
textinput=fileto('d:\TECCTC\TECCTC1.txt')
textinput=chrtran(textinput, '*0_', spaces)
textinput=alltr(strtr(textinput, spaces, carriage))
strtof(textinput, 'temp.txt')
sele 1
appe from temp.txt sdf for word<spaces
go top
do while not eof()
for i=1 to 2
position=recn()
twowords=twowords+alltr(word)+spaces
skip
endfo
twowords=twowords+carriage
if position<recc()
go position
else
exit
endif
enddo
strtof(twowords, 'temp.txt')
sele 2
appe from temp.txt sdf
replace all freq with 1
inde on bigram tag bigram
total to temp on bigram
zap
appe from temp
copy to bigram.txt sdf
modi file bigram.txt
```

Appendix D: Top 30 Most Frequently Used Collocations from the Two Corpora

Top 30 most frequently used collocations from the NECCD							
Rank	W1	W2	F(1)	F(2)	F(1,2)	MI	Log-likelihood
1	dependence	reduce	54	1016	1016	6.45	93.65
2	board	directors	2677	1225	925	5.23	13265.83
3	executive	chief	1617	1500	853	5.05	12524.14
4	estate	real	740	1923	529	12.45	7784.64
5	arbitration	law	4256	5139	310	3.08	2180.96
6	arbitration	clause	4256	644	276	5.91	3139.72
7	supply	chain	1531	581	273	7.52	3763.93
8	respect	with	639	31601	272	3.01	1986.44
9	create	jobs	1400	2187	262	5.68	2830.18
10	cards	credit	507	3497	251	6.4	3025.7
11	development	economic	2358	3664	246	4.09	2081.26
12	consistent	with	439	31601	229	3.3	1797.16
13	applicable	law	789	5139	226	5.06	2256.48
14	place	take	2381	4113	221	3.75	1761.83
15	unemployment	rate	706	5014	220	5.22	2251.3
16	business	entity	9239	577	219	4.62	2075.07
17	minister	prime	483	323	216	9.69	3776.12
18	pay	off	3366	2345	213	4.01	1775.42
19	party	third	2250	1181	210	5.56	2226.92
20	comment	declined	491	424	208	9.22	3429.82
21	mutual	fund	706	2682	208	6.04	2367.96
22	fiscal	year	903	8625	206	3.98	1731.98
23	customer	service	1226	3167	203	4.97	1976.25
24	management	risk	3941	3006	203	3.36	1501.58
25	union	European	641	1296	199	7.16	2595.94
26	deficit	trade	658	4776	195	5.21	1990.5
27	stage	at	445	22223	192	3.54	1543.32
28	depend	on	224	36801	190	3.78	1744.9
29	agency	awarding	1992	235	186	7.89	2755.76
30	law	provisions	5139	1031	186	4.39	1660.81

Top 30 most frequently used collocations from the TECCTC

Rank	W1	W2	F(1)	F(2)	F(1,2)	MI	Log-likelihood
1	venture	joint	1223	2869	974	5.36	13995.44
2	exchange	stock	4115	3030	947	3.36	10235.5
3	intellectual	property	866	2897	853	4.85	13138.54
4	crisis	financial	1354	5893	839	3.01	9916.18
5	renovation	technological	63	824	824	6	66.92
6	bilateral	relations	2320	1708	364	5.76	3997.97
7	manager	general	1278	4438	357	5.21	3650.33
8	fund	custodian	5186	378	352	6.72	4787.11
9	application	documents	4221	1981	349	4.62	3240.84
10	crisis	global	1354	3162	348	5.58	3737.48
11	volume	export	2004	10534	347	3.27	2560.04
12	commercial	administrations	4554	636	345	6.13	4093.32
13	competent	authorities	1776	2716	344	5.39	3585.19
14	gas	natural	1102	1034	339	7.45	4628.07
15	partner	trading	846	3377	338	6.12	3942.15
16	push	forward	573	1401	338	7.95	4953.14
17	enhance	cooperation	1415	9105	332	3.92	2770.55
18	application	materials	4221	4021	329	3.51	2517.67
19	comply	with	338	34628	327	4.04	3264.76
20	policy	monetary	2527	647	327	6.88	4218.11
21	press	conference	877	2055	325	6.73	4072.16
22	prime	minister	330	3228	322	7.47	4793.19
23	speed	up	649	11747	322	4.64	3113.89
24	implementation	measures	2116	7108	311	3.61	2430.53
25	individual	income	1244	4286	306	5.08	3057.16
26	taxation	bureaus	2445	552	305	7.06	4030.24
27	consumption	energy	1764	3088	304	5.04	3002.9
28	government	provincial	7648	1456	304	4.01	2563.68
29	application	form	4221	1924	293	4.41	2623.85
30	development	sound	14024	812	287	3.89	2420.03

Appendix E: Source codes for the *Perl* programme of text chunk segmentation and computation of vocabulary growth

```
##This programme segments all the commercial texts used in this study (exemplified with the
TECCTC) into 2000-word text chunks, then makes wordlists and computes vocabulary growth.
@filename=glob("tecctc\*.txt"); #####
foreach $file(@filename){
open(F,"$file") or die ("Can't open file.\n");
read(F,$text,90000000);
$cumutext=$text;
}
open(R,">tecctcwordlist.txt") or die("Can't creat file\n"); ####
open(W,">tecctcvocgrowth.txt") or die("Can't creat file\n"); ####
use Lemmatizere;  ##lemmatizere removes ord >126 or <48
use Text::Tabs;
$tabstop=30;
mkdir('tecctcwordlist'); #####
$|=1;
$cumutext=~s/[\n\t]/ /g;
$cumutext=~s/["'"]/" /g;
$cumutext=~s/['`']/ /g;
$cumutext=~s/--/ /g;
$cumutext=~s/ - / /g;
$cumutext=~tr/ / /s;
$nonchar=chr 41377;
$cumutext=~tr/$nonchar//s;
$cumutext=~s/^ | $//;
$tokennumber=($cumutext=~tr/ / /s)+1;
$filelength=length($cumutext);
$chunknumber=int $tokennumber/2000;
$remain=$tokennumber%2000;  ##check remaining words and then add each of the words from
$addwordround till the end
$addwordround=$chunknumber-$remain;
for ($i=1;$i<$chunknumber+1;$i++){
if($i%100==0){
system(cls);
print "$i chunks have been produced ...\n";
}
}
$textchunk=$cumutext;
if($i<$addwordround){
## beginning of I LOOP
```

```

$textchunk=~s/((\S+ ){2000}).*/$1/;
} else {
$textchunk=~s/((\S+ ){2001}).*/$1/;  ##when $i equals $addwordround, add one of the remaining
words.
}
push(@chunkarray,$textchunk); ##for randomizing chunks
$chunklength=length($textchunk);
#$cumutext=~s/^\s+//;
$cumutext=substr($cumutext,$chunklength,$filelength);
#$cumutext=~s/$textchunk(.*)/$1/;
} ##end of I LOOP
##The following randomizing text chunks.
$arraylengthb=$arraylength=$#chunkarray;
for($j=1;$j<$arraylength+1;$j++){
$fengwordlist=">tecctwordlist\tecctwordlist".$j.".txt";  #####
open(G,"$fengwordlist") or die("Unable to create file.\n");
if($j%10==0){
system(cls);
print "$j word lists have been produced ...\n";
}
$elementnum=int rand($arraylengthb)+1;
$arraylengthb--;
$textchunk=$chunkarray[$elementnum];
splice(@chunkarray,$elementnum,1);##remove the already selected file name in array
$textchunkb=getshortform($textchunk);
$textchunkb=markforeignword($textchunk);
makewordlist($textchunkb);
getvocgrowth();
}
printcumuwordlist();

#sub getshortform convert short forms into full forms
sub getshortform{
$text=shift();
$text=lc($text);
$text=~s/Isn't/is not/g;
$text=~s/Aren't/are not/g;
$text=~s/Wasn't/was not/g;
$text=~s/Weren't/were not/g;
$text=~s/Don't/do not/g;
$text=~s/Didn't/did not/g;
$text=~s/Doesn't/does not/g;
$text=~s/Haven't/have not/g;
$text=~s/Hasn't/has not/g;
$text=~s/Hadn't/had not/g;
$text=~s/Won't/will not/g;

```

```

$stext=~s/wouldn't/would not/g;
$stext=~s/can't/can not/g;
$stext=~s/couldn't/could not/g;
$stext=~s/there's/there is/g;
$stext=~s/there're/there are/g;
$stext=~s/we're/we are/g;
$stext=~s/they're/they are/g;
$stext=~s/it's/it is/g;
$stext=~s/i've/i have/g;
$stext=~s/we've/we have/g;
$stext=~s/you've/you have/g;
$stext=~s/they've/they have/g;
$stext=~s/shouldn't/should not/g;
$stext=~s/he'll/he will/g;
$stext=~s/they'll/they will/g;
$stext=~s/she'll/she will/g;
$stext=~s/i'd/I would/g;
$stext=~s/he'd/he would/g;
$stext=~s/we'd/we would/g;
$stext=~s/she'd/she would/g;
$stext=~s/that's/that is/g;
return($stext);
}
sub markforeignword{
my($word,$word1,$foreignword,$text);
open(Z,"multiword.txt") or die("File does not exist.\n");
read(Z,$foreignword,7000);
my($text);
$text=shift();
@multiword=split(/\n/,$foreignword);
foreach $word(@multiword){
$word1=$word;
$word1=~s/ / ^/g;
$text=~s/\b$word\b/$word1/gi;
}
return($text);
close(Z);
}
sub makewordlist{
my($text);
$text=shift;
($wordlist,$wordnumber)=lemmatize($text);
#@wordlist=%$wordlist;
#$textvocsiz=int $#wordlist/2+1;
$textvocsiz=keys(%$wordlist);
$subtitle="Text: text chunks".$.j.".txt\nText size: $wordnumber\nVocabulary size: $textvocsiz\n";

```


[illegible]

Appendix F: Source codes for the *Perl* programme of lemmatisation

```
package Lemmatizere;
use Exporter;
@ISA=("Exporter");
@EXPORT=("lemmatize");
sub lemmatize{
my($textinput,$word,$dicinput,$wordform,
$tokennumber,$lemma,@tempwordlist,%dichash,%wordlist,%lemmatemp);
$textinput=shift();
#$textinput=~s/\-\\n//g;
$textinput=~s/[ , . : " ' ! ? ( ) \ ] / /g;
$textinput=~tr/[,._?";'!()><+&%*{}=~\\|\\n\\t\\[\\]\\\\@#\\$\\ / /s;
$textinput=~s/^ | $//g;
@tempwordlist=split(/ /,$textinput);
#foreach $word(@tempwordlist){
#EMPTYSTRING:foreach $word(@tempwordlist){
foreach $word(@tempwordlist){
#$word=~s/^W//g;
#$word=~s/ //g;
$tokennumber++;## if(not $word eq "");
#next EMPTYSTRING if($word=~m^W+ / or $word=~m^b\d+\b / or $word eq "");
if(ord($word)>47 and ord($word)<126){
$word=lc $word;
$word=ucfirst($word);
$tempwordlist{$word}++;
}
}
}
open(LEMMAFILEHANDLE,"lemmadic.txt")or die("File does not exist.\n");
read(LEMMAFILEHANDLE,$dicinput,900000);
%dichash=split(/\\n /,$dicinput);
foreach $wordform(sort(keys(%dichash))){
$wordform=ucfirst($wordform);
if(exists($tempwordlist{$wordform})) {
$lemma=$dichash{$wordform};
$lemmatemp{$lemma}+=$tempwordlist{$wordform};
delete($tempwordlist{$wordform});
if(exists($tempwordlist{$lemma})) {
$lemmatemp{$lemma}+=$tempwordlist{$lemma};
```

```

delete($tempwordlist{$lemma});
}
}
}
%wordlist=(%tempwordlist,%lemmatemp);
$tempwordlist=();
%lemmatemp=();
%dichash=();
#The following statement passes back a referenced wordlist hash and a scalar
#variable containing number of word tokens in text.
return(\%wordlist,$tokennumber);
close(LEMMAFILEHANDLE);
}
1;

```


Appendix G: Source codes for the *Perl* programme of computing

keyword growth

```
##This programme computes the increase of keywords in a corpus (exemplified with the NECCD) as
the number of text chunks increase.
open(A,"key_neccd.txt") or die ("Can't open file.\n"); #####
read(A,$keyword,9000000);
open(B,">kgrowth_neccd.txt") or die("Can't open file.\n"); #####
$keyword=~s/\n/ 1\n/g;
%keyword=split(/\n /,$keyword);

@filename=glob("neccdwordlist\*.txt");#####
foreach $file(@filename){
open(F,"$file") or die ("Can't open file.\n");
read(F,$wordlist,90000);
$wordlist=~s/[s\S]+?-+?\n//;
$wordlist=~tr/ / /s;
%wordlist=split(/\n /,$wordlist);

foreach $word(keys %wordlist){
$word=lc $word;
if (exists($keyword{$word})) {
$keywordgrowth++;
delete($keyword{$word});
}
}
$tokennum+=2000;
system(cls);
print "Number of files processed: $tokennum...\n";
$growthdata="$tokennum $keywordgrowth\n";
push(@growthdata,$growthdata);
}
foreach $data(@growthdata){
print B $data;
}
```