FINDING THE FORMULA:
FORMULAIC LANGUAGE USE IN HONG KONG PRIMARY SCHOOL ENGLISH TEXTBOOKS

by

THURSTAN STEVEN RUSSELL

A thesis submitted to

The University of Birmingham

for the degree of

MASTER OF ARTS BY RESEARCH

Department of English Language and Applied Linguistics

School of English, Drama and American & Canadian Studies

College of Arts and Law

The University of Birmingham

September 2017

ABSTRACT

This thesis examines the use of multi-word lexical phrases, sometimes referred to as formulaic language, in textbooks used in Hong Kong primary schools. To identify whether such phrases are as equally represented in Hong Kong English textbooks as they are in a textbook used by native English speakers in the United Kingdom, formulaic language is identified, counted and compared between texts. Identification is based on a set of pre-determined criteria, and frequency of the formulaic sequences is established using corpus linguistic methodology. An identification criteria value is given to all identified formulaic sequences based on how many criteria are used to identify them as formulaic. The suitability and usefulness of the formulaic sequences are discussed in relation to teaching and learning English as a foreign language. Results suggest that Hong Kong English textbooks contain less formulaic sequences than the textbook used in the United Kingdom and the identification criteria values are lower. Furthermore, qualitative analysis of formulaic sequences found in the Hong Kong English textbooks reveal a disparity in the structure and use of some sequences compared to sequences found in corpora of naturally produced English, raising the question of what constitutes native-like English.

ACKNOWLEDGEMENTS

I would like to express my thanks and appreciation to the College of Arts and Law, University of Birmingham, for awarding me a Masters Scholarship to undertake this research.

Thanks also to my dedicated supervisors, Dr Gareth Carrol and Professor Jeannette Littlemore, who have given me invaluable guidance and inspiration at every step of the process, and to the academic staff in the Department of English Language and Applied Linguistics, who are always willing to offer their help and support.

Finally, to my wonderful family, whose continuous encouragement and reassurance is an endless source of motivation.

CONTENTS

LIST OF APPENDICES

LIST OF TABLES AND FIGURES

ABBREVIATIONS

EFL	English as a foreign language
L1	First language
L2	Second language
HK	Hong Kong
UK	United Kingdom
BNC	British National Corpus
MI	Mutual information
COBUILD	Collins Birmingham University International Language Database
EDB	Education Bureau of the Government of the Hong Kong Special Administrative Region
COCA	Corpus of Contemporary American English
NOW	News On the Web Corpus

CHAPTER 1

INTRODUCTION

1.1 Background to the Thesis

Teacher: At the end of the day, it is important that you use English as often as possible.
Student: Why is the end of the day better than other times?

This thesis is motivated by my experience as a teacher of English as a foreign language (EFL). In classes, I found that misunderstandings in communication with my students often arose around the use of phrases, rather than individual words. Sometimes this was the students' use of phrases, which could be confusing due to direct translations from their first language (L1), but very often it was my own use of phrases, that were misunderstood by the students due to their idiomatic or colloquial meanings. Explanations of individual word definitions were often insufficient to explain the misunderstandings and it was apparent that strings of words and their meanings were often connected, rather than operating individually. Although vocabulary is traditionally thought of and described in singular lexical units, such as dictionaries listing words individually, it often includes multi-word lexical units (Schmitt and Carter, 1997). In the above example, the meaning of the phrase 'at the end of the day', can be taken for granted by native speakers of British English to introduce a fact after everything else has been considered. However, to learners of English who are unfamiliar with the phrase, it could easily be taken literally as meaning the point in time at which the day ends. These phrases are often fixed, and using them in the wrong situation or in the wrong order can sound unusual and hinder smooth communication. Binomials are a good example of words that often go together. In the example of *fish and chips,* all three words can be used individually, but when used together in British English, they create a very specific meaning of the traditional British battered dish. However, if the phrase were to be used in the alternative sequence of *chips and fish,* it would become odd

sounding and could even affect how the listener perceives the meaning of the phrase. Academic studies in linguistics have helped to explain and contextualise the importance of such phrases in language. Knowledge of multi-word vocabulary, such as idioms and other fixed expressions, is necessary for 'speaking a language with any degree of fluency' (Gibbs, 2008, p.697). Sinclair (1991) described the importance of word collocation, that is, the regular and natural occurrence of certain words with others. He explained that language users have a large choice of semi-preconstructed phrases within their lexicon which can be selected for use in the same way that a singular unit can be, and he described this as 'the idiom principle'. Identification of such patterns in language was made easier by the development and growth of the field of corpus linguistics. This area of linguistic research compiles and observes patterns in computerised databases of naturally occurring language, and research using corpus linguistic methods have added to the now extensive literature on phraseology. Important research to come out of this was the development of *Pattern Grammar* (Hunston and Francis, 1999), which detailed how words are used in a series of characteristic and recurrent grammatical patterns. Identifying language patterning has had implications for foreign language learning and teaching with the development of educational materials based on phraseological collocation (Willis, 1990; Lewis, 1993; O'Keeffe, McCarthy and Carter, 2007).

Language patterns have subsequently been described in the context of formulaic language, focusing on memory, storage and retrieval (Wray, 2002). Such sequences are estimated to account for up to half of naturally produced language (Erman and Warren, 2000; Oppenheim, 2000; Foster, 2001) and are linguistically challenging to non-native speakers, presenting an obstacle to native-like proficiency (Carrol and Conklin, 2014). Formulaic sequences are prevalent in language as a means for participants to socially relate to one another in conversation and also as a way of reducing linguistic misunderstanding (Wray, 2012) which is made more likely by novel constructions. The use of formulaic phrases in language has

significant implications for learners of EFL. A low level comprehension and production of such phrases could result in social awkwardness and misunderstanding when communicating with native speakers. Studies have endeavoured to calculate how much naturally produced language is made up of prefabricated sequences and how much of it is singular and novel. A problem that researchers have faced has been the identification of formulaic sequences. What constitutes formulaic language is paramount to identification and as a consequence, results from studies have been highly inconsistent, ranging from 15% (Rayson, 2008) to more than 50% (Erman and Warren, 2000). In spite of such inconsistencies, and even without having a concrete idea of exactly how much naturally produced language is made up of such prefabricated phrases, it is clear that formulaic language is very important for learning, storing and producing language, both for native speakers and non-native speakers.

If formulaic sequences are as prevalent in naturally produced language as some studies suggest, then it seems that learning such sequences should be prioritised in foreign language learning in order to achieve a natural command of the target language. Greater exposure to commonly used formulaic sequences could enhance a learner's ability to process and produce the desired language and aid communication. If formulaic sequences could enable more efficient learning generally in EFL, then use in primary school learning materials may increase that efficiency even further. Although controversial, the 'critical period hypothesis' (Lenneburg, 1967) states that the optimum age for learning language is between two and early teens. Studies have produced inconsistent results to either confirm or refute this hypothesis (Flege et al., 1999; Singleton, 2005), but children attending school daily are usually better placed for learning than adults (Marinova-Todd et al., 2000). Focusing on formulaic language at the primary school stage may help to avoid some of the obstacles that learners can later face with word selection, resulting in a greater efficiency in the use of the target language.

1.2 Thesis Aims and Questions

The central aim of this study is to examine formulaic language use in EFL primary school textbooks. Hong Kong (HK) primary school English textbooks will be used in the study due to the importance of English in HK and the fact that, perhaps as a result of the historical connection of HK to the United Kingdom (UK), English is taught in all HK primary schools.

The focus of the study will centre on two specific research questions:

1. Are formulaic sequences represented to the same extent in Hong Kong primary school English textbooks as in a textbook intended for native English speakers in the United Kingdom?

2. Are the formulaic sequences used in Hong Kong primary school English textbooks appropriate and useful, based on collocational strength?

In establishing how much formulaic language is used in HK English textbooks, a clear working definition of formulaic language needs to be developed, and so an important component of this research aims to discover what exactly counts as a formulaic sequence. Various aspects of formulaic language will be discussed and previous research will be reviewed in the process of forming a definition. With a working definition established, data analysis will be carried out on various texts taken from the HK English textbooks in order to calculate proportions of formulaic language within the texts. When the proportions have been calculated, the prevalence, or otherwise, will be discussed in relation to levels of formulaic language found in a primary school textbook used by native English speakers in the UK.

The British National Corpus (BNC) (BNC Consortium, 2007) will mainly be used to establish the collocational strength of the words within the identified formulaic sequences through a statistical measure – Mutual Information (MI) score. Analysing the collocation strength of the

sequences should help to give an understanding of their relevance and usefulness to learners of English.

1.3 Overview of the Thesis

Following on from this introductory chapter, Chapter 2 will examine the historical context of formulaic language and how related studies of phraseology, corpora and language patterning have helped to position formulaic language as an important and growing subfield of linguistics and psycholinguistics. It will examine the various approaches that can be taken to formulaic language research and discuss existing research that has sought to identify and calculate formulaic language in both spoken and written language, as well as studies exploring how formulaic language could be applied to EFL learning. English use in HK society and education will be discussed, as will the role of English as a global language. Consideration will be given to what constitutes native-like ability and use of English and the degree to which that is important or not. Chapter 3 will detail the methodology used in this thesis to determine the extent to which formulaic language can be found in the HK primary school EFL textbooks chosen for this study, as well as in texts taken from a primary school textbook used in the UK and intended for use by native speakers of English. With the aid of previous research, a working definition of formulaic language will be established for use in this study through a list of identification criteria, and the method of counting will be explained. Selection of the textbooks to be analysed will be justified with an overview of the relationship between the HK Government, textbook publishers and schools. The collection and processing of the data will be summarised. In Chapter 4, a detailed description of these results will be shown and discussed through both quantitative and qualitative analysis. Results will show how formulaic sequences were identified, and a comparison will be made between formulaic sequences found both in

the HK primary school EFL textbooks and the textbook used in the UK. In addition, analysis of specific formulaic sequences found in the HK texts will be carried out using corpora and the results will be examined and discussed in relation to their suitability, or otherwise, in EFL learning. Finally, Chapter 5 will conclude the thesis by discussing the implications of these results to HK EFL learning specifically and EFL learning in general, as well as the direction that future research in this area can take.

CHAPTER 2

LITERATURE REVIEW

This chapter will set out the background to this study. It will discuss previous research and literature on formulaic language, as well as linguistic theories, studies and literature that preceded and led to the theoretical concept of formulaic language as it is known today. The understanding of the role that formulaic language plays in language generally, and language acquisition specifically, has wide reaching implications for language teaching and learning. Its relevance is applicable to both first language (L1) and second language (L2) teaching and learning across languages, but for this study, emphasis will be placed on formulaic language use in the English language, especially in relation to the teaching and learning of English as a foreign language (EFL). The widespread use of English as a global lingua franca, as well as region specific variations in English, make the study of formulaic language in English particularly interesting. Although EFL teaching and learning is prevalent throughout the world, and the aims and implications of this project could be applied to EFL in any context, data collection for this study focused on English textbooks used in Hong Kong (HK) primary schools. The reasons for choosing HK as a case study will be discussed later in this chapter, in addition to an overview of the use of English in HK education and society, and its historical significance to the region.

2.1 Formulaic language in context

Formulaic language research is a vast and continuously evolving area of linguistics. It encompasses a wide range of studies that have been approached from various perspectives and linking those studies coherently is a difficult and complicated task. It has been noted that formulaic language research activity is ever increasing, with greater interest in the topic being

evident from the many conferences now held around this theme (Wray, 2012). The explosion of research, which often utilizes computerized corpora (Ellis et al., 2008; Paquot and Granger, 2012; Garnier and Schmitt, 2015) and psycholinguistic techniques such as eye-tracking (Siyanova-Chanturia et al., 2011; Carrol and Conklin, 2015; Carrol et al., 2016) suggests that formulaic language research has become a popular and well-established field of linguistics.

2.1.1 Historical beginnings

While the potential implications of increased formulaic language research offer exciting possibilities, recognition of fixed phrases in language is not a recent phenomenon at all. Wray (2002, pp.7-8) cites observations of fixed phrase use in cases of aphasia as long ago as the mid-nineteenth century. An awareness of the importance of fixed phrases existed outside of medicine and academia too. The writer, George Orwell took exception to the use of what he described as 'worn-out metaphors…merely used because they save people the trouble of inventing phrases for themselves' (Orwell, 1946). He famously examined how fixed phrases could ultimately come to convey people's thoughts in the fictional 1949 book 'Nineteen Eighty-Four' (Wray, 2009, p.44).

Formulaic phrases have been used in the political arena too, as an attempt to control thought in historical dictatorships, totalitarian societies and Marxist regimes. Lenin was quoted to have said 'people for the most part…don't know how to think, they only learn words by heart' (cited in Ji, 2004, p.83). In Communist China, Mao Tse Tung, implemented a policy of 'linguistic engineering', reflecting the idea that language is so closely tied to thought, that controlling one would ultimately control the other. Formulaic phrases such as 'our Party is a great Party, a glorious Party, a correct Party' were taught in schools not only in Chinese, but also in English (Ji, 2004, p.87). The use of formulaic phrases in politics can still readily be seen in examples

such as ‘strong and stable government’ a phrase which was used extensively in the 2017 Conservative election campaign.

2.1.2 Influences on formulaic language research

Lexicographers and phraseological researchers have understood for a long time that single words are not always the appropriate unit of lexical description (Gibbs, 2007, p.698). In 1957, the linguist, John Firth, famously said ‘you shall know a word by the company it keeps’, recognising at that time that ‘the placing of a text as a constituent in a context of situation contributes to the statement of meaning’ (Firth, 1968, p.179). In 1954, Hornby published ‘A Guide to Patterns and Usage in English’, recognising that ‘a knowledge of how to put words together is as important as, perhaps more important than, a knowledge of their meanings’ (Hornby, 1954, p.v). In relatively more recent times, a seminal paper on fixed phrases came from Pawley and Syder (1983). They observed that speaking naturally in a language requires more than just an understanding of the grammar of that language. A sentence can be said in many different ways while still being grammatically correct, but only one version may sound natural to native speakers of that language. An example by Pawley and Syder is the sentence ‘I want to marry you’. Although this could be accurately expressed as ‘I wish to be wedded to you’ or ‘I desire you to become married to me’, these alternatives would sound odd and unnatural (Pawley and Syder, 1983, p.196). They argue that native-like fluency of a language is achieved by having knowledge of ‘lexicalised sentence stems’ which are ‘retrieved as a whole’ and are either complete or partial expressions with slots for variations in tense, noun or pronoun (pp.208-210). Sinclair (1991) further developed the idea that lexical choices contain ‘a large number of semi-preconstructed phrases that constitute a single choice’ in what he described as ‘the idiom principle’. However, he felt that this idiomaticity in language could not

account for language use alone and contrasted it with 'the open choice principle', describing some language as being 'the result of a very large number of complex choices' with grammar being the only restriction (Sinclair, 1991, pp.109-110).

Much of Sinclair's work was set within the context of the 'Collins Birmingham University International Language Database' (COBUILD) project, set up in 1981, that aimed 'to build a corpus of current English and to use this to produce a monolingual learner's dictionary' (Moon, 2009, p.4). 'The Collins COBUILD English Language Dictionary' was first published in 1987 and had a big influence on commercial dictionary publishing (Moon, 2009, p.4). The COBUILD project largely utilised corpus linguistics, the assembling of a computer database of naturally produced written and spoken language. These real-life examples of lexical collocation were then used to produce 'The Collins COBUILD English Language Dictionary'. The ground-breaking work undertaken by the COBUILD team and the continuous development of corpus linguistics has had an important bearing on formulaic language research, providing a useful methodology in helping to identify formulaic sequences. The role of collocates was not the only consideration for the COBUILD team. In examples such as the preposition 'of', Sinclair (1991, p.82) noted that the contribution of such a word is in 'its participation in grammatical structure'. Examples like this needed to be classified in a way that represented this 'meeting of lexis and grammar' (Sinclair, 1991, pp.81-98).

A central aim of the COBUILD project, therefore, was to identify patterns in grammar in order to allow language learners to not only understand a word, but also to use it. The association between the grammatical patterns in which a word is placed and the meaning of the pattern was recognised. Explicit grammar codings were developed and applied: a sequence of elements detailing the pattern in which a word is used (Hunston and Francis, 1999, pp.32-33). For example, the verb 'fill' is presented as follows in the more recent 'Collins COBUILD Advanced Learners Dictionary' (2006), with the capital V (verb) representing the verb 'fill':

‘fill’ – V n with n (*e.g. fill a saucepan with water*)

V n (*e.g. filled a flask*)

V with n (*e.g. filled with tears*)

V (*e.g. was filling*)

What was to become known as ‘Pattern Grammar’ defined patterns relating to individual words as ‘all the words and structures which are regularly associated with the word and which contributes to its meaning’ (Hunston and Francis, 1999, p.37). Pattern grammar was further developed in a series of two volumes: ‘Collins COBUILD Grammar Patterns 1: Verbs’ (Francis et al., 1996) and ‘Collins COBUILD Grammar Patterns 2: Nouns and Adjectives’ (Francis et al., 1998). The series was ambitious in its aims ‘to show all the patterns of all the lexical items in the Collins COBUILD English Dictionary, and within each to show all the lexical items that have a pattern’ (Hunston and Francis, 1999, p.35).

Patterns in grammar have also been described in terms of constructions. Constructions are the central aspect of the broader approach of construction grammar, defined as follows:

> ‘Constructions are stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns… Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist.’
>
> (Goldberg, 2003, p.219)

Construction grammar takes the view that the structure of a language pattern influences the meaning of the phrase, separately from the lexical meaning of the words it contains. For example, a usually intransitive verb such as ‘sneeze’ can be seen to cause motion in the sentence ‘Pat sneezed the napkin onto the floor’ by simply having the same pattern of other motion causing constructions such as ‘Pat pushed the napkin onto the floor’ (Goldberg, 2005).

Construction grammar has many similarities to pattern grammar, but takes a more theoretical approach than the data-driven methodology used with pattern grammar (Littlemore, 2009, p.164). Research in grammar patterning and construction clearly overlaps formulaic language research and it is sometimes difficult to draw a line between lexical patterning and grammatical patterning. In order to on focus on analysis of data collected for this study, attention will be given to formulaic language in terms of lexical patterning.

Nattinger and DeCarrico (1992, pp.32-38) pointed out that prefabricated speech exists in a far wider context than researchers had previously thought. They saw what they called 'lexical phrases' as existing on a variable continuum and accepted that there is a great deal of variation to them. They categorized lexical phrases by using four structural criteria based on 'length and grammatical status', 'whether the phrase has a canonical shape or non-canonical shape' (that is, whether it has a typical grammatical structure or not), 'whether the phrase is variable or fixed' and 'whether the phrase is continuous or discontinuous, that is, whether it consists of an unbroken sequence of words or whether it is interrupted by variable lexical fillers' (Nattinger and DeCarrico, 1992, p.38). Variability of seemingly fixed phrases was also discussed by Moon (1998) in a detailed work on fixed expressions and idioms. Using a database of written and spoken corpora, Moon (1998) identified variation in individual words contained in fixed expressions (such as variations in the verb or noun) and 'systematic variations' in the whole structure, as in the following examples:

'stick/stand out like a sore thumb' (individual verb variation)

'a piece/slice of the action' (individual noun variation)

'learn the ropes' and *'show someone the ropes'* (whole structure variation)

(Moon, 1998, pp.124-139).

These examples from corpus analysis show that although fixedness is a feature of formulaic sequences, variation was found in 40% of fixed expressions in the database, '[calling] into question the whole notion of fixedness' (Moon, 1999, p.121) and giving credence to the idea of a scale of variability.

2.1.3 Recent approaches to formulaic language research

Within the last twenty years, Alison Wray has been an influential figure in the field of formulaic language research and its development. In addition to carrying out wide-ranging research into formulaic language use herself, Wray has also endeavored to bring together findings from a variety of formulaic language research (Wray, 2012). An important contribution, from the paper with Michael Perkins, has been to offer the following clear working definition of a formulaic sequence:

> 'a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar'
>
> (Wray and Perkins, 2000, p.1)

Wray and Perkins (2000) also discussed the many functions that formulaic language serves, and the advantages for the language user. One of the main reasons put forward for the natural use of formulaic language is that it offers advantages in processing for the speaker. Using formulaic sequences saves on processing effort and frees the speaker to concentrate on other things, such as conversation evaluation and other activities (Wray, 2002, p.69). The benefit, however, may not only be for the speaker. Using formulaic sequences may also aid comprehension for the hearer. Common discourse markers help to give structure to the conversation, allow the speaker to stay on track and aid the listener in comprehension. It can also give the speaker an identity, either signalling individuality or establishing position as a

group member, relating speaker to hearer. A speaker can also use certain formulaic sequences as a means 'to manipulate the hearer into a desired action or perception' (Wray, 2002, p.93).

Emphasis on the processing aspect of language is highlighted in the definition given by Wray and Perkins (2000), heralding a new focus for formulaic language research, with consideration given to the cognitive processes and advantages involved in the use of formulaic sequences. These studies placed formulaic language research studies comfortably within the field of psycholinguistics and utilised methodologies such as eye-tracking to measure the processing of formulaic sequences in both native and non-native speakers (Carrol and Conklin, 2015; Carrol et al., 2016; Siyanova-Chanturia et al., 2011).

2.2 Formulaic language as a term

As mentioned in the previous section, the conept of formulaic language is not new, it has existed previously under many different linguistic descriptions. Terminology has historically created a sense of confusion in formulaic language research with different descriptions being used to describe the same observations, while at the same time, the same description has been used to describe different structures (Moon, 1998, p.2). Although the origins of formulaic language research can be seen in previous linguistic theories and studies, the concept as a subject in its own right is relatively new and has seen a large increase in interest from researchers in recent years (Wray, 2012, p.232). With increased interest has come an even greater increase in terminology, and understanding exactly what formulaic language is and how it should be defined has become a compulsory first step in any related research. Wray (2002, p.9) identified over fifty different terms used to describe formulaic language. Table 1 (p.15) lists a selection of some of these commonly used terms and shows the extent to which formulaic language research is approached from a wide variety of linguistic perspectives, and as a result,

can appear to be very disjointed. Formulaic language research has been applied particularly in the areas of phraseology, corpus linguistics, psycholinguistics, clinical linguistics and language acquisition.

chunks	formulaic sequences	multiword units
clichés	fossilised forms	phrasemes
collocations	frozen metaphors	preassembled speech
conventionalised forms	gambits	ready-made expressions
fixed expressions	holophrases	semipreconstructed phrases
formulaicity	idioms	set phrases
formulaic speech	lexical phrases	familiar expressions

Table 1- Alternate terms for formulaic language (adapted from Wray, 2002, p.9)

For the purpose of this research, I will refer to the general combined concept of these terms as 'formulaic language', while describing each string as a 'formulaic sequence', based on the Wray and Perkins (2000) theoretical definition.

2.3 Approaches to formulaic language research

Durrant and Mathews-Aydinli (2011) identify three separate approaches towards formulaic language research - 'phraseological', 'frequency-based' and 'psychological'. These three approaches, although sometimes overlapping, differ dramatically and have a significant bearing on reaching a definition of formulaic language (Durrant and Mathews-Aydinli, 2011, p.59). It is perhaps partly because formulaic language research has been approached from these three different angles that research and findings have sometimes differed dramatically.

2.3.1 A phraseological approach

A phraseological approach to formulaic language focusses on the structural, non-compositional and fixed nature of formulaic sequences. Non-compositionality refers to the inability to break down the various parts of a phrase in terms of semantics and function. For some phrases, only knowledge of the whole phrase, not the individual words, can aid comprehension, as the sequence cannot be interpreted literally (Read and Nation, 2004, p.32). Examples of this include idioms such as 'he took me to the cleaners' or 'she has a chip on her shoulder'.

The fixedness of a phrase describes the extent to which some of its lexical or grammatical components can be exchanged. Moon (1998), taking evidence from corpus studies, found that fixedness was complicated and more variable than might be expected. Non-compositionality and fixedness vary depending on the phrase, with some phrases being more non-compositional and fixed than others. As well as being non-compositional, the same sequence can also be compositional, such as 'kick the bucket', which can have a literal or figurative meaning depending on the context (Cacciari, 2014). 'Kick the bucket' is an often-cited example of a phrase on the non-compositional and fixed end of the scale, allowing limited variation and lacking semantic transparency if previous knowledge of the phrase is unknown. Cutler (1982) suggested that the amount of time an idiom has been present in the language reflects the compositionality, so that an older idiom will be more syntactically frozen. However, there is some evidence to suggest that multi-word expressions are more flexible than expected, sometimes behaving 'as cohesive units that cannot be internally modified' (Cacciari, 2014, p.268).

The study of phraseology itself is also not fixed, and what phraseology encompasses can depend on different authors' definitions (Gries, 2008, p.4). A phraseological approach often relies on native speaker intuition to determine whether a phrase is formulaic or not. However,

care must be taken when using such an approach as intuition can be highly subjective. Two researchers analysing the same dataset by intuition alone may end up with very different results. Taking a phraseological approach to formulaic language can sometimes inevitably overlap with one or both of the other approaches.

2.3.2 A frequency-based approach

The study of phraseology was bolstered by the development of large text corpora (Sinclair, 1991; Moon, 1998; Hunston and Francis, 1999). The evolution of corpus linguistics has allowed researchers to confirm their intuitive hunches with extensive real language data. Two approaches can be applied to a frequency-based approach – 'corpus-based' and 'corpus-driven'. The corpus-based approach describes a methodology that uses corpora to test out existing linguistic ideas and theories. With this method of research, a predetermined research question can be examined using corpora to test the hypothesis, with the corpus acting as a methodological tool. However, a corpus-driven approach allows researchers to extract data from corpora to inform the research. In this case, the corpus is more than a database to back up existing ideas, the research and conclusions directly reflect data that is extracted from the corpus (Tognini-Bonelli, 2001).

Nearly all corpus-driven research includes a description of frequency, and findings are often reported in such terms. The tendency for words to collocate can be measured in terms of frequency by statistical measures such as 'Mutual Information'(MI), 'MI3', 'Z-score', 'T-score', 'Log-likelihood' and 'Dice coefficient' (Hoffmann et al., 2008, p.145). However, the frequency-based approach to formulaic language can sometimes be problematic. While frequency-based evidence can show that words often collocate, that does not necessarily prove that those sequences are formulaic. Similarly, just because a sequence may have low frequency

or even no frequency in corpora does not mean that it lacks formulaicity (Wray, 2002, p.31). For example, the idiom 'kick the bucket' is clearly a formulaic sequence, if non-compositionality and fixedness are the criteria for identification. However, in the BNC, it has a frequency of only 0.07 instances per million words, with only 7 examples found in just two different texts. All of the examples recorded are in the context of educational texts explaining what the phrase 'kick the bucket' actually means. The reliability of frequency as an indicator of a formulaic sequence is also dependent on the type of corpus being used. The source of data and the number of samples may be too limited to determine formulaicity (Hunston, 2002, p.72; Wood, 2010, p.110).

2.3.3 A psychological approach

Formulaic language research has not only existed in the domain of linguistic studies. Psychological and medical research going back to the 1860s has shown that the brain's left hemisphere is related to spoken language (Wood, 2015, p.6). Studies on stroke and brain injury patients suffering from aphasia (an impairment to language caused by injury or damage to the brain) have been carried out from a formulaic language perspective (Lum and Ellis, 1994; Stahl and Van Lancker Sidtis, 2015; Van Lancker Sidtis and Postman, 2006). Indeed, formulaic language has been described as a feature of aphasia for far longer than it has been observed in linguistics (Wray, 2002, p.217). Research has suggested that patients with aphasia can process certain formulaic phrases better than novel constructions and the rhythm that can be associated with some sequences (especially songs or rhymes) is likely to have an impact on retention of those phrases (Van Lancker-Sidtis and Rallon, 2004, pp.209-210). These studies have provided a large amount of evidence to propose that formulaic sequences are stored and accessed in the same way as individual words (Wray, 2002, pp. 236-237).

This interdisciplinary approach to formulaic language research has directed linguistic researchers to take a psycholinguistic approach to their studies, with focus on a dual cognitive processing system where formulaic phrases are stored, accessed and retrieved from the long-term memory, while novel phrases are constructed using 'grammatical rules to select and arrange lexical items into strings' (Van Lancker Sidtis, 2012). Studies have also used experimental methods, such as self-paced reading tasks (Millar, 2011) and eye-tracking (Carrol and Conklin, 2015; Carrol et al., 2016; Siyanova-Chanturia et al., 2011) to better understand language processing. Eye-tracking technology has been described as giving 'researchers a window to the mind' (Conklin and Pellicer-Sánchez, 2016, p.453), allowing the researcher to measure a participant's eye movements while reading. Data can be recorded to show the movement (saccades), stops (fixations) and re-reading (regressions) by the participant on the text. It is assumed that the length of fixation reflects the processing effort; the longer the fixation, the more processing required. A second assumption is that what the reader fixates upon is what is being considered (Conklin and Pellicer-Sánchez, 2016, p.454). Eye-tracking studies are able to show quantifiable differences in processing times when reading formulaic sequences and novel expressions. For example, a study by Siyanova-Chanturia et al. (2011) measured eye movements of native speakers and non-native speakers when they were reading stories containing figurative idioms, literal idioms and novel expressions. They found that native speakers read idioms faster and did not need to re-read as frequently as they did when reading novel expressions (Siyanova-Chanturia et al., 2011, pp.264-265).

Underwood et al. (2004) first used eye-tracking to measure formulaic language recognition in a study that presented formulaic sequences within a text and measured native and non-native participants' fixations on the component words within the sequence. The component words were then presented in non-formulaic sequences and fixations were compared. The findings

revealed that native speakers fixated less on the words when they were used within a formulaic sequence than when they were not.

The three different approaches to formulaic language research all have certain benefits and drawbacks. The best approach for a researcher may be to consider, and perhaps employ, a combination of the three approaches (Wood, 2015, p.32). Gyllstad and Wolter (2016) utilized all three complementary approaches in their study which used a psycholinguistic visual semantic judgement task to record processing times for free combinations and collocations. The collocations were defined based on the phraseological approach to semantic transparency, and the reaction times were also predicted using corpus values. They concluded that collocational processing studies should include phraseological definitions of semantic transparency 'as a purely frequency-based approach will fail to capture this important variable' (Gyllstad and Wolter, 2016, p.317).

2.4 Identifying formulaic Language

If repetition is customary when emphasizing importance, then it could be said that the three main factors in formulaic language research are identification, identification, identification. In addition to previous studies which have referred to formulaic language with varied terminology, it is also the case that studies have applied different criteria to the identification of formulaic language (Erman and Warren, 2000; Van Lancker-Sidtis and Rallon, 2004; Wray and Namba, 2003). Although this can have the advantage of expanding the scope of formulaic language research, it has also had the consequence of producing wide ranging results in studies that have sought to quantify levels of formulaic language.

It is an important part of the process in formulaic language research to identify which sequences are relevant to the study. Some formulaic sequences are easier to identify than others,

semantically opaque phrases, such as some idioms, metaphors or proverbs, are perhaps the best examples. ‘Kick the bucket’ and ‘spill the beans’ are regularly quoted sequences in formulaic language studies, but their relevance in naturally occurring language is limited. In research relating to teaching and learning EFL, a researcher must decide whether such formulaic sequences are relevant to that study if native speakers hardly ever use them. However, it should be noted that not all idioms, metaphors and proverbs are infrequent or lacking semantic transparency. For example, the idiom ‘at the end of the day’ is an idiom that is frequently used in everyday English with 7.73 instances per million words in the BNC compared to only 0.07 instances per million words for ‘kick the bucket’.

Other examples of phrases that are often labelled as formulaic language, are:

1. Collocations

e.g. catch a cold.

Collocations are words that frequently co-exist with each other, such as binomials like ‘fish and chips’. Corpus linguistic methodology has helped to identify collocations, but their exact definition is still hard to pinpoint and define because of the sheer numbers of words that do regularly collocate. Even if a collocation has a low frequency of only two instances, it could be argued, either through statistical measures such as MI or through native-speaker intuition, that the words collocate.

2. Phrasal Verbs

e.g. look for your keys.

Phrasal verbs are very common in the English language and are a very difficult part of the language for learners to master. Phrasal verbs often have more than one meaning, they are grammatically complex, and they usually have collocational associations with other words (Kurtyka, 2001, p.29).

3. Lexical bundles

e.g. in order to…

Lexical bundles are described as units of function, rather than meaning, 'which serve to characterize particular types of discourse'. They are phrases of three or more words identified through a corpus and are often described in research focusing on academic formulaic language (Wood, 2015, p.45).

4. Lexical phrases

e.g. a _____ ago (a year ago).

Lexical phrases were described by Nattinger and DeCarrico (1992) and as such, the definition seems a little outdated now. However, lexical phrases are still mentioned sometimes in literature. They are described as 'chunks of language of varying length' (Nattinger and DeCarrico, 1992, p.1) being more non-compositional than novel expressions. They can contain slots for novel words to fill.

5. Metaphors

e.g. he's a monster when he's hungry.

Metaphors are phrases which link domains, having a different contextual meaning, which can be understood through comparison to the basic meaning. They differ from idioms in that they are not fixed lexical phrases, but as with idioms, learners of English often find metaphors difficult to comprehend and master (Littlemore et al., 2011).

6. Proverbs

e.g. don't put the cart before the horse.

Proverbs, similarly to idioms and metaphors, are often semantically opaque. Unlike idioms, which fulfil constituent roles in sentences, they act as independent statements, containing advice, instructions or warnings (Wood, 2015, p.47).

The above list is not exhaustive. Other examples of multiword units that could be described as formulaic sequences are rhymes, songs, catchphrases and movie or political quotes. However, these generic labels for phrases that are considered formulaic can cause problems in identification as they can often overlap. An example is the collocation 'catch a cold', which could be seen as partially idiomatic if compositionality is a criterion for idiom. As you don't literally 'catch' a cold, then it is also metaphorical. The example 'can you lend me a hand?' could be defined as either a metaphor, an idiom or a collocation through different interpretation too. Deciding what is to be defined as formulaic language and how that should be done is critical to formulaic language research. Chapter 3 discusses in detail issues relating to identification and the identification criteria that were applied to this research.

2.5 Formulaic language acquisition

Considering the potential advantages of using formulaic language, such as faster processing and relatability to the listener, understanding how formulaic language is acquired by a learner is of considerable importance to linguistic research. The largest amount of research into formulaic sequences has been in 'postchildhood L2 and foreign language learning' (Wray, 2012). However, despite the growing interest in formulaic language, there has been surprisingly little research into its acquisition (Wood, 2015, p.67), especially L1 acquisition.

2.5.1 Formulaic language in first language acquisition

A traditional view of child L1 acquisition is that children begin learning to speak by imitating the words that they hear adults use. These imitations are very accurate but surprisingly become worse as the children go through the process of reorganising in line with the adult system

(Foster-Cohen, 1999, pp.35-36). This imitation, of course, is not based on any knowledge of lexis or syntax and when a child says '/wandæt/ (want that)' while pointing at a toy, they are simply imitating what they have heard. Whether this 'sound' is one or two words is surely not relevant to the child. By viewing child L1 acquisition from the point of view of formulaicity, there can be an awareness that even though the adult may hear the child utter several words, it may only be one unit for the child (Peters, 1983, p.5). This is perhaps a good representation of formulaic language and may have some bearing on understanding how it is first acquired. The child may later come to learn the meaning of the words when used independently, but learning the phrase initially can help the child quickly get what they want. Hakuta (1974) suggested that such imitations play an important part in language acquisition and are not only an incidental phase on the way towards competency. Wray (2002) suggests that as a child first hears and acquires language, the issue of analysing irregular strings may not even apply, with the child only analysing those strings that require it. This 'needs-only analysis' assumes that as little language should be broken down as possible, with strings only being analysed when and if necessary. Following this theory, formulaic sequences could remain unanalysed from the child's earliest experience through to adulthood.

2.5.2 Formulaic language in second language acquisition and EFL

Wong Fillmore (1973) studied Spanish-speaking children learning English and found that the acquisition of prefabricated speech was at the centre of language acquisition. Hakuta (1976) emphasised the importance of formulaic sequences in a study of L2 acquisition especially in relation to the early stages of acquiring language, when learning formulaic sequences could kick start the learning process to a point beyond the lexical and grammatical knowledge. This direct route to language acquisition is very important to the learner from a motivational point

of view (Hakuta, 1976, p.333). These early studies identifying the importance of language patterning have influenced the way in which linguists view the acquisition and processing of language and much formulaic language research has focused on the implications for L2 teaching and learning (Willis, 1990; Lewis, 1993; O'Keeffe, McCarthy and Carter, 2007).

Studies have shown the importance of formulaic language use for a natural and fluent proficiency of English, but evaluating which formulaic chunks should be presented in EFL has been a great challenge. A limited number of studies (Shin and Nation, 2008; Simpson-Vlach and Ellis, 2010; Martinez and Schmitt, 2012) have attempted to compile lists of useful formulaic language but no study has yet focused specifically on how formulaic language could benefit primary school EFL learners. It has been argued that formulaic patterns of vocabulary should be given as much focus as any other forms of vocabulary in the language classroom (Hatami, 2015, p.125), although identification of which formulaic sequences to be taught remains problematic.

Corpus-driven research stemming from the COBUILD project sought to develop a series of educational materials based on corpus research. In addition to the ground-breaking *Collins* 'COBUILD English Language Dictionary', a series of English Course books (Willis, 1988a, 1988b; Willis, 1989) were developed, but they were formatted as self-help books rather than developed as a teaching strategy. The comprehensive series of two volumes: 'Collins COBUILD Grammar Patterns 1: Verbs' (Francis et al., 1996) and 'Collins COBUILD Grammar Patterns 2: Nouns and Adjectives' (Francis et al., 1998) attempted to direct teachers in a logical usage of pattern grammar in the classroom context (Francis et al., 1996, p.xiii), but finding an approach to teaching pattern grammar remains a challenge. Michael Lewis (1993; 2002) described an approach to English language teaching and learning based on the direct targeting of prefabricated lexical phrases. However, although activities were described in the books for teaching and learning lexical phrases, there was no indication of which lexical

phrases should be learnt. With so many formulaic sequences available, the problem remains as to which ones should take priority.

Wray (2012, pp.235-236) believes that native speakers are aware of the value of formulaic language and therefore raises the question why adult learners do not immediately turn to multiword phrases when learning an L2. The answer, Wray believes, may be in the educational culture that compels the learner to seek control over the language. Mature learners, through experience, have a greater understanding of how language works and not content with learning a chunk will actively attempt to break the sequence down. Child learners of an L2 should not be bound by such cultural restrictions, and therefore formulaic language teaching and learning may have more success with a younger age group.

Some types of formulaic language may be more problematic for specific learners of English than others and could therefore warrant more explicit learning. For example, as the French language does not contain phrasal verbs, learning them may present more challenges and the inclusion of phrasal verbs in learning materials for French learners of English may be more relevant than other learners of English. The lack of semantic transparency of formulaic sequences, such as the case with idioms, is one of the main obstacles to understanding, and learners will rely on literal meanings of individual parts of the sequence in the absence of a knowledge of the whole (Kecskes, 2000). Non-compositionality of idioms can result in listening and speaking errors in all learners. The substitution of even a single word in a formulaic sequence can sound odd (Littlemore, 2009, p.172) and hinder communication. Errors may occur accidentally and appear insignificant such as ‘a drop in an Ocean’ instead of ‘a drop in the Ocean’ (Mahmoud, 2002) but ‘malformed’ formulaic language can interfere with language processing (Millar, 2011). Errors can also occur due to partial transference of an idiom from the learner’s L1 as in the case of a Spanish learner using the phrase ‘spread the voice’ instead of ‘spread the news’ (Irujo, 1986, p.287). Misunderstandings as a result of

influence from the L1 can also take place when a similar idiom exists in the learner's L1 but contains a different meaning, as in the case of 'red-faced' meaning embarrassed in English but furious in Arabic (Mahmoud, 2002). This was described by Weinreick (1968) as 'interference'. Interference can also bring benefits to the learner as different languages can share the same idioms with the same or similar semantic meaning, resulting in easier comprehension and production (Irujo, 1986).

2.5.3 Formulaic language teaching strategies

Research into vocabulary teaching and learning has started to investigate the acquisition of multi-word expressions away from the tradition of teaching single words (Pellicer-Sánchez, 2017). In an attempt to establish a formulaic language teaching strategy, several studies have developed materials for the purpose of teaching English as a foreign or second language (L2).

Shin and Nation (2008) compiled a list of the highest frequency collocations from spoken texts in the BNC. Using a set of six criteria, they found a large number of collocations that would be among the most frequent 2000 words, if there were no distinction between single or multi-word units. The criteria that they used were:

1. 'Each pivot word was a word type. That is, the different word forms 'book' and 'books' were treated as different pivot words and investigated separately, rather than treating 'book' and 'books' as one word family'.
2. 'The pivot word had to be a noun, a verb, an adjective, or an adverb. Adverbial particles like 'up' as in 'get up' were treated as pivot words because they were adverbs'.
3. 'All the pivot words had to occur in the most frequent 1,000 content words of English according to the spoken word frequency list by Leech, Rayson, and Wilson (2001)'.

4. ‘Each collocation had to occur at least thirty times in ten-million running words’.
5. ‘Each collocation should not cross an immediate constituent boundary… immediate constituents are components that immediately make up larger parts of a sentence’.
6. ‘Different senses of collocations with the same form were counted separately’.

(Shin and Nation, 2008, pp. 341-343)

The criteria that Shin and Nation (2008) applied took an interesting approach in treating different words forms separately, rather than considering all forms under one lemma. Also by breaking up sequences into immediate constituents, the strength of collocations within a sequence could be separately identified. The list of collocations that was compiled during the study provided a useful list of high frequency collocations that ‘could be usefully taught in an elementary speaking course’ (Shin and nation, 2008).

Another list was created specifically for academic speech and writing (Simpson-Vlach and Ellis, 2010), again using frequent patterns in corpora of written and spoken language. The ‘Academic Formulas List’ used several corpora and MI measures to list the most frequent academic phrases. Martinez and Schmitt (2012) aimed to develop a list that would be a guide for L2 teachers and learners. Their list of 505 phrases was compiled using not only the BNC frequency based measures, but they also highlighted meaning or function as one of the identification criteria, as well as non-compositionality. Unlike the previous two studies, they combined a phraseological methodology with a frequency-based approach. Following on from the ‘Academic Formulas List’ (Simpson-Vlach and Ellis, 2010), Ackermann and Chen (2013) developed the ‘Academic Collocation List’. The list utilised the ‘Pearson International Corpus of Academic English’ comprised exclusively of academic English from five native English-speaking countries. Like Martinez and Schmitt (2012), they did not rely solely on a frequency-based approach. After corpus analysis, other criteria focusing on parts of speech were manually

applied to the data. The data was then reviewed for pedagogical relevance, resulting in a pedagogical list of 2468 items.

More recently, Garnier and Schmitt (2015) developed 'The PHrasal Verb Pedagogical List (PHaVE List)'. This list focused on phrasal verbs, which the researchers claim learners encounter 'in every 150 words of English'. In addition, it is suggested that phrasal verbs are 'highly polysemous' and are therefore of great importance to English language learners (Garnier and Schmitt, 2015). The 150 phrasal verbs chosen for analysis were taken from lists compiled in previous studies on phrasal verbs (Biber et al., 1999; Gardner and Davies, 2007; Liu, 2011). Although the list only contained 150 items, Garnier and Schmitt (2015) claim that these 150 phrasal verbs cover 62.95% of all phrasal verbs in the BNC. A further contribution that this study made was to take into account all of the different meanings of the phrasal verbs, bearing in mind their polysemous nature. All of these studies have claimed that the developed lists can allow teachers and learners a starting point and direction in teaching and learning formulaic language. Conducting formulaic language acquisition studies in a real classroom setting is the ideal environment to carry out research and it is important to apply such lists to actual classroom situations for their effectiveness to be established. This research has helped to move formulaic language in EFL further forward and opened the doors for the application of such lists in real EFL learning situations.

2.6 English as a global language

The research contained within this thesis will focus on the English level and ability specifically in HK, but the relationship of English to all areas in the world is pertinent to any linguistic research looking at English language use and acquisition. Table 2 (p.30) shows which countries the United Kingdom (UK) Government classifies as native English-speaking countries in

relation to citizens not needing to prove knowledge of English for immigration purposes (UK Government, 2017).

Antigua and Barbuda	Dominica	St Kitts and Nevis
Australia	Grenada	St Lucia
The Bahamas	Guyana	St Vincent and the Grenadines
Barbados	Ireland	Trinidad and Tobago
Belize	Jamaica	United Kingdom
Canada	New Zealand	United States of America

Table 2 – Native English-speaking countries (adapted from: UK Government, 2017)

Many countries on this list can often be overlooked in discussions on English speaking countries, with emphasis often placed on only a few, such as the UK, the United States of America and Australia. More alarmingly, there are many countries (like HK) that do not have a majority of native English speakers, but nevertheless have many speakers of English as a second or foreign language.

Kachru (1985) described 'three concentric circles of world Englishes' – the inner, outer and expanding circles. The 'inner circle' contains and describes native English-speaking countries or regions, where English is the majority L1, such as those in Table 2. The 'outer circle', sometimes known as the extended circle, includes countries or regions with political or historical connections to English speaking countries, such as previous colonies of the UK. Areas in the outer circle use English as one of their languages, and English has a high status in the region. HK is an example of a region belonging to the outer circle. The 'expanding circle' recognises that although countries or regions belonging to this circle have no direct association with native English-speaking countries, English is used widely within them as an international language or lingua franca. This expanding circle is ever increasing (Kachru, 1985, pp. 12-13).

Kachru's description may be generalized and continuously changing, but it effectively highlights the complexity of the English language. Thinking of the language as simply British, American or Australian is extremely inadequate in understanding the many dimensions of the English language.

Consideration must be given to the role of English as a global language, and how that co-exists with other languages used in conjunction with it. Attitudes that see phenomena such as using two languages in the same conversation or even in the same sentence (known as code switching) as incorrect or inappropriate behaviour, are changing. In the past, code switching has been viewed negatively, with terms such as 'Singlish' (Singapore English), 'Chinglish' (Chinese English) and 'Spanglish' (Spanish English) used critically (Meyerhoff, 2006, p.122) as it is sometimes thought to be the result of a lack of knowledge in one of the languages, requiring borrowing from another (Myers-Scotton, 1995, pp.47-48). It is still the case that when bilinguals or multilinguals become aware that they are code switching, they can feel the necessity to apologise. Considering the linguistic ability required for multilingual code switching, this is an unjustified response (Holmes, 2008, p.46).

Recent studies have highlighted the benefits of multilingualism in society and in the classroom (known as 'translanguaging') in relation to factors such as 'identity performance, lesson accomplishment, and participant confidence' (Creese and Blackledge, 2010, p.112). Translanguaging differs from code switching by acknowledging the valuable relationship between the two languages that are used, and recognising that a combination of the languages is a natural and normal way for bilinguals and multilinguals to communicate (Garcia and Wei, 2014, p.22-23).

It is clearly important that an L2 is taught and learnt as an addition to, not a replacement of, the learner's language repertoire. The claims that knowledge and use of formulaic language can

result in ‘native-like competency’ (Pawley and Syder, 1983) need to be considered carefully with respect to examining exactly what native-like competency is and whether it is necessary for non-native speakers to possess. In terms of the English language in a global context, and in respect of the number of global Englishes in existence, ‘native-like’ seems to be an inadequate term. However, there also seems to be a conflict between the theoretical issues that, for example, translanguaging research is concerned with and the reality of the global demand for learning English. Countries such as China have seen ‘an explosion in the demand for English’ (Cortazzi and Jin, 1996, p.61) due to the belief that a good command of English will help with education and career. A ‘native-like proficiency’ is often what is demanded by English learners. One of the main aims of this research is to respond to that demand by examining the effectiveness of formulaic language use in EFL textbooks.

2.7 English language use in Hong Kong

As a result of HK’s British colonial past, English was the government’s only official language until 1974 when Chinese was also made an official language. For this historical reason, English remains an official language and is still used in government, law, business and throughout the society, such as on street and road signs (Setter et al., 2010). Due to the prevalence of English in HK, it is taught as an L2 in all HK kindergartens, primary schools and secondary schools, and is even the main medium of instruction in some secondary schools. In spite of this, a recent study (Bacon-Shone et al., 2015) found that nearly 90% of HK people describe Cantonese as their mother tongue and the study further claims that the competency levels of English language in HK are low. HK, then, finds itself in an interesting position of having English as an official language, but with the vast majority of HK Chinese using Cantonese as ‘the dominant language of the home and informal communication with friends and peers’ (Li, 1999, p.70).

Bacon-Shone et al. (2015) therefore propose that spoken and written English should receive more attention in HK schools. The HK Government also appears to take this view as the Education Bureau of the Government of the Hong Kong Special Administrative Region (EDB) states on its website that it seeks to improve student's language skills in English, as well as the two dialects of Chinese which are also used. Cantonese is the traditional Chinese dialect used in HK, but since the return of power to China in 1997, Mandarin, which is the official Chinese dialect used throughout mainland China, has also become more widespread (Setter et al., 2010). It can be seen that English language plays a unique role in HK, and although it has a significant historical and practical function within the society, the English language skills of HK people have been identified as low. Despite the reported low levels of English in HK, the situation has to be taken in the context of expectations. In the past, all HK university students and those in jobs dealing with the public were expected to be bilingual in both Chinese and English. In addition, the perceived decline in English language standards could be accounted for by an emergence of an interlinguistic HK form of English (Joseph, 1997. pp.61-64). The existence or status of an established and recognised HK English is debated in linguistics, but a modern HK English language variety is used in HK by mostly native Cantonese speakers borrowing from the Cantonese language (Cummings and Wolf, 2011). Lexical borrowing may be the result of a speaker falling back on their L1, but often it will occur when the L2 does not contain an appropriate word for the description. L2 speakers may also often borrow single nouns and adopt pronunciation from their L1 (Meyerhoff, 2006, p.44).

2.7.1 English language in Hong Kong education

Due to the status as an official language in HK, English is taught as an L2 in HK schools, along with Mandarin Chinese. It is sometimes even used as the main language of instruction in some schools and most university classes are conducted in English. However, its status is not

reflected in the way it is taught. It is treated as a foreign language, rather than an L2, but due to its British colonial past, emphasis is strictly on British English (Cummings and Wolf, 2011) rather than other varieties such as American, Australian or a local variety of HK English. English in HK education has undergone many changes in the last twenty years. Before 1997, when HK was still a British colony, English was the medium of instruction for nearly all secondary school and university students. However, in reality, classrooms often used Cantonese to introduce English language materials due to the difficulty for students to effectively learn through an L2. As tertiary education increased dramatically from the late 1980s, this situation became increasingly common throughout universities too. This has ultimately led to a growing criticism, especially in business and professions, of a steady decline in English standards (Evans and Green, 2007, p.4). Despite this, the HK Government aims for the society to be bi-literate in Chinese and English and trilingual in Cantonese, English and Mandarin (also known as Putonghua). In doing so, it takes an active role in regulating language learning, introducing language proficiency testing for all teachers of English and Mandarin, unless they possess a relevant language degree (Glenwright, 2005). Through its website, the EDB makes publicly available extensive guidelines on the English language curriculum for pre-primary, primary and secondary education. The following extract taken from the 'English Language Curriculum Guide (Primary 1-6)' (EDB, 2004, p.4) explains the EDB's main aims regarding English language in HK primary education:

'English Language Education aims to provide primary school learners with a wide range of contexts and learning experiences to:
• develop their English Language proficiency;
• enhance their personal and intellectual development; and
• extend their understanding of other cultures through the English medium'

2.7.2 Formulaic language in Hong Kong English language education

In an attempt to reach the English language aims set out by the EDB, the 'English Language Curriculum Guide (Primary 1-6' (EDB, 2004) gives a detailed account of what should be included in the English curriculum. This serves as a guide not only to teachers, but also to textbook publishers. In the guide, reference is made to 'formulaic expressions' and 'structural patterns', (EDB, 2004, pp.46-48) suggesting that the EDB values formulaic language learning in L2 acquisition, but a definition of what is meant by a formulaic expression is not given, leaving interpretation open to publishers of textbooks. Examples of formulaic expressions are listed to aid Key Stage 2 (Primary 4-6 / aged 9-11) students in 'interpersonal communication' skills. According to the curriculum guide, formulaic expressions can be used to:

1. Make and respond to suggestions
e.g. Let's go to Stanley this weekend.
That's a good idea.
I'm sorry I can't.

2. Show agreement or disagreement
e.g. Yes, I agree.
No, I don't think so.

3. Open telephone conversations
e.g. Hello. May I speak to Tony, please?

4. Identify oneself in telephone conversations
e.g. Speaking.
This is Peter.

5. Show concern
e.g. What's wrong?
Take care.

6. Express and respond to good wishes
e.g. Merry Christmas.
Same to you.

7. Begin and end formal letters
e.g. Dear Mr. Lee,
Yours sincerely

Adapted from: EDB (2004, p.46)

Structural patterns are also mentioned in the curriculum guide. In the primary curriculum (P1-P6 / Key Stage 1 and 2) nine structural patterns are recommended to be taught, and it is suggested that 'familiarity with these patterns help learners in primary schools construct sentences by analogy and edit their own writing' (EDB, 2004, p.47).

One of the nine patterns recommended is:

'It/There/This + Verb (Be) + Subject'
e.g. 'It is a monster'
(EDB, 2004, p.47)

This more traditional subject/verb/object example differs from the formulaic nature of pattern grammar (Hunston and Francis, 1999) in that it does not appear to be based on real life examples taken from corpora. Although the structure *it + verb + noun* is a common one (256.15 instances per million words in the BNC), there are no examples of the specific structure 'it is a monster' in the BNC. This raises the question of the usage of authentic language in the practice of teaching and learning formulaic language and patterns. It is interesting that formulaic expressions and structural patterns have been highlighted by the EDB as a useful learning tool, but how do teachers and textbook publishers respond to these limited examples and how do the guidelines translate into actual teaching and learning practice? Grammatically irregular expressions such as 'same to you' and 'speaking' are exactly the type of expressions that may be misunderstood by learners of a language, but how can they be effectively presented as a teaching strategy? Chapter 4 will discuss examples of formulaic expressions and structural patterns that have been explicitly presented in the HK textbooks analysed as part of this research.

CHAPTER 3

METHODOLOGY

The original intention of this project was to take a mostly quantitative analytical approach to the research as the aim was primarily to establish whether there was a below average level of formulaic language used in Hong Kong (HK) English as a foreign language (EFL) primary school textbooks, compared with textbooks intended for native English speakers in United Kingdom (UK) primary schools. To do this effectively, formulaic language needed to be identified in the textbooks and then counted and compared.

3.1 Identifying formulaic language

The main obstacle to formulaic language research lies in the identification of formulaic sequences and the criteria in which it is classified. Sinclair (1991, pp. 109-110) described language as the two distinct and contrasting choices of 'the idiom principle' with language use involving selection of 'semi-preconstructed choices', versus 'the open-choice principle' with the only constraint on language choice being 'grammaticalness'. These contrasting ideas are sometimes described 'as forming a continuum from the most free combinations to the most fixed idioms' (Howarth, 1998, p.35).

When a speaker produces a sentence to describe how he/she feels, there are only a few constructions that can be chosen. Options available are in fact quite limited to constructions such as:

I am
I feel } hot, cold, happy, sad etc.

Does this mean that ‘I am’ is a formulaic sequence? If collocational frequency is the judging criteria, it could be described as formulaic, with ‘I am’ having an MI score of 6.78 in the British National Corpus (BNC), suggesting a strong collocation. The MI score gives an indication of the strength of collocation between two words. The lower the score (closer to zero), the more likely it is that the two words occur together by random chance. Hunston (2002, p.71) suggests that ‘an MI-score of 3 or higher can be taken to be significant’. However, if formulaicity is judged by frequency alone, then meaningful analysis would be hindered by an overabundance of formulaic sequences. The following sequences are examples of phrases that are considered to strongly collocate, based on their MI score in the BNC, but would not be considered formulaic by other criteria such as semantic opacity:

MI score for ‘room’ collocating with ‘tidy’ is 3.53
MI score for ‘meal’ collocating with ‘tasty’ is 9.24
MI score for ‘you’ collocating with ‘do’ is 4.10

This is another example of the difficulties experienced in identifying formulaic sequences. Frequency-based approaches alone may over-identify sequences as formulaic (as above), while at the same time missing sequences that intuition may consider to be formulaic. For example, the phrasal verb ‘pull over’, meaning move to the side of the road and stop, appears intuitively to be a good example of formulaic language due to its lack of semantic transparency. However, in the BNC, the MI score for ‘over’ collocating with ‘pull’ is only 2.33. If 3.0 or above is taken as significant, then it may be said that ‘pull’ and ‘over’ do not strongly collocate. Therefore, identifying formulaic language by native-speaker intuition can be a useful method to employ, but it must be applied with caution. Wray (2002, pp.21-23) identifies five problem areas of using intuition for identification. Firstly, intuition can only be used on small quantities of data. Manually identifying sequences from large datasets would be impractical, unlike computer searches of corpora. Secondly, identification by intuition is a laborious task and inconsistencies

may occur due to tiredness or changes in judgement over time. This may be combatted by utilising multiple researchers, but it can also create a third problem of inconsistency in personal judgements. Fourthly, even if all judges are using the same methods of analysis, boundaries of formulaic sequences can be undefined and there may not be only one answer to seek. Finally, using intuition may favour subjective insights while neglecting knowledge that we may not have. There are clearly advantages and disadvantages to both a frequency-based approach to identification of formulaic language and an approach based on intuition. A combination of both approaches seems to be the soundest solution to addressing disadvantages from each.

3.1.1 Identification in previous studies

Using intuition for identification of formulaic language by a single researcher can be seen as 'dubious from a modern scientific perspective' and would benefit from one or more of the following:

- 'a pre-determined definition of a formulaic sequence'
- 'a second person replicating the identification using the same definition'
- 'a panel of judges agree on identification of a formulaic sequence'

(Read and Nation, 2004, p.29)

In determining the method of identification for this research, in the absence of multiple researchers, a helpful starting point is the established definition of a formulaic sequence offered by Wray and Perkins (2000, p.1):

> 'a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar'

This description helps the researcher to have a working definition of formulaic language, but it should be used 'fairly loosely as a coverall' (Wray, 2002, p.9). A more detailed set of criteria

need to be applied to define formulaic language for analysis purposes. Wray and Namba (2003, p.27) describe this methodology as 'a criterion-based approach', and this will help inform the methodology of this research. Formulaic language includes, among other things: idioms, proverbs, phrasal verbs and binomials. These pre-determined, generic sub-types of formulaic language categories could be used to group formulaic sequences, but this approach is not straightforward and does not help in the identification process. The categories can easily overlap and although some sequences may appear to clearly belong to a category ('kill two birds with one stone' is an idiom), many formulaic sequences, and the categories that they belong to, are not so well defined ('I see' – meaning I understand).

One of the issues in formulaic language research, specifically in the area of identification and counting, is that establishing criteria is largely dependent on the point of view of the researcher and the aims of the research. It is perhaps for this reason that studies which have sought to quantify levels of formulaic sequences in language have yielded such inconsistent results (Biber et al., 1999; Erman and Warren, 2000). However, as long as the same criteria are applied to all the texts in the same study, then meaningful comparison can still take place within the context of that study. Comparisons between separate studies will remain difficult unless a standardised, established set of criteria becomes widely accepted.

Van Lancker-Sidtis and Rallon (2004) identified and counted 'formulaic expressions' in a screenplay and found that nearly 25% of the text was formulaic. The formulaic expressions were identified manually and placed into one of three categories: speech formulas, idioms and proverbs. The definition of speech formula was relatively specific to the context of the text being described as 'highly dependent on conversational context and often [serving to] move the dialog forward' (Van Lancker-Sidtis and Rallon, 2004, p.211). For this research, the definition of speech formula seems too general for identification. It is conceivable that many sequences could be defined as speech formula. In addition, idioms and proverbs are generic

labels which do not offer much insight into the identification of sequences, and are categories that can potentially overlap. It is difficult to separate a sequence such as ‘a leopard cannot change its spots’ from an idiom, proverb or metaphor.

Another study of formulaic language identification and counting is that of Erman and Warren (2000). This study is often cited in the literature in relation to the claim that ‘more than half (around 55 per-cent) of a text will consist of prefabricated language’ (Erman and Warren, 2000, p.50), highlighting the potentially huge importance of formulaicity in language and in language learning.

They defined a ‘prefab’ as:

> ‘a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization.’
>
> (Erman and Warren, 2000, p.31)

The criterion they used for identification mostly relied on that of:

> ‘restricted exchangeability [whereby]…at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity… [It] may also imply the blocking of certain syntactic variability which is normally possible.’
>
> (Erman and Warren, 2000, p.32).

A good example given of this is using the sequence ‘I’m afraid’ to give bad news, as in the sentence ‘I’m afraid I can’t help you this weekend.’ This cannot be exchanged for a synonym such as ‘scared’ or ‘frightened’ while keeping its intended meaning. However, it is possible to see potential complications with the definition of restricted exchangeability. It appears to give considerable latitude in categorising phrases as prefabs. An example given of a prefab is the phrase ‘up here’ which can be categorised as such since it cannot be reversed to ‘here up’ (Erman and Warren, 2000, p.32). It is important, however, to stress the difference between non-

compositionality and basic grammatical restriction. There is a possible risk that this criterion could be applied to many syntactic structures which are not formulaic in nature but are restricted by standard rules of grammar. The string ‘white table’ does not keep its meaning when it is reversed to ‘table white’, but that does not mean it is a formulaic prefab. Also, it should be highlighted that non-compositionality exists on a scale. To paraphrase Erman and Warren’s definition of restricted exchangeability, a word (or more) in a sequence cannot be replaced by a synonym without changing the meaning, function or idiomaticity. The important feature here seems to be a change in idiomaticity, as it could be argued that ‘it’s raining cats and dogs’ could be changed to ‘it’s raining heavily’ without changing the meaning or function, although the idiomaticity would be changed. Defining idiomaticity could also be problematic, however, without specific criteria. In some examples, it is possible to also keep the idiomaticity of a sequence while changing a word, as in the example ‘a piece of the action’ changed to ‘a slice of the action’ (Moon, 1998, p.126). These examples show that even with an established criterion, interpretation can differ, and the use of multiple criteria could be beneficial. It could be hypothesised that the more identification criteria applied to a sequence, the more reliable the identification. The criteria of restricted exchangeability, used with caution, can still be a useful tool in determining formulaicity, but in discussing restricted exchangeability, the idea of fixedness in formulaic language should be considered. Only a small amount of formulaic language is fully fixed (Wray, 2002, p.34). Many formulaic sequences allow for some level of morphological and lexical variation such as changes in verb tense, pluralisation of nouns, and variation in or adding of individual words, as in the following examples:

Tom got really carried away.
Don’t get too carried away.
Stop getting carried away.

When applying criteria to identification of formulaic sequences, it is important to ‘be maximally explicit about which parameter settings are adopted in order to (i) render their

definitions maximally precise and (ii) allow researchers from other frameworks to more easily recognize potential areas of overlap, or indeed conflict' (Gries, 2008, p.3). Some of the most comprehensive criteria for identification of formulaic language have come from studies that have looked at different aspects of formulaic language use. Perhaps the most comprehensive was a study by Wray and Namba (2003) which focused on formulaic language identification in spoken language, identifying formulaic language used by a bilingual child. Initial identification for this was based on native-speaker intuition, which can often be 'treated with suspicion in scientific research' (Wray, 2002, p.20). To provide a framework for their intuition and to help justify their choices, Wray and Namba (2003) developed the following detailed list of eleven criteria:

A) 'By my judgement there is something grammatically unusual about this wordstring.
B) By my judgement, part or all of the wordstring lacks semantic transparency.
C) By my judgement, this wordstring is associated with a specific situation and/or register.
D) By my judgement, the wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves.
E) By my judgement, this precise formulation is the one most commonly used by this speaker/writer when conveying this idea.
F) By my judgement, the speaker/writer has accompanied this wordstring with an action, use of punctuation, or phonological pattern that gives it special status as a unit, and/or is repeating something s/he has just heard or read.
G) By my judgement, the speaker/writer, or someone else, has marked his wordstring grammatically or lexically in a way that gives it special status as a unit.
H) By my judgement, based on direct evidence or my intuition, there is a greater than chance-level probability that the speaker/writer will have encountered this precise formulation before in communication from other people.
I) By my judgement, although this wordstring is novel, it is a clear derivation, deliberate or otherwise, of something that can be demonstrated to be formulaic in its own right.
J) By my judgement, this wordstring is formulaic, but it has been unintentionally applied inappropriately.
K) By my judgement, this wordstring contains linguistic material that is too sophisticated, or not sophisticated enough, to match the speaker's general grammatical and lexical competence.'

(Wray and Namba, 2003, pp.29-32)

These criteria were not intended to be restricting in any way and their purpose was as an aid to help the researchers to justify their intuitive reasoning for identification of formulaic language.

The phrase ‘by my judgement’ is a reminder that, even with detailed criteria, identification is a very subjective process. The criteria were developed specifically for the identification of formulaic sequences in the spoken language of a Japanese-English bilingual child and so most in the list are relevant to formulaic language in speech. Others can be applied generally to formulaic language research and will be used to aid identification of the textbooks in this study. Specifically, the following criteria from Wray and Namba (2003) will be applied:

1. There is something grammatically unusual about the wordstring.

   Although complete grammatical irregularity is rare in formulaic language, this criterion allows for ‘any wordstring that the regular grammar of the language cannot easily generate’. For example, ‘If I were you’ uses a subjunctive that is no longer in general use, but contained within the fixed phrase (Wray and Namba, 2003, p.29). Wood (2010, p.112) also points out that ‘formulaic sequences tend to be syntactically irregular’.

2. Part or all of the wordstring lacks semantic transparency.

   A common trait of formulaic sequences is the lack of semantic transparency when all the parts of the sequence are used together. Idioms such as ‘a red herring’ are a good example of phrases completely lacking semantic transparency, while ‘it’s raining cats and dogs’ contains at least some of the literal meaning. Some phrases such as ‘it’s a small world’, appear to have a clear meaning, but contain an underlying message which may not be easily understood, in this case to express a coincidence (Wray and Namba, 2003, p.29). Other sequences, such as ‘for some time’, may be ‘deceptively transparent’ (Martinez and Schmitt, 2012, pp.308-309), leading a learner to consider the meaning to be a short time, when in fact it is the opposite.

3. The wordstring is associated with a specific situation and/or register.

Some formulaic language has particular associations with specific situations or register. At the beginning of a speech for example, it is common to start with the phrase ‘ladies and gentlemen’ or to wish someone ‘many happy returns’ on their birthday.

4. The wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves.

   In some ways, this criterion overlaps the previous two but additionally applies to phrases that are ‘routinely employed for a specific act’ (Wray and Namba, 2003, p.30). Examples of this might be set phrases used in ceremonies or procedures such as marriage vows or an oath in a court of law.

3.1.2 List of identification criteria for this research

The first general criteria established for this research is that the formulaic sequence must contain more than one word. This may seem like an obvious requirement when considering words often associated with formulaic language such as ‘phrase’, ‘sequence’, and ‘string’, but in the study by Van Lancker-Sidtis and Rallon (2004), single words were included as ‘speech formulas’ if they ‘constituted a full expression’ (Van Lancker-Sidtis and Rallon, 2004, p.212). An example given for this is the word ‘right!’ which in a certain context could be synonymous with the expression ‘oh I understand’. In EFL, however, the focus is on learning or teaching words that naturally occur together, and so only multi-word sequences were included. In addition, to be categorised as formulaic language, the sequences had to fulfil one or more of the following criteria, adapted from the studies previously mentioned:

A) Have restricted exchangeability
B) Be grammatically irregular
C) Lack semantic transparency
D) Be associated with a specific situation and/or register
E) Perform a function other than the meaning of the words themselves

3.2 Counting formulaic language

In determining the method for counting formulaic language within the textbooks, there were two considerations. Should the sequences be counted as one unit, or should all the words within the sequence be counted?

e.g. Once upon a time, there was a little girl who was a goody two shoes.

Assuming the underlined sequences in the above sentence have been identified as formulaic, how should they be counted? The sentence contains 15 words and 9 of those words have been identified as belonging to a formulaic sequence. Therefore, it could be said that 9/15 (60%) of the words are formulaic.

It is also possible to consider each formulaic sequence as a single 'morpheme equivalent unit' (MEU). Wray (2008, p.12) coined the term MEU in an attempt to standardise the theoretical definition of formulaic language and defined it as:

> 'a word or word string, whether complete or including gaps for inserted variable items, that is processed like a morpheme, that is, without recourse to any form-meaning matching of any sub-parts it may have.'
> (Wray, 2008, p.12)

In this way, it could be said that there are 3 units of formulaicity in the sentence above. If the non-formulaic parts of the sentence are counted individually, then the result would show that 3/9 (30%) units are formulaic. Clearly, these results are very different, and it seems that describing only 30% of the above sentence as formulaic is a misleading representation, at least in terms of volume. Therefore, the first method of counting has been adopted for this research. Although there are different options for counting, by using the same method for all textbooks, meaningful comparison can be achieved.

Some formulaic language has lexical flexibility within the sequence.

e.g. give me a break

In the above example, the pronoun 'me' can be replaced with any object pronoun or a noun.

e.g. you never give her a break/why don't you give John a break?

These types of 'slots' do not have complete semantic flexibility (Schmitt and Carter, 2004, p.7). Although there is some flexibility as to which pronoun or noun can be used, there needs to be a pronoun or noun in that position of the sequence for it to keep its idiomatic meaning. Therefore, in this example, even though the pronoun requires an open choice it would still be counted as belonging to the formulaic sequence in this study.

e.g. give me a break (4/4)

However, in some circumstances, a formulaic sequence may have a word included which is not necessary for the idiomatic meaning to remain.

e.g give me a darn break

In this example, 'darn' has been added for emphasis but is not part of the formulaic sequence. Erman and Warren (2000, pp.34-35) described these optional parts of the sequence as 'extensions' and disregarded them in their study, while including open slots that are necessary for the sequence to be complete. For this research, optional extensions, such as 'darn' in the above example, will not be counted as part of the formulaic sequence.

e.g. give me a darn break (4/5)

The primary school textbooks used in this research typically contain, among other things, a combination of questions, pictures, cloze passage exercises (passages containing omitted words to be filled in by the students), tables, charts and complete texts. Successfully compiling an

organised and systematic corpus of textbook data required determining which parts of the textbook to focus on. As a significant portion of the textbooks consisted of complete texts, often included for reading comprehension tasks, it was decided to focus on these texts in the study as the aim is to highlight formulaic sequences in the language presented and not the suitability of questions or exercises presented within the textbooks.

3.3 The Hong Kong textbooks

This study aims to look at formulaic language in primary school EFL textbooks. The age range of primary school students in HK is approximately 6-12 years old, with six different year groups (P1-P6). In deciding which primary level to focus on, it was considered that the last year group of primary school (P6) would be advantageous as it is supposed that the higher level of language presented will be more suitable for analysis. The last year of primary school is also a bridging point to secondary school and stands midway between the first year of primary school (age 6) and the last year of secondary school (age 18). It is hoped, therefore, that although the research focus is specifically on formulaic language use in primary school EFL textbooks, the results and implications may be applicable to both primary and secondary school EFL textbooks.

The Education Bureau of the Government of the Hong Kong Special Administrative Region (EDB) makes available a recommended textbook list on its website (EDB, 2017). This covers all subjects in kindergartens, primary schools and secondary schools. For primary school English language textbooks, the EDB recommends 12 different printed textbooks, from 4 different publishers. In addition, one online resource is also on the recommended list. It advises schools to select textbooks with consideration to students' needs and abilities, cost and suitability of supplementary exercises, and 'in accordance with the related curriculum guides'

and the ‘‘Guiding Principles for Quality Textbooks’’ (EDB, 2016a). The guiding principles, as set out on the EDB website state that:

> ‘Being important sources of reading for students, quality textbooks help develop students’ ability to learn through reading. The amount and quality of the texts to be included therefore deserve greater attention.’
>
> (EDB, 2016b)

To gain insight into which of the EDB recommended English textbooks are most widely used in HK, a survey of HK primary school websites was carried out. The results established that out of the 50 schools surveyed, books were used from three of the four publishers, with only 8% of schools not using textbooks recommended by the EDB.

It was decided that analysis of textbooks from the three publishers, whose books are used by the schools surveyed and are recommended by the EDB, would be carried out to give a representative dataset of HK English textbooks currently used in HK primary schools. In the interest of anonymity, the textbooks will be referred to as:

- Hong Kong textbook 1 (HKTB 1)
- Hong Kong textbook 2 (HKTB 2)
- Hong Kong textbook 3 (HKTB 3)

3.4 English textbooks used in the United Kingdom

In contrast to the EDB’s recommendation of specific textbooks to be used in HK Schools, ‘there has been no tradition of direct State approval of textbooks in England, and currently there are no processes in place’ (Oates, 2014, p.10). According to the Government’s current Minister of State for School Standards, Nick Gibb, the UK has an ‘ideological hostility to the use of textbooks, particularly in primary schools’. He believes that, since the 1970s, textbooks in schools have been replaced with worksheets which has had a detrimental effect on education standards (Gibb, 2014, p.2). Research by Mullis et al. (2012) (cited in Oates, 2014, p.7) found

that, in England, only 10% of teachers used textbooks for teaching Maths, compared to 70% in Singapore and 95% in Finland. For teaching Science, the statistics given by Martin et al. (2011) (cited in Oates, 2014, p.7) were 4% in England, 68% in Singapore and 94% in Finland. It appears that the situation is similar for English textbooks. One primary school teacher revealed that worksheets and free online resources make up the content of the English classes with focus on reading for learning, but there are no recommended English textbooks to be used in UK primary schools. There are, however, English revision textbooks available for the age range 10-11, mainly to be used for 11+ preparation for children who aim to take grammar school entrance examinations. The textbook used in this study is typical of those revision practice textbooks and was selected to stand as a comparison to the three HK textbooks. It was decided that analysis of only one UK textbook would be sufficient as the main focus of the study is to determine the amount and suitability of formulaic language in HK English textbooks. The textbook shall be referred to as the United Kingdom textbook (UKTB).

3.5 The pilot study

After careful consideration of the criteria for identification of formulaic language, the method of counting, what to count and selection of the textbooks, it was decided that an initial pilot study should be carried out to test the suitability of the criteria of identification on actual texts from the textbooks. The pilot study analysed five texts (684 words) from HKTB 1 and three texts (618 words) from the UKTB. Fewer texts were used from the UKTB because they contained more words. Formulaic language was identified in both textbooks using the five different criteria. It was then counted and compared.

3.5.1 Results of the pilot study

The five texts from HKTB 1 showed that an average of 14% of the language consisted of formulaic sequences (Table 3) while the 3 texts from the UKTB contained 18% (Table 4).

Text number (HKTB 1)	Total no. of words within a formulaic sequence	Total number of words within the text	Percentage of formulaic language
Text 1	21	136	15%
Text 2	14	183	8%
Text 3	17	154	11%
Text 4	12	121	10%
Text 5	23	90	26%
All 5 texts	87	684	14%

Table 3 – Formulaic language identified in the HK textbook

Text number (UKTB)	Total no. of words within a formulaic sequence	Total number of words in the text	Percentage of formulaic language
Text 21	29	177	16%
Text 22	28	161	17%
Text 23	52	280	19%
All 3 texts	109	618	18%

Table 4 – Formulaic language identified in the UK textbook

On the surface, these results appear to be quite similar but there are some factors that need to be considered. Firstly, the amount of formulaic sequences in the HK texts appears to be less consistent than the UK texts. The three UK texts range from 16% to 19% while the range of the HK texts is between 8% and 26%. Secondly, when looking at the formulaic sequences found in the HK texts (Table 5, p.52), they are sometimes repeated within texts as that sequence is the explicit focus of instruction. The sequence is artificially repeated to promote learning through repetition. For example, in HKTB 1, text 1, ‘want to/wants to’ is repeated 6 times in the same text. If the percentages are adjusted to allow for this repetition and ‘want to’ is only

counted one time, then the percentage of formulaic language in HKTB 1, Text 1 would be 9% instead of 15%. This issue means that comparison of percentage of formulaic language between the HK texts and the UK texts may not be fair or equal.

Text	Formulaic language identified		
1	want to/wants to (repeated 6 times) grow up lots of		all day spend all day
2	think of (opinion) so that (repeated 2 times) had to		need to at school wants to
3	lots of it's like need to	so that going to brush [his] teeth	not surprisingly brothers and sister
4	wants to set off moved to tears		found out as [Po] would say
5	looking for there are lots of won't believe your eyes.		got into trouble as usual (repeated 2 times) making sacrifices giving up

Table 5 – Formulaic language identified in texts from Hong Kong textbook 1

3.5.2 Modification of identification criteria

An issue that arose with the identification criteria during the pilot study was that criteria D and E appeared to overlap too much. Those sequences that fitted criterion D (be associated with a specific situation and/or register) such as 'your majesty' also fitted criterion E (perform a function other than the meaning of the words themselves). As a result, criterion D and criterion E were conflated to become a modified criterion D. Therefore, the new set of criteria to be applied to the main analysis of this research will be as follows:

A) Have restricted exchangeability
B) Be grammatically irregular
C) Lack semantic transparency
D) Perform a function other than the meaning of the words themselves/associated with a specific situation and/or register

3.5.3 The need for corpus analysis in identification

During the pilot test, the five identification criteria that were applied to the texts helped to justify selection by intuition. However, it was discovered that some sequences that intuitively appeared to be formulaic, such as 'lots of' (HKTB 1, text 1), did not fit any of the five criteria. The only explanation to justify their inclusion in the list of formulaic language was that they appeared to have a strong collocation. The sequences were placed into an unnamed criterion and the collocation was tested using the British National Corpus (BNC). This is achieved by looking at the Mutual Information score (MI score), with a score of 3 or above indicating 'the amount of non-randomness' to be significant (Hunston, 2002, p.71). All of the formulaic sequences, except one, that were put into the unnamed criterion achieved an MI score higher than 3 (Table 6). As a result, a new criterion of 'strong collocation' was introduced, with the requirement that the collocation must produce an MI score of 3 or above in the BNC. As the new criterion was not part of the original set of criteria it will be labelled criterion S. Testing the collocational frequency of the sequences after they have been selected by intuition serves as a verification of the intuition, in cases where no other identification criteria apply, but a limitation of using this method is that sequences may be overlooked in the selection process.

Formulaic sequence	collocates	MI score (BNC)
want to	want (to)	4.50
wants to	wants (to)	4.51
lots of	lots (of)	4.94
need to	need (to)	4.11
brush [his] teeth	brush (teeth)	7.50
as usual	(as) usual	4.98
a bit of	(a) bit / bit (of)	5.12 / 3.32
worry about	worry (about)	7.68
came out	came (out)	4.92
terrified of	terrified (of)	2.85
unable to	unable (to)	5.38
just as	just (as)	3.43
appeared to	appeared (to)	3.72

Table 6 - 'Strong collocation' MI scores (BNC)

It was decided that for the main analysis, all formulaic sequences that are identified using only the criterion of strong collocation would be verified using a corpus, with an MI score of 3 or above taken as the threshold.

Three corpora have been used for measuring collocational MI scores in this research with the BNC being chosen as the main corpus. Although a limitation of the BNC is the relatively small size of the corpus (around 100 million words), data is collected from a wide range of sources making it a well-balanced reference corpus. Approximately 90% of the data is written material taken from genres such as books, newspapers, magazines and school essays. The remaining 10% of data has been taken from spoken material from different contexts. Spoken material is much more difficult to compile in a corpus and this is reflected in the ratio. As the name British National Corpus suggests, the BNC is representative of British English which, depending on the research context, could be seen as an advantage or a limitation. As the language emphasis in Hong Kong is on British English, due to the historical colonial context (Cummings and Wolf, 2011), a British corpus was considered to be the most suitable reference to use. Another advantage of using the BNC is that it is accessed through the BNCweb online interface (BNC Consortium, 2007), which is user-friendly, fast, and powerful (Hoffmann et al., 2008). It offers many functions allowing the researcher to conduct a variety of linguistic analyses if necessary, relating to the source of the data (Hoffmann et al., 2008). The BNC has been used successfully in previous studies identifying formulaic sequences (Gardner and Davies, 2007; Martinez and Schmitt, 2012).

Although British English is favoured in Hong Kong, some of the qualitative analysis conducted for this research examined sequences that were not found in the BNC. In order to investigate possible influence from American English or other Englishes, two additional corpora were used for this part of the research: the Corpus of Contemporary American English (COCA) (Davies, 2008), a large and varied corpus of American English containing more than 520 million words,

taken from a wide range of spoken and written sources, and the News On the Web (NOW) corpus (Davies, 2013) containing an extremely large database of 4.8 billion words. The NOW corpus takes material from web-based sources and is added to by approximately 5-6 million words per day. This provides an up-to-date and relevant corpus reflecting contemporary language taken from 20 different English speaking countries. The wide variety of English represented in the NOW corpus again raises the question of what is meant by native-like English and what is to be taken as the standard.

3.5.4 The need for a qualitative approach

The pilot test showed that identifying and counting formulaic language in the HK texts and the UK texts can yield some interesting results and discussion points, but as the purpose of the texts in the HK textbooks appears to be different from the UK textbook, insofar as formulaic sequences are explicitly presented and repeated, it may be difficult to offer a fair and equal comparison. It was revealed that the formulaic language used in both sets of texts has important differences and that a closer, more qualitative analysis of the individual formulaic sequences and how they are presented will be necessary to fully examine the use of formulaic language in HK English textbooks. In the five HK texts, sequences are presented as teaching points, showing understanding of the importance of formulaic sequences, but clearly the formulaicity of these sequences differs from some sequences in the UK texts. An example of how formulaic sequences are presented in the HK texts is that ‘want to be’ is presented as a key structure. In HKTB 1, text 1, ‘want to’ (not ‘want to be’) was identified through the ‘strong collocation’ criteria as a formulaic sequence. Analysis of the data in the main study will need to adopt a mixed methods approach, examining the texts through qualitative analysis, as well as the quantitative identification and counting.

3.6 Collection and processing of data in the main study

The texts contained in the three HK textbooks and the UK textbook were transcribed and assigned line numbers for easy identification. Formulaic sequences contained in the texts were identified using the criteria established after the pilot study:

A) Have restricted exchangeability
B) Be grammatically irregular
C) Lack semantic transparency
D) Perform a function other than the meaning of the words themselves/associated with a specific situation and/or register

S) Strong collocation

The following conditions were applied in the processing and presentation of the data:

1) Identified formulaic sequences were underlined. In sequences where a lexical item is needed in the sequence but may have variability, the item was counted as part of the sequence but placed in square brackets. For example, in the sequence ‘do your homework’, ‘your’ can be replaced by a different lexical item such as ‘my’, ‘your friend’s’, or ‘John’s’, so the sequence is presented as:

‘do [your] homework’ (3 items)

2) Where grammatical variability, such as tense variation, is possible the variation was not bracketed as in the examples:

‘there is’ (2 items)

‘there are’ (2 items)

3) In sequences that contain additional lexical items which are not necessary for the phrase to be formulaic, the additional item was not underlined and not counted as part of the sequence, such as:

'spend many hours' (2 items)

4) Where contractions, such as 'isn't', are included in the formulaic sequence, the contraction was counted as two words to represent 'is not'.

5) If a compound noun is written as two words (washing machine), it was counted as two words. If it is hyphenated (mother-in-law) or written as one word (teacup) it was counted as one word.

3.6.1 Establishing collocational strength using the BNC

In order to verify those sequences that were identified as strongly collocating (S), the sequences were searched in the British National Corpus (BNC) to establish the MI score. Those with an MI score of 3 or above were deemed to be strong collocations. Corpus analysis was then carried out on all formulaic phrases irrespective of how they were initially identified, so that an MI score was given to all formulaic sequences. Frequency of sequences, given as frequency per million words, was also noted. Frequency is often normalized and described as frequency per million words in order to allow for comparison across a corpus or between corpora. Identifying strong collocation of two-word sequences was achieved by using one of the words as the search word (node) and finding the collocational MI value of the other word based on the likelihood of it appearing next to the node. Searches in corpora can be conducted to include a span of a certain number of words. Sometimes a collocational search may include a span of 3 spaces either side of the node. It was considered that from the point of view of this research, in looking at specific formulaic sequences, the search should be conducted in relation to the sequence. For

example, if a researcher wanted to discover how often people use the phrase ‘white egg’ in relation to ‘brown egg’, and simply searched ‘egg’ as the node, the results in the BNC would show that the MI score for ‘white’ is 4.81 and the MI score for ‘brown’ is 3.23. It could be concluded from this that ‘white eggs’ are discussed more than ‘brown eggs’, but further analysis shows that in 85% of the instances ‘white’ occurs one space to the right of ‘egg’ (as in ‘egg white’) but it only occurs one space to the left (as in ‘white egg’) in 2% of instances. ‘Brown’, on the other hand occurs one space to the left of ‘egg’ in 33% of instances. There seems to be justification for searching within a specific span from the node in formulaic language research as formulaic sequences are often discussed in terms of non-compositionality, and therefore allowing for flexibility of searching three spaces either side of the node may give misleading results.

Another issue with corpus search methods is the issue of whether to search a single specific lexical unit or a lemma. In corpus linguistics, a lemma is the headword of a set of wordforms (such as a PLAY is the headword of play, plays, playing, played) combined with a linguistic tag such as part of speech. If a search is conducted on the lemma ‘play_VERB’, all forms of the verb ‘play’ (play, plays, playing, played) will be included in the search, without including other parts of speech, such as the noun ‘play’ (Hoffmann et al., 2008). Searching a lemma can be very useful to identify and compare differences in use of wordforms. Sometimes formulaic sequences may allow variations in wordforms, such as ‘look/looking/looked over’, but by the very non-compositional nature of formulaic sequences, variation may be rare or impossible. Sinclair (1991, p.8) suggests ‘there is a good case for arguing that each distinct form is potentially a unique lexical unit’. Shin and Nation (2008) treated different word forms separately in their study on high frequency collocations. Searching a lemma may give an incongruent answer to the question being asked. In the sequence ‘button my lip’, if a simple search of the lemma {lip} is conducted in the BNC, the MI score of ‘button’ is 2.67, but if the

specific lexical unit ‘lip’ is searched, the MI score is 4.14. Intuitively, a researcher may claim to know which sequences allow for variation, but a uniform search method should be applied for all examples. For this research, the specific lexical unit was searched, not the lemma. It is feasible that there can be limitations to whichever search method is employed.

3.6.2 Assigning identification criteria (IC) value to all formulaic sequences

Although only using a small dataset, the pilot study highlighted the wide variety of formulaic sequences present. In the five HK texts, 16 formulaic sequences were identified only through the ‘strong collocation’ criteria compared to seven sequences in the UK texts. Although the pilot study was based on a very limited sample, it was interesting to note the criteria through which the different formulaic sequences were identified. To categorise the sequences in terms of how they were identified as being formulaic, each sequence was labelled with the criterion (A-D or S) that was used to identify it. Overall, it was found that many of the sequences fulfilled more than one criterion and in these cases, the sequence was assigned each relevant criterion so that some sequences were identified using only one and others with identified using all five. If a sequence was identified using more than one criterion, then all relevant criteria were used to label the sequence. It was considered that knowing which criterion was used to identify a sequence, and how many criteria were used, may be relevant for further analysis, and so an identification criteria (IC) value was assigned to each sequence. If a sequence was identified using only one criterion, it was assigned an IC value of 1 and if five criteria were used then the assigned IC value was 5. As formulaicity does not necessarily indicate high frequency (Wray, 2002, p.31), those sequences that fulfil one or more identification criteria (A-D) may, or may not, also be strong collocations. Therefore, all the formulaic sequences found in the texts were searched in the BNC to find the MI score and establish whether they could also be identified

through criterion S. Those with an MI score of 3 or above were additionally assigned the label of strong collocation (S).

3.6.3 Extra data collected for qualitative analysis

As well as containing passages of texts, the HK textbooks also contain individual sequences which are presented outside of the passages as teaching points and 'useful expressions'. These sequences are introduced either before or after a text and then repeated throughout the text. In addition, formulaic expressions are listed at the back of two of the HK textbooks (HKTB 2 and 3). These expressions were transcribed and incorporated into the textbook corpus to be used as data for a closer qualitative analysis.

CHAPTER 4

RESULTS AND DISCUSSION

The texts from the three Hong Kong (HK) textbooks and the United Kingdom (UK) textbook (UKTB) were transcribed and compiled in a corpus containing a total of 54 texts amounting to 10,136 words (Table 7). Due to the size, the corpus of texts is not included in this thesis but is available upon request. The HK textbooks contained 42 texts in total amounting to 7539 words. The UKTB contained 12 texts with 2597 words.

Textbook	No. of texts	No. of words
HKTB 1	20	3677
HKTB 2	9	2016
HKTB 3	13	1846
UKTB	12	2597
Total	54	10136

Table 7 - Number of texts and words in textbook corpus

After compiling the corpus of texts from the three HK textbooks and the UKTB, analysis of the data was achieved through a mixed methods approach, using both quantitative and qualitative methods of analysis.

4.1 Quantitative Analysis

The first stage of analysis was quantitative and involved the identification and counting of formulaic sequences in all the texts using the following criteria, outlined in detail in Chapter 3:

A) Have restricted exchangeability
B) Be grammatically irregular
C) Lack semantic transparency
D) Perform a function other than the meaning of the words themselves/associated with a specific situation and/or register

S) Strong collocation

4.1.1 Collocational strength analysis

After formulaic sequences were manually identified in the HK textbooks and the UKTB using the above criteria, those sequences that were only identified using strong collocation were tested for collocational strength. Of the 50 sequences that were initially identified through criterion S alone, only one was found to have an MI score of less than 3 – 'terrified of' (MI score 2.85) (UKTB, Text 47). This sequence was subsequently removed from the list of formulaic sequences in the data. Interestingly, this was the same sequence that was identified during the random sampling in the pilot test. After the strong collocation sequences had been verified, all other formulaic sequences that had been identified through other criteria were also tested for collocational strength to see whether criteria S could be applied. To give an example of how the sequence 'finds out' was tested in the BNC, the word 'finds' was searched as the node and then the MI score of 'out' being a collocation was recorded as 4.38 suggesting a strength of co-occurrence. In this case, the criteria S was then additionally applied to the sequence 'finds out'. Table 8 shows an example of how the sequences were listed and scores recorded. The full table listing all identified formulaic sequences and their MI scores are shown in Appendices 1-4 (pp.104-118). The word without brackets shows the node and the word in brackets shows the collocate. The number of instances that a sequence occurs in a corpus can also be measured in normalized frequencies which are calculated to allow for variations in size of different corpora (Hoffmann et al., 2008, p.71).

Formulaic Sequence	Collocation Node (collocate/s)	MI (BNC)	Frequency per million words
HKTB 1 - Text 7			
one day	(one) day	4.76	42.89
hears about [it]	hears (about)	4.37	0.16
finds out	finds (out)	4.38	1.15
giving up	giving (up)	4.42	4.98

Table 8 - Examples of MI collocation scores

Normalized frequencies can be useful when comparing phrases between different varieties of English in different corpora, or comparing frequency of phrases between spoken and written data. Normalized frequencies of the sequences (per million words) were also recorded during data collection, but just reporting the normalized frequency without comparison to other data cannot reveal much, and so focus was given to reporting MI scores.

Two-word sequences can be easily searched for collocational strength in a corpus, by searching the node and checking for the collocate, but finding collocational scores for multi-word sequences is more problematic. The method used for this stage of the research was to search the MI score for sections of the sequence. In their study on high frequency collocations, Shin and Nation (2008, p.342) analysed sentences in terms of 'immediate constituents', dividing firstly into phrases and progressively down to a minimal two-word collocation. Although breaking down a sequence appears to be counter-intuitive to the chunking quality of formulaic language, it was useful in determining if indeed all parts of the sequence strongly collocated. For example, the sequence 'was nowhere to be found' (UKTB, Text 50) was analysed as in the table below.

Multi-word sequence	MI score (BNC)
(was) nowhere	3.63
nowhere (to)	3.01
nowhere (x be)	3.14
nowhere (x x found)	4.93

Table 9 - Multi-word sequence analysis 1

In the example, each part of the sequence collocates with another part suggesting that although a 'compositional' analysis was applied, the sequence appears to be non-compositional. In contrast, the following analysis of the sequence 'all will become clear' (UKTB, Text 55) (Table 10, p.64) suggests that collocational likelihood of 'become' and 'clear' (MI score 4.9283) is much stronger, and perhaps more fixed, than 'all' and 'become' (MI score -2.0938). From a collocational point of view, it suggests that 'become clear' is more readily identified as a

formulaic sequence than ‘all will become clear’. This is also reflected when comparing the frequency of the phrases, with ‘become clear’ occurring 2.04 instances per million words while ‘all will become clear’ occurs 0.01 instances per million words.

Multi-word sequence	MI score (BNC)
(all x) become	-2.09
Become (clear)	4.93

Table 10 - Multi-word sequence analysis 2

However, from a phraseological perspective, the sequence ‘all will become clear’ lacks a certain amount of semantic transparency as it has the meaning that everything will be revealed. This perhaps highlights the limitation of using only a frequency-based methodology as a means of formulaic language identification.

4.1.2 Identification criteria (IC) value

After identification criteria (IC) values were added to all formulaic sequences based on the criteria used to identify them, they were listed in tables (Appendices 5-8, pp.119-133). Table 11 shows an example of how the sequences were listed, with identification criteria and an IC value representing the number of criteria that were used to identify the them.

Formulaic Sequence	Identification Criteria (IC)	IC Value
HKTB 1 - Text 7		
1. one day	A, C, D, S	4
2. hears about [it]	A, S	2
3. finds out	A, C, S	3
4. giving up	A, C, S	3

Table 11 - Example of identification criteria (IC) value tables

By assigning an IC value to each formulaic sequence, it was possible to easily identify the number of criteria that had been used to identify each sequence. This method of assigning an IC value has not been used before in formulaic language research, and its relevance (if any) needs further exploration. Sequences are often described in terms of non-compositionality or

fixedness and positioned on a cline in relation to those labels. However, in general, researchers have not classified formulaic sequences according to criteria values. A higher IC value does not necessarily mean that a sequence is more formulaic, as it depends on how formulaicity is identified and measured. However, it could be hypothesised that a higher IC value could be an indication that the sequence is more recognisable as a formulaic sequence and could potentially have more relevance to teaching and learning English as a foreign language (EFL). For example, if a sequence has restricted exchangeability, is grammatically irregular, lacks semantic transparency and performs a function other than the meaning of the words, then it could be argued that this sequence would be of greater benefit to a learner than a sequence that fulfils less criteria, as without learning it the meaning would most likely be unclear if the learner encountered it. If it also strongly collocates, then its importance as a sequence to learn would seem even greater. For example, 'want to' from HKTB 1 (Text 1) can only be identified through the criterion of strong collocation. If every sequence is given a value of 1 for each criterion, then 'want to' has a formulaic value of 1 as it only fulfils criterion S. By comparison, the sequence 'keep it in check' from the UKTB (Text 47), fulfils criteria A, B, C and S and therefore scores 4. However, fulfilling more identification criteria may not automatically show that the sequence is more beneficial to a learner as many idioms may score high in criteria A-D due to their non-compositionality, but their frequency in everyday language may be low. To assess the benefits of such values to a learner of English, they would need to be taken in context with frequency measures. This is an area that could potentially be explored in future research.

The average IC value of each sequence in each of the HK textbooks was calculated by dividing the total IC values by the total number of formulaic sequences. The values from the three textbooks were then calculated to determine the average IC value for formulaic sequences in all the HK textbooks together (Table 12, p.66). The average IC value of formulaic sequences in all the HK textbooks was 2.2 compared to an average IC value of 3 in the UKTB (Table 13).

Textbook	Total IC value	Total number of formulaic sequences	Average formulaic value
HKTB 1	304	142	304÷142 = 2.1
HKTB 2	271	114	271÷114 = 2.4
HKTB 3	145	66	145÷66 = 2.2
All HK textbooks	720	322	2.2

Table 12 - Formulaic value of sequences in Hong Kong textbooks

Textbook	Total formulaic value	Total number of formulaic sequences	Average formulaic value
Average from HK textbooks	720	322	2.2
UKTB	385	130	3

Table 13 - Comparison of formulaic value of sequences between HK and UK textbooks

Each formulaic sequence was further categorised based on IC values so that a comparison could be made, for example, of how many sequences had an IC value of 4 in the UK textbook compared to the HK textbooks. The complete analysis is shown in Appendix 9 (pp.134-135) and a summary can be seen in Tables 14 and 15.

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
HKTB 1	37	52	49	4	0
HKTB 2	20	44	38	11	1
HKTB 3	10	38	14	3	1
Total (out of 322 sequences)	67	134	101	18	2
%	21%	42%	31%	6%	1%

Table 14 - IC value categories

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
HK Textbooks	21%	42%	31%	6%	1%
UKTB	6%	25%	39%	27%	3%

Table 15 - IC value category comparison between HK Textbooks and UKTB

In the HK textbooks, 21% of all identified formulaic sequences only had an IC value of 1 compared to just 6% of all identified formulaic sequences in the UKTB. In contrast, 27% of all

formulaic sequences in the UKTB had an IC value of 27%, compared to only 6% in the HK textbooks. The figures show that more than half of the formulaic sequences in the HK textbooks had IC values of 1 and 2, whereas more than half of the sequences in the UKTB had IC values of 3 and 4. Sequences with IC values of 5 were rare in both HK textbooks (1%) and the UKTB (3%). The formulaic sequences that were identified using all five identification criteria, and therefore had an IC value of 5, are shown in Table 16. Intuition may suggest that all of the sequences, with the possible exception of 'fair's fair', which has a low normalized frequency in relation to the other sequences, are useful expressions that could benefit a learner of English. The MI scores, especially for 'yours faithfully' and 'yours sincerely' are high. It is apparent from the examples that they have been identified using criterion D (perform a function other than the meaning of the words themselves/associated with a specific situation and/or register) and it could be argued that this criterion could be limiting to a sequence if the situation that it is associated with is uncommon. For example, 'yours faithfully' and 'yours sincerely' are associated with letter writing, a practice that is increasingly rare. As such, it may be that sequences with an IC value of 4 may be more relevant than those with an IC value of 5.

Formulaic sequence	Textbook (Text)	MI (BNC)	Frequency per million words
once upon a time	HKTB 2 (Text 23)	4.47	1.63
once upon a time	HKTB 3 (Text 34	4.47	1.63
fair's fair	UKTB (Text 46)	5.19	0.11
yours faithfully	UKTB (Text 51)	13.34	1.55
yours sincerely	UKTB (Text 54)	14.27	8.04
get in touch	UKTB (Text 55)	5.93	3.93

Table 16 – Formulaic sequences with IC value 5

Although the dataset is too small to draw solid conclusions, there does seem to be an interesting difference between the IC values in the HK textbooks and the UK textbook as illustrated by the graph (figure 1, p.68). Further research on IC values of formulaic sequences would help to contextualise these findings.

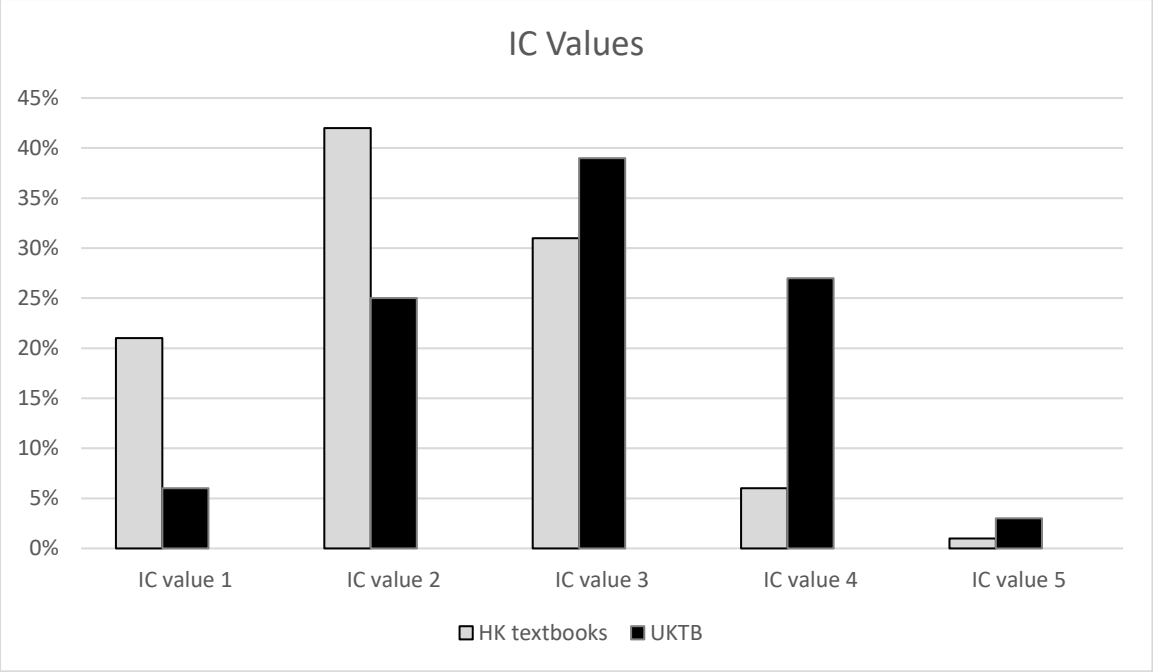

Figure 1 – Comparison between IC value categories

4.1.3 Formulaic sequences present in HK and UK textbooks

One of the key aims of this thesis was to investigate whether formulaic sequences are represented to the same extent in primary school EFL learning materials as learning materials intended for native English speakers, or whether there is a need for greater emphasis on such sequences. After identifying which sequences were formulaic in the textbooks, a percentage of formulaic language present in each of the textbooks could be calculated (Tables 17 and 18, p.69). The formulaic sequences in the UKTB accounted for 15% of the language used. Although this figure is at the low end when compared with previous research (Biber et al., 1999; Erman & Warren, 2000; Foster, 2001), it does fit within the range and compares with a study by Rayson (2008) who also found formulaic language at 15% using Wmatrix corpus analysis. HKTB 2 was at a similar level with 14%, while the other two were relatively lower at 9% each. Considering that many formulaic sequences are repeated in the HK textbooks to

emphasize the teaching point, such as the sequence ‘want(s) to’, which accounts for 6 out of 10 formulaic sequences in HKTB 1 (Text 1), the average figure of formulaic language in the HK textbooks at 10% does seem particularly low.

Textbooks	Total no. of words within a formulaic sequence	Total number of words within the texts	Percentage of formulaic language
HKTB 1	323	3677	9%
HKTB 2	286	2016	14%
HKTB 3	166	1846	9%
Total	775	7539	10%
Average	258	2513	10%

Table 17 - Percentage of formulaic language identified in all Hong Kong textbooks

Textbooks	Total no. of words within a formulaic sequence	Total number of words within the texts	Percentage of formulaic language
Average from HK textbooks	258	2513	10%
UKTB	385	2597	15%

Table 18 - Percentage of formulaic language identified in Hong Kong and UK textbooks

These results do appear to suggest that formulaic language is not as well represented in the HK textbooks as it is in the UKTB. The results also correlate with the higher IC values found in the formulaic sequences in the UKTB. The HK textbooks contain less formulaic sequences, and those that are present, have less IC values.

Further examination of a larger dataset would give a clearer indication of whether the textbooks analysed for this research are typical of HK English textbooks. Incorporation of EFL textbooks used in other countries would be beneficial in establishing whether this is a pattern throughout EFL learning materials in general.

4.2 Qualitative Analysis of the Hong Kong textbooks

Although percentages can give an indication of how much or how little formulaic language is represented in the textbooks, closer analysis of the formulaic sequences can reveal more about the type of formulaic sequences that are being presented both through the texts and outside of the texts.

4.2.1 Authenticity of texts

Although the HK textbooks have been compared with a UK textbook, there are clearly some important differences to note between the two. Firstly, although all of the textbooks are intended for primary students of age 10-11 years old, it would not be unusual to expect the language level to be different between books intended for native speakers and non-native speakers of English. This is evident in the texts that are used in the different books. The UKTB uses authentic texts sourced from literature, extracts taken from well-known literary authors such as C. S. Lewis and Walter de la Mare. As the aim of the texts in the HK textbooks seems to be to highlight teaching points, it appears that the majority of texts are tailor-made and written or compiled by the authors of the textbooks. This has the disadvantage of being both unauthentic and inconsistent between textbooks, relying on the ability of the textbook author. An example of this can be seen in texts 31 and 32 from HKTB 3 where three teaching points of 'would rather … than …', 'prefer … to …' and 'prefer' are deliberately written into the two texts. It appears to lose authenticity and although it intends to show a grammatical pattern, it actually results in an unnatural formulaic structure by being forced into the text. An example from HKTB 3, text 31 is:

'mum and I would rather watch emus running around than feed kangaroos'

By searching this sequence in the BNC, only 14 examples are found. By examining these naturally produced examples, it can be seen that this structure appears to be used to express that the worst thing imaginable (the first action) would be a better option than the second action. Out of the 14 examples in the BNC, half of them have 'die' as the first action.

e.g. 'I would rather die than not vote Labour'

This highlights the danger of presenting formulaic sequences or grammatical patterns in an inauthentic context. The Education Bureau of the Government of the Hong Kong Special Administrative Region (EDB) guidelines state the advantage of using authentic texts in textbooks:

> 'Text types also provide authentic and meaningful contexts within which learners can learn how grammar works for purposeful communication. For example, when learners read a recount of a school function held a year ago, they see how the past tense is used to record past events.'
>
> (EDB, 2004, p.16)

The EDB's recognition of using authentic language is highlighted in specific advice to teachers, such as in the example below where advice is given about authenticity in passive constructions:

> 'Teachers should not plan for the learning and teaching of the passive constructions of all possible patterns. It is recommended that passive constructions of simple language forms be introduced to learners only when there is a genuine communicative reason for using the passive.'
>
> (EDB, 2004, p.48)

Where passive constructions are presented in the HK textbooks, they do appear to be natural, suggesting the EDB guidelines have been effective regarding passive forms. In all cases, the subject of the verb is either unknown or unmentioned as is a standard usage of passive constructions (Berry, 2015, p.185).

e.g. 'many towns and villages are now left with no food' (HKTB 2, text 27)

e.g. 'the first modern Olympic Games were held in 1896' (HKTB 3, text 35)

However, it is difficult for the EDB to micromanage every aspect of language teaching and to identify what constitutes an authentic text. Even if authentic texts are used, there it is problematic to find enough of the target language in texts to allow incidental learning to take place. Despite government recommendations for publishers to use authentic texts, the potential to learn the target language is small unless materials are specifically designed (Webb et al., 2013, p.114). Research has found that words need to be encountered at least 5-8 times before real learning can take place (Pellicer-Sánchez & Schmitt, 2010, p.44) and this has been shown to be the case for multi-word units as well. Webb et al. (2013) examined the effect of repetition of collocation on EFL students in Taiwan by conducting a reading and listening task which exposed the students to different repetitions of collocations. The study found that repeated exposure to collocations significantly increased learning. Pellicer-Sánchez (2017) also conducted a study on incidental learning of collocations through reading on English language learners from a variety of language backgrounds. The collocations were pseudowords of high frequency nouns, from the BNC, and their adjective collocates. Results confirmed that collocations can be learnt through incidental reading when the target language is repeated, emphasising the need for modified texts to allow enough repetition to take place. This highlights the difficulty in finding a balance between maintaining authenticity of a text and providing enough exposure to the target word or sequence.

4.2.2 'Odd-sounding' sequences

The EDB's recognition of and guidance on the importance of authentic use of passives appears to have translated to the textbooks. However, there are many examples of sequences in the textbooks that intuitively sound odd. Text 41, 'A Guitar's tale', recounts the story of a

personified guitar, told in first person from the guitar's perspective. Although creative, it is possible that such a structure could be confusing for the intended age group and ability. In the example, 'now, I am given to a cheerful child' (HKTB 3, Text 41), a passive structure is used, but the tense and storytelling perspective seem unnatural. This intuition is supported by searching the sequence in the BNC. There are 4 occurrences of 'I am given to' and they are all followed by 'understand', creating a formal formulaic sequence as in the example 'I am given to understand that you fully agreed.' It is perhaps as a result of this formulaic sequence that, from a British point of view at least, the sequence in the Hong Kong text intuitively sounds odd. In the larger Corpus of Contemporary American English (COCA), there are 11 occurrences and 6 of those are also 'I am given to understand', with the other examples also taking the formulaic meaning 'given to' - describing the way a person behaves. The following sentences were also selected from the texts as containing intuitively odd sounding sequences and analysed.

a) *'it is deep grey in colour'* (HKTB 2, Text 24).

Listed below are the first ten results after running a search of 'in colour' in the BNC:

blood. They are sometimes greyish ***in colour****, sometimes a dull red or brown.*
although the veining is greenish ***in colour*** *and the paste is very yellow.*
his face was frighteningly lacking ***in colour****, he was trembling, and his lips*
Double Gloucester is rich orange ***in colour*** *and is made from the morning*
his blue eyes, not so different ***in colour*** *from that storm-tossed lake*
As I always like to work ***in colour*** *even when sketching,*
hands high and a perfect mulberry ***in colour****. It was a thick-necked, elegant*
I suppose it's best to do it ***in colour****. We've got some of the*
have a CD that would be ***in colour****. So I did the CD, put it*
tend to be a little darker ***in colour*** *and stronger in flavour.*

The results reveal that the actual name of a colour only precedes 'in colour' 4 times. In two of the cases, the colour has the suffix 'ish' is added to the colour ('greyish' and 'greenish'). In the other two cases, both of the colours have dual meanings ('orange' and 'mulberry'). It is likely that 'in colour' is used in these two cases to emphasize the meaning is colour (not fruit). COCA

confirms the results with only 4 occurrences of a colour preceding ‘in colour’, two of which have the suffix ‘ish’ (‘reddish-brown’ and ‘brownish’), one which has a dual meaning (‘rose’) and the last one is ‘tan-white’. It seems that a colour is not usually followed by the sequence ‘in colour’ in both British and American English, unless the colour needs emphasizing.

b) *‘It (smart card) lets people have a convenient daily life’* (HKTB 2, Text 38).

The BNC has no occurrences of descriptive adjectives preceding ‘daily life’, only possessive adjectives, prepositions and articles. This may account for the odd sounding nature of the sequence in British English, but COCA has 2 occurrences of descriptive adjectives preceding ‘daily life’ (‘ordinary’ and ‘normal’).

c) *‘I plan to do revision after I have dinner’* (HKTB 3, Text 39).

In the BNC, ‘do’ does not collocate with ‘revision’ and in COCA there is only 1 occurrence, which is an imperative. However, ‘plans’ has an MI score of 4.80 in the BNC when collocating with ‘revise’ and ‘plan’ has an MI score of 4.16 in COCA. It seems that a more formulaic sequence would be ‘I plan to do some revision’ or ‘I plan to revise’, and this is a good example where an opportunity to present a strongly collocating sequence in the text has been missed. It is not the only such case in line 3904. ‘After’ has an MI score of 5.29 in the BNC when collocating immediately before ‘dinner’ and 5.21 in COCA. ‘Have’ only has an MI score of 2.33 in the BNC and 2.29 in COCA. More importantly, the BNC and COCA have no occurrences of using ‘I have’ between ‘after’ and ‘dinner’. The formulaic structure of this sentence in both British and American English, based on the corpora, should be ‘I plan to revise after dinner’ not ‘I plan to do revision after I have dinner’. It could be argued that this difference is unimportant and is only meaningful to British and American English speakers, but this example goes to the heart of formulaicity and reveals why a seemingly meaningful sentence could sound odd to native speakers. It may also be the same case in all varieties of English as when the sequence is searched in the News On the Web (NOW) corpus, which continuously

takes data of global Englishes from the internet and currently contains around 4.8 billion words, there is only one occurrence, shown below:

*' I just have to call and let them know, and **after I have dinner**, I always go there and eat'*

d) *'The government has decided to sweep ice-cream men off the streets'* (HKTB 3, Text 43).

In this case, there are no occurrences of 'sweep + off the streets' in either the BNC or COCA. One possible reason for these odd-sounding sequences could be due to interference from the first language (L1) resulting in the sentence structure or writing style being influenced by traditional Chinese, the majority written language in Hong Kong. In trying to determine if this is the case, the sequences were shown to a native Cantonese speaker / traditional Chinese writer who was asked whether they could be directly translated into traditional Chinese. The sequences made perfect sense to the Cantonese speaker/ traditional Chinese writer who was able to directly translate them word for word as in figure 2.

a) It is dark grey in colour.
它 是 深 灰 色 。

b) It lets people have a convenient daily life.
它 讓 人們 有 一個 方便的 每一天 生活 。

c) I plan to do revision after I have dinner.
我 計劃 去 做 溫習, 之後 我 有/ 吃 晚飯 。

d) The government has decided to sweep ice-cream men off the streets.
政府 已經決定 去 掃蕩 雪糕小販 離開 街道

Figure 2 - Odd-sounding sequences directly translated into Chinese

This illustrates that interference from the L1 can be an important factor in the odd-sounding nature of these sequences. Many formulaic sequences can sound odd as a result of interference from the L1 on the second language (L2). Interferences, also known as transfers, between languages are quite frequent and can result in situations such as using the word order from the L1 in the L2, incorrectly inserting determiners, using incorrect prepositions and importing the structure and meaning of a word or expression from the L1 (Grosjean, 2010, pp.70-72).

4.2.3 Pseudo formulaic sequences

As well as the HK textbooks containing examples of odd-sounding sequences, they also contain what this study refers to as pseudo formulaic sequences. This is where sequences are specifically being presented in the textbooks as structures to learn, but are in fact not formulaic in naturally produced language from corpora. For example, HKTB 2 (Text 25) contains the sequence ‘don’t pick the plants’. In the BNC, the MI score for ‘plants’ collocating 2 spaces after ‘pick’ is 0. However, the MI score for ‘flowers’ collocating after ‘pick’ is 5.01. These scores show that it is unnatural to use the verb ‘pick’ with the object ‘plants’ but normal with the object ‘flowers’. These results are confirmed in different corpora. In COCA, the MI score for ‘plants’ collocating after ‘pick’ is 0, while the MI score for ‘flowers’ collocating after ‘pick’ is 3.31. In the NOW corpus, the MI score for ‘plants’ collocating after ‘pick’ is 0 and the MI score for ‘flowers’ collocating after ‘pick’ is 3.29.

HKTB 2 (Text 27) contains the sequence ‘the airport is shut down’. In the BNC, ‘shut down’ does not collocate with ‘airport’ up to 3 spaces. The word ‘closed’, however, has an MI score of 4.31, and as such appears to be a more formulaic sequence to use in this context. By looking at the most frequent uses of ‘shut down’ in the BNC, it appears to be used mostly in a permanent sense of closing, not temporarily due to weather as in the example from Text 27. For example,

‘the workshop had looked certain to shut down after the local authority withdrew its funding’. However, when verifying these results with COCA and the NOW corpus, ‘shut down’ does collocate with ‘airport’ with MI scores of 4.66 (COCA) and 3.67 (NOW). The following examples suggest that ‘shut down’ refers to a permanent action in British English, but perhaps in American English and other world Englishes it can be used for temporary situations:

‘A terminal at the Newark, New Jersey Airport was shut down for more than an hour because of a baby’ (COCA).

‘Chengdu's airport was shut down for more than an hour Monday because of rain’ (NOW).

These results highlight the issue of different Englishes and the difficulty in establishing what native English is and the standard that learners of English should strive towards. A British English speaking EFL teacher may well tell a student that an airport can only be shut down permanently, while an American English speaking EFL teacher may give a different instruction.

A final example of a pseudo formulaic sequence is from HKTB 3 (Text 36) ‘the Beijing 2008 Olympic Games have also drawn much attention from people all over the world’. The BNC corpus reveals that ‘from’ has an MI score of 1.91 for collocating with ‘drawn attention’ but ‘to’ has an MI score of 5.30. The word ‘from’ appears to be used in a different context to ‘to’ when collocating with ‘drawn attention’ as the following example shows:

‘All the hype surrounding Microsoft Corp's Windows NT has drawn attention away from the fact that …’ (BNC).

The use of the word ‘away’ suggests that attention is drawn from one thing to another whereas using ‘to’ only suggests attention on one thing. Perhaps a more natural sequence for the example in the textbook would be ‘the Beijing 2008 Olympic Games have attracted people from all over the world’. MI scores from COCA and the NOW corpus also show that ‘to’

collocates more frequently with ‘drawn attention’ (MI score of 3.83 in COCA and 3.59 in the NOW corpus) than ‘from’ (MI score of 1.79 in COCA and 2.79 in the NOW corpus). All of these examples are emphasised as teaching points in the textbooks. While it is good that collocations and formulaic sequences such as these are being taught in the HK textbooks, there is clearly an incongruence with both native speaker intuition and varied corpora.

4.2.4 Suitability of language and presentation

Among the most commonly represented formulaic sequences in the HK textbooks are idioms. As common characteristics of idioms are non-compositionality and semantic opacity, they are among the most difficult formulaic sequences for language learners to acquire. However, some idioms are infrequently used in everyday language and so it could be argued that they should take less priority in learning than other more frequent formulaic sequences like strong collocations, such as ‘want to’ and phrasal verbs, such as ‘look for’. Perhaps the infrequent use of idioms in everyday language is one reason that idioms are difficult to acquire. Despite this, idioms have a strong presence in the HK textbooks and often seem to be associated with humour, as they are presented as jokes in two of the textbooks. Although humour can be a useful learning tool, the semantic and structural complexity of idioms may warrant more direct instruction. The students may have the impression from the way the idioms are presented in the textbook that they are only used in humorous contexts. In HKTB 2, the idioms are presented as jokes in the following question and answer format.

‘Q: Why does the monitress <u>go bananas</u>?
A: Because her classmates are naughty.’
(MI score 4.72 / 0.09 instances per million words [BNC])

If the meaning of the sequence ‘go bananas’ was unknown to a learner, it would perhaps be difficult to establish the meaning from this statement. Assuming a ‘monitress’ is to act as a

supervisor in the class, then go bananas could mean 'become strict' or similar possibilities. Although the humour has the potential to confuse students, research by Reuterskold and Van Lancker Sidtis (2012) suggests that the nonliteral nature of idioms makes the learning process faster. However, the study also found that children store the form and meaning of idioms more effectively when they are exposed to them in a natural context. The novelty effect created by the semantic opacity of idioms may have the potential to enhance learning but the context should remain natural. As idioms are overwhelmingly used in speech (McCarthy, 1998, p.145), presenting them in a natural context in a textbook is challenging. Some other idiom jokes from HKTB 2 are listed below.

'Q: Why does the cook always spill the beans?
A: Because he keeps telling people others' secrets.'
(MI score 12.25 / 0.23 instances per million words [BNC])

Although the MI score of 'spill the beans' is high in the BNC, showing strong collocation, out of the 23 instances, none of them are presented in a question or joke format.

'Q: Why can't we play jokes on a snake?
A: Because we can't pull its leg!'
(MI score 4.71 / 0 instances per million words [BNC])

There are no instances of 'pull its leg' in the BNC although if the lemma 'pull' is searched with '+ leg', there are 23 instances. The first ten lines (below) show that of the seven idiomatic uses of 'pull + leg', they are all used in present or past continuous tense with 'my' or 'your' between 'pull' and 'leg'. This is a good example of a formulaic sequence being presented in an unnatural context.

*they've had a good patrol.' 'You're **pulling my leg**!' Did he really expect her to believe*
*just the right length to stretch taut, **pulling the leg** slightly back. The old man nodded.*
*Irish.' 'It's all right, I'm only **pulling your leg**.' 'You can afford to.' 'I have*
*round the inside of his lips. She was **pulling his leg**. She always gave as much as she*
*position for a second or two. Then **pull each leg** away in opposite directions as far as it*
*who suddenly snarled, 'Don't you **pull my leg**, I'm not an animal.' The chain snapped*
*my smile, to confess that he had been **pulling my leg**; but his brooding face was drained of*

*Charles Greenwich London Are you **pulling my leg**? I rather hope you are. — Ed. Dear*
*'Don't worry, I'm Labour. I was only **pulling your leg**' His face collapsed with relief, and and said, 'No, not really, pet. Just **pulling your leg**.' The ground floor windows*

'Q: How did the farmer know the young man was the thief?
A: Because he caught him red-handed.'
(MI score 12.3 / 0.23 instances per million words [BNC])

Although the context and structure of this idiom seems more natural, it is likely to be quite difficult to understand the meaning from the context alone. The idiom is accompanied by a picture in the textbook showing a thief stealing strawberries and having red hands. If this were the first time to encounter the idiom, idiomatic meaning may be difficult to understand.

'Q: Why is the letter 'A' lazy?
A: Because it's never as busy as a 'bee'.'
(MI score 0 / 0.0 instances per million words [BNC])

The are no instances of the idiom 'busy as a bee' in the BNC. However, 'busy' is a collocate of 'bee' with seven instances of 'busy bee', and two instances of 'busy little bee'

'Q: Why does the little girl frame her dog?
A: Because she thinks it is as pretty as a picture!'
(MI score 4.76 / 0.14 instances per million words [BNC])

The context of 'pretty as a picture' seems quite complex here and lacking humour. It also takes the meaning of 'picture' perhaps too literally by using the context of a 'frame'. In all 14 of the examples in the BNC, none refer to actual pictures. In most of the contexts, the idiom is used as a direct complement such as:

e.g. 'Francesca was as pretty as a picture'

4.2.5 Formulaic expressions

One of the aims of this research is to establish the extent to which formulaic language is represented in Hong Kong primary school textbooks and how it is presented. The Government of the Hong Kong Special Administrative Region takes an active role in offering guidelines to publishers and teachers and in officially recommending textbooks for use in primary schools. The EDB specifically mentions the use of formulaic expressions in its 'English language Curriculum Guide' (2004, p.46) (figure 3), but gives no clear definition as to what is meant by formulaic expressions, leaving interpretation open to the publishers. Formulaic expressions are highlighted and listed at the back of two of the three textbooks, but the two lists appear to be quite different showing flexibility for the publishers to decide which formulaic expressions are to be included. The direct inclusion of formulaic expressions, as stated in the EDB's 'English Language Curriculum Guide' (2004), has been followed but interpreted differently.

Formulaic Expressions for Interpersonal Communication
KS2

	Examples
Use formulaic expressions to	
• make and respond to suggestions	Let's go to Stanley this weekend. That's a good idea. I'm sorry. I can't.
• show agreement or disagreement	Yes, I agree. No, I don't think so.
• open telephone conversations	Hello. May I speak to Tony, please?
• identify oneself in telephone conversations	Speaking. This is Peter.
• show concern	What's wrong? Take care.
• express and respond to good wishes	Merry Christmas. Same to you.
• begin and end formal letters	Dear Mr. Lee, Yours sincerely,

Figure 3 – Formulaic expressions for interpersonal communication KS2
(extract taken from EDB English language Curriculum Guide, 2004, p.46)

The list at the back of HKTB 2 continues with the theme of idioms, again presented in a humorous format and with the same issues of semantic opacity as the idioms previously discussed. In the example, '*I'm going to button my lips. I can't tell you the secret',* the idiom is shown with 'lips' as plural. The BNC shows no collocation of 'button' and 'lips' and no instances of 'button my lips'. However, the MI score of the singular noun 'lip' collocating with button is 6.87, although there are only two instances in the BNC. COCA shows the MI score as 4.05 with only seven instances, but again there are no instances of 'button your lips'.

In HKTB 3, the lists of formulaic sequences are presented as 'useful expressions' which are to be used in the context of giving a presentation and having a discussion. They are not idioms but expressions to be used in a specific situation as shown in some examples below:

'Hello, Mr/Mrs/Miss/Ms … and our classmates.
First, I'd like to tell you about …
Next, I want to talk about …
Take a look at this chart. It tells us …
To sum up, …
Do you have any questions?
Thank you very much.'

'OK. Let's start our discussion …
Shall we start now?
What do you think about this?
That's true, but …
Pardon?
I'm sorry. Could you say that again?
So we've decided to …'

The method of presenting formulaic sequences at the back of a textbook, where traditionally vocabulary lists of individual words might have been, seems like a good idea. However, choosing the suitability and age-appropriateness of the sequences remains an issue, and the context of a presentation seems to be aimed at a higher ability level than the language presented in the rest of the book. In a book that explicitly teaches vocabulary such as 'wealthy', 'handsome' and 'selfish', and past continuous structures such as 'he was lying on the ground',

the phrase ‘take a look at this chart. It tells us …’ seems to be at odds with the other examples. Similarly, the phrase ‘do you have any questions?’ would be a useful formulaic sequence in a university presentation or a meeting, but not necessarily in a primary school class of 10-11-year olds. The BNC has four instances of ‘do you have any questions?’: two of them from the spoken English contexts of a business meeting and a council planning meeting, one from a book on interview techniques and the other from dialogue of an interview in fiction. An interesting feature of the ‘useful expressions’ is that they appear to be sequences that would naturally be used in spoken rather than written English. The concept of formulaic expressions seems to have been interpreted as ‘spoken expressions’ by the publisher as the contexts are of a presentation and a group discussion. There are important differences between spoken and written English, and in an EFL context, the issue of how to highlight these can be problematic. Written grammar is, by definition, easy to present in a textbook full of writing, but spoken grammar is more difficult. Spoken English is unplanned and relies to a degree on the context of the conversation for meaning (Carter & McCarthy, 2006, p.164). As such, it is difficult to teach spoken formulaic phrases from a textbook as presenting them in written form is a contradiction which could be confusing to students.

The following example is taken from HKTB 2 and shows a dialogue between two students deciding when to arrange a school trip:

‘Student A: Shall we visit Mai Po Nature Reserve on 15th January?
Student B: I’m afraid not. 15th January is a Monday. Mai Po Nature Reserve opens on neither Mondays nor Thursdays.’

Corpus analysis in the BNC shows that when checking the structure of ‘verb \ neither \ + \ nor’, only the verbs ‘be’ ‘have’ ‘do’ and the modal ‘could’ come before ‘neither + nor’, in the first ten instances as shown on the following page.

or bad luck, the physical world ***is neither for nor*** *against the actions of the individual*
appear to be crinoids that ***have neither stalk nor*** *rootlets and are lying in a*
of various Arab slivers, America ***has done neither itself nor*** *Israel a service. Those whom*
on absolute certainty. Herluin ***is neither small nor*** *young. And why should any brother*
perfectly ordinary people, who ***were neither Pharaohs nor*** *priests, full brother-sister*
Too Late For a moment or two I ***could neither breathe nor*** *move. Then I felt my fear*
see the back of it. There ***was neither joy nor*** *peace in the house and she foundto*
allreviews can be discounted as ***being neither criticism nor*** *evaluation. The nuggets of*
cope with the problem, though it ***was neither clear nor*** *agreed what the basic problem*
which has been received and it ***is neither necessary nor*** *logical, simply because the*

Intuitively, it seems that the phrase 'opens on neither Mondays nor Thursdays' would sound more natural with the verb 'open' after 'neither' as in ''neither opens on Mondays nor Thursdays', but most natural would be 'It doesn't open on Mondays or Thursdays'.

4.2.6 Region specific English

Some of the sequences found in the HK textbooks appear to be odd-sounding to a native-speaker's intuition, but regional specific English has to be considered. As discussed in Chapter 2, there are many varieties of global English and what may sound unnatural to a native speaker of one variety may sound normal to another. For students, teachers and publishers of the textbooks that they use, a decision must be reached as to what standard will be taught and learnt and how strictly that standard is adhered to. In HKTB 3, the verb 'visit' is used as a teaching point with different nouns used in the sequence. Table 19 shows the different nouns and the MI scores of the different collocates, from the BNC, COCA and the NOW corpus.

visit (node) + collocate	MI - BNC	MI - COCA	MI - NOW
visit (a palace / palaces)	1.55 / 0	0 / 0	1.68 / 2.65
visit (a temple / temples)	3.82 / 4.99	2.01 / 3.02	3.70 / 4.44
visit (a museum / museums)	3.67 / 4.80	3.21 / 4.96	2.62 / 4.10
visit (a water market / water markets)	0 / 0	0 / 0	0 / 0
visit (the police college / police colleges)	0 / 0	0 / 0	0 / 0
visit (amusement park / amusement parks)	0 / 0	0 / 0	0 / 0

Table 19 – Collocates of 'visit'

There are similarities in all corpora for the nouns ‘museum’ and ‘temple’, but ‘palace’ does not collocate with ‘visit’ in COCA. Furthermore, the nouns ‘water market’, ‘police college’ and ‘amusement park’ do not collocate with ‘visit’ in any of the corpora. Does that mean that the sequence is ‘wrong’? This question has no easy answer and it is important to consider both points of view when examining the use of formulaic language in different regional contexts. Hong Kong has a locally well-known police college and so ‘visiting’ that in the contexts of school trips may be a normal situation and discussed as such.

When looking at the most frequent nouns that are used with ‘visit’, names of cities and countries dominate in the BNC. However, in the more up-to-date COCA and the internet-focused NOW corpus, websites are the most ‘visited’ collocates. This also highlights the changing nature of language and collocations, and the need for corpora to be updated.

When considering natural-sounding sequences, lexis must also be taken into account. Different regions of the world have differences in things such as customs, jobs and activities. An English sequence that has roots in the L1 of a region may be commonly used in that region but not elsewhere and not in native English-speaking countries. HKTB 1 has a chapter discussing Chinese New Year, which suggests the collocations of ‘health’, ‘happiness’ and ‘wealth’ for the verb ‘bring’. Table 20 shows the MI scores for those words collocating with ‘bring’.

Collocates of ‘bring’	BNC	COCA	NOW
health	-1.7055	n/a	n/a
happiness	5.02	4.26	4.93
wealth	n/a	2.21	3.20

Table 20 – Suggested collocates of ‘bring’ in Hong Kong textbook

These MI scores suggest that although ‘bring happiness’ appears to be a formulaic sequence, ‘health’ is not a collocate of ‘bring’ in any of the corpora used. ‘Bring wealth’ seems not to be used in British English but is more common in American English and frequent in global English. The results are inconsistent and again emphasise the need for caution in labelling a

sequence ‘native-like’. Even if the sequences were not present in any corpora, would this mean that they should not be used in Hong Kong English? If these phrases are said in Chinese during festivals, it is natural that translations or near translations would be used in English if there is no native English alternative.

4.3 Formulaic language in the UK textbook

The UKTB uses literary texts, but in what sense are they authentic? It can be argued that literature often seeks to be creative and veer away from formulaicity that may be present in everyday language use, especially speech. Textbooks used for native English speakers have a very different function from those intended for learners of the language. Perhaps teaching creativity in language is one of the goals of textbooks used in the UK. The following sequences from the UKTB are not represented in the BNC corpus and could also be described as odd-sounding.

a) ‘*he got his living’* (UKTB, Text 46)

The MI score of ‘make’ collocating with ‘living’ is 4.1633, but with ‘get’ it is -0.8501.

b) *‘Christmas was coming on’* (UKTB, Text 46)

There are 9 examples of the sequence ‘Christmas was coming’ in the BNC, but none of them have ‘on’ after ‘coming’.

c) *“And how, pray, did you come to enter my dominions?”* (UKTB, Text 48)

‘Pray’ does not collocate with ‘how’ or ‘did’ in the BNC. The word ‘dominions’ is used very formally in relation to royalty and governance in the BNC but there are 164 occurrences.

Creativity with language exists in many situations from advertising to joke telling (Carter, 2004, p.18) and is a part of everyday life. These examples highlight the difficult balance

needed to both respect the fixed nature of formulaic sequences and at the same time have the ability to know when and how to adapt them.

4.4 Discussion

Although the central aim of this research was to count formulaic language in the HK textbooks, the sequences first needed to be identified. Some of the identification issues that are often discussed in formulaic language research and literature (Wray and Namba, 2003; Durrant, and Mathews-Aydınlı, 2011; Martinez and Schmitt, 2012) were encountered during the analysis stage of this study, even though a dual identification approach of intuition and collocation was taken. While considering the results of the research, it is useful to re-visit the identification criteria and discuss some of the issues that were faced with identification.

4.4.1 Intuition and identification

The method of identification used for this research was firstly to use a set of five identification criteria to justify native speaker intuition. The criteria that were used proved invaluable in the process and allowed a consistent method of identification to be carried out over a period of time. However, native speaker judgement can be very subjective and thresholds of judgement can change even with a single researcher (Wray, 2002. p.23). Despite having a prepared set of criteria, interpretation can be inconsistent. Through applying the criteria during this research, it was discovered that boundaries of interpretation also need to be firmly established. Some of the issues with the identification criteria were as follows:

A) Have restricted exchangeability

This criterion proved very useful for identifying fixed sequences, but the nature of fixedness is that it exists on a scale and it is difficult to restrict a phrase to the description of either fixed or not. As with the other criteria used, the purpose is to aid the researcher in justifying intuition, but it may be the case that intuition is not enough to identify fixedness. Corpus studies show that variation of fixed expressions is widespread (Moon, 1999, p.120), and perhaps beyond the scope of intuition. The definition of restricted exchangeability given by Erman and Warren (2000, p.32) is that 'at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity'. Without the use of corpus linguistic evidence, it is not always clear whether a replacement of a word with a synonym has changed the meaning, function or idiomaticity. Changes may be very subtle and undetectable by intuition. An example of a sequence from HKTB 2 (Text 29) was 'period of time'. It was considered whether 'length of time' could replace the sequence 'period of time', but when analysing the sequences in the BNC (as shown below), it was discovered that in the first ten instances 'short' is used twice and 'small' and 'limited' once as descriptive adjectives for 'period of time'. For the sequence 'length of time', the adjectives 'considerable', 'great', 'uncertain' and 'reasonable' are used.

*high-calorie food in a relatively small **period of time** to supply our daily requirements, thus*
*is likely to be performed for a longer **period of time** (duration). Environmental events*
*for their jobs for only a limited **period of time**. On this basis, some 7 per cent*
*services they provide. After a **period of time** we will then ask local groups to*
*that you can attain that goal over a **period of time**. If you have trouble sorting out what*
*to the proceedings on the Bill for a **period of time** equal to the duration of the*
*to work overseas for a short **period of time**, it is common for employers*
*and classified accidents over a **period of time**. Comparisons with targets and*
*they should be given a short **period of time** to clear the mess up, or we will*
*successfully applied over a **period of time**, that have made the centre a more*

*number of files, for a very considerable **length of time**. They were monotonous in the sense*
*capitals, if he expected to stay for any **length of time**, to equip himself with a mass of*
*choice of exercise, controlling the **length of time** for each, selecting the background*
*quite good! So, he's only had the same **length of time** doing this as you have so, by the end*
*Organisation standard criteria. The **length of time** to disease progression and survival*

*the service is usually related to the **length of time** customers are actively connected to*
*no more for a great and uncertain **length of time** and had left me, and the castle, in*
*according to such criteria as **length of time** on the waiting-list, current*
*the purchaser is allowed a reasonable **length of time** from when he discovers it to be*
*follow if you are to stick with it for any **length of time**. A diet that involves uncommon*

In other examples, it was considered whether replacing a sequence such as the phrasal verb 'find out' with a single lexical item, such as 'discover', disproved restricted exchangeability. In this case, it was decided that if the phrase were to be completely transformed into an unrecognisable phrase or a single unit, then this would not constitute replacement.

B) Be grammatically irregular

Although grammatical irregularity has been used as an identification criterion for this research and some formulaic sequences do fit this category, it is useful for a researcher to first consider what English grammar is. Grammar can be described from different viewpoints. Prescriptive grammar explains how a language should be used and descriptive grammar describes how the language is actually used by people who speak and write it (Huddleston, 2005, p.4). If the global nature of English is to be accounted for, descriptive grammar can become complicated. What is grammatically regular in one variety of English may not be in another. Consider the responses to the questions below:

Q: When will you eat breakfast?
A: I have eaten it already. (British English)
A: I ate it already. (American English)

Even the same variety of English, such as British English, may have many different variations. Trudgill (2016, p.90) gives examples of dialect that is used in Norfolk, UK, where the 's' reflecting the third person singular has been omitted from the end of the verb as in 'he drive very fast' and 'she like that a lot'. Although this localised grammar shows irregularity from

standard English, it does not identify the sequence as formulaic. The identification criterion of being grammatically irregular should be based on standard English, but it should also be recognized that the boundaries of standard English can be obscure.

C) Lack semantic transparency

To a certain degree, lack of semantic transparency of a formulaic sequence would be better assessed by a non-native speaker rather than by native-speaker intuition. After all, every word, or phrase, lacks meaning until the meaning is learnt. As it may be expected that a native-speaker knows (and possibly has known since childhood) the meaning of a formulaic sequence, the semantic opaqueness may not be apparent. A non-native speaker may be a far better judge as to which sequences appear semantically transparent or opaque. Even when considering the meaning from a non-native perspective, establishing a lack of semantic transparency is not clear-cut. During the identification process, the sequence 'make money' (HKTB 1, Text 13) was identified as lacking semantically transparency. This was based on the assumption that the meaning of 'make' is to build or put something together, not as in this case 'earn'. However, it is not necessarily the case that a learner of English only knows one meaning of 'make'. When looking up the word 'make' in the 'Oxford Advanced Learner's English-Chinese Dictionary', the first entry is:

'to create or prepare something by combining materials or putting parts together… [e.g.] to make a table/dress/cake'

Further down the list is the entry:

'to earn or gain money… [e.g.] she makes $100,000 a year'

It could easily be the case that a learner focuses on the example of ‘make money’ which, it could be argued, may be more relevant than ‘make a table/dress/cake’. In this case, which use of ‘make’ lacks semantic transparency? Of course, the first meaning of ‘make’ is not a fixed phrase in respect of the many nouns that can be used as an object, such as ‘make a table/chair/bed/desk’, whereas ‘make’ as in ‘money’ has a degree of restricted exchangeability. Another issue that needs addressing when considering semantic transparency is the context of the sequence. Some formulaic phrases can be both compositional and non-compositional. ‘Kick the bucket’, for example, has both the idiomatic meaning of die and the literal meaning of kicking the bucket (Cacciari, 2014). Shin and Nation (2008) treated different meanings of the same collocation as though they were different sequences, and it seems to be an important part of the identification process. Establishing a lack of semantic transparency requires reference to the context, but sometimes establishing a hidden meaning may require corpus analysis too. In the sequence ‘live with him’ (HKTB 3, Text 34) the full sentence is:

‘The Beast let the man go, but he wanted one of his daughters to live with him’

The sequence ‘live with him/her’ appears to be quite complex. For example, if Mary tells her friend, ‘I’m living with John now’, the friend may very well assume that Mary is now in a relationship with John. If this is the case, the sequence appears to be literal, but at the same time carrying a hidden meaning or assumption. This would make the sequence lacking in semantic transparency to a certain degree. In the BNC, the first ten instances of the sequence ‘live with him’ (shown below) reveal that in four cases, the sequence ‘live with him’ carries the hidden assumption of a relationship, while the other six cases do not.

the sufferer daily (and if they did not **live with him** or her this often involved several visits
if he's asking for the child to go to **live with him** [pause] and you oppose that then
woman, wanted the children to **live with him**. The local authority was considering
talk to Woolley, you couldn't **live with him**. Killion felt hatred flare inside his
when he advertised for a woman to **live with him** on a desert island. Lucy Irvine
thing she wanted, she said, was to **live with him** in some godforsaken bog, cut off from

father and step-mother also came to **live with him** for a time. Richard Baxter only
decided to give up my lodgings to **live with him**. Well, I did a moonlight flit and stole
I think you might do better to **live with him** for a while, before you actually tie the
civilized beings who are forced to **live with him**, because he fascinates the inferior

In the example from HKTB 3, it is open to interpretation whether the sequence carries a hidden semantic meaning or whether it is literal. Given the context of a children's story, it was decided that, in this case, the sequence did not lack semantic transparency.

D) Perform a function other than the meaning of the words themselves/associated with a specific situation and/or register

Once again, this criterion, although helpful in understanding the role of some formulaic sequences and aiding in identification, needed some further consideration and clarification. An example, such as 'your majesty' (UKTB, Text 48) can easily have this criterion applied. It both performs a function other than the meaning of the words (used to respectfully address a monarch), and is associated with a specific situation. However, some examples were more difficult to identify through this criterion. It could be argued that many sequences are situation specific, and again this may only be confirmed through examining examples in a corpus. In HKTB 2, Text 29 contains the sequence 'heavy rain' and 'heavy rainfall'. Using intuition alone, it was considered that 'heavy rainfall' had more association with climate, while 'heavy rain' (although clearly talking about weather) can be used in more general situations. To test the accuracy of this intuition, the two sequences were searched in the BNC. Based on these results, it can be seen that in British English 'heavy rain' (with 2.29 instances per million words) is used more frequently than 'heavy rainfall' (with 0.2 instances per million words). Whereas 'heavy rain' is used in both written English (2.4 instances per million words) and spoken English (1.34 instances per million words), 'heavy rainfall' is only used in written English (0.2 instances per million words). This suggests that 'heavy rainfall' is more specific to a written register, and therefore criterion D is more applicable. Examining the first ten instances of each

sequence, also confirms that ‘heavy rainfall’ is exclusively used when discussing climate and its effects, while ‘heavy rain’, although used in the context of weather forecasts in three instances, is also used in various other contexts, such as describing the scene of an accident.

*appearing not to be reliant on **heavy rainfall**. There are two or possibly three*
*was greatly influenced by the **heavy rainfall** and the restrictions imposed by*
*not prevented overflow after **heavy rainfall**, however. Friends of the Earth*
*path has stayed dry, despite some **heavy rainfall**, though one or two spots are beginning*
*When this happens, extremely **heavy rainfall** may result along the desert coast of Peru.*
*Tisa, rise in areas where **heavy rainfall** is supplemented by melt water from*
*poor soils, steep slopes and **heavy rainfall** obtain. It is successful where temperate*
*a 20% chance each day of a good **heavy rainfall** lasting D8 hours. This chance rises to*
*said: ‘The problem with **heavy rainfall** and flash floods is that a lot of*
*clay/loam soils. However, the **heavy rainfall** could make this an uncertain crop so*

*on the surface after periods of **heavy rain** though such flooding is of short*
*she didn't know what to do. The **heavy rain** — an inch in less than four hours*
*Ireland will have continuous and **heavy rain** but gradually become more showery*
*fifteen bulbs stolen, and when that **heavy rain** comes, it fills the little cups and goes*
*‘We-ell, they're forecasting **heavy rain**,’ Simon pointed out, ‘but if I*
*The accident happened in **heavy rain** on the A 424, three miles north of*
*gale force winds and extremely **heavy rain** outside the building in which the*
*and crashed out of the running. **Heavy rain** stopped everyone after 18 laps, so it was*
*red-rose county by six wickets. **Heavy rain** had prevented any play on the first day*
*with thunder likely. Some **heavy rain** over Scotland will move north, then*

In this case, intuition was justified through the use of corpus analysis, and criterion D was applied to ‘heavy rainfall’ but not ‘heavy rain’. Corpus analysis also revealed that the MI score was higher for ‘heavy rainfall’ (MI score 9.26) than ‘heavy rain’ (MI score 8.8).

All of the identification criteria are in danger of being open to interpretation, with the potential for different researchers to apply different thresholds and boundaries. For this reason, applying IC values to formulaic sequences takes on an even greater importance. In theory, the more criteria that can be applied to a sequence, the more conclusive the identification should be. If a sequence has a value of 3, there seems less chance that it has been identified with over reaching boundaries than a sequence with a value of 1. Applying an IC value can help to strengthen the identification, but it is important that before engaging in formulaic language identification, the

criteria are defined in detail and the thresholds are set as firmly as possible. In should be emphasized that despite these issues with identification, it was found that the criteria used to identify formulaic sequences in this study proved effective and helpful in justifying native speaker intuition.

4.4.2 Corpus analysis

The fifth identification criterion was that of strong collocation, which was justified by establishing the MI score of a word collocating with another word. If the MI score was 3 or above then it was assigned the category of strong collocation. Sequences of more than two words were broken down into two-word collocations, giving an indication of which parts of the sequence collocated the most. For example, the sequence 'that's another story' (UKTB, Text 56), was broken down as follows:

'that's another story'	(that x) another	MI score - 0.35
	(is) another	MI score - 2.26
	another (story)	MI score - 4.24

In this example, 'that' and 'is' are not collocates of 'another' but 'story' is. When establishing collocation of sequences, the difficulty can be that some parts of the sequence may collocate while others do not. When identifying a sequence of more than two words through strong collocation, it had to be decided whether to assign the criterion in the case where part of the sequence collocates but another part does not. In cases such as these, it was decided that they did not qualify as strong collocations, but again this is open to interpretation. This method, however, was useful in identifying sequences within sequences. Through this approach, it could be seen that the more lexical units that were added to a node (search word), the more likely the sequence was of collocating. In the example, 'all day long' (UKTB, Text 50), 'all' was a

collocate of ‘day’ (MI 3.65), but ‘long’ was not (MI 2.48). However, ‘long’ was a collocate of ‘all day’ (MI 7.32), working in a similar way to predictive text on a mobile phone. Once ‘day’ is added to ‘all’, ‘long’ becomes a more likely option.

Another area that required clarification during the analysis stage was how much of a window span should be searched. The window span is the amount of words that are searched either side of the node. As discussed earlier in chapter 3, it was decided that for this research, the span would be based on the identified sequence so that in the case of ‘look for’ (HKTB 1, Text 20), the span would be one window to the right. However, in some cases, where there is variability in a sequence, it is difficult to know how much variability is allowed for a sequence to remain formulaic. If a search was made on the sequence ‘tell x about x’, deciding the number of windows would be problematic. For example, all of the following examples have the structure ‘tell x about x’:

tell (John) about (Mary)
tell (his friend) about (the party)
tell (all of his friends and classmates) about (his new school)

The BNC, COCA and the NOW corpus are all limited to a window span of five spaces left or right, so in the last example, even the maximum span would be insufficient. To a certain degree, intuition has to be relied upon in such cases.

The research highlighted the large degree of possibilities available when conducting corpus searches, such as window span, node choice, and deciding whether to search the lemma of a word (such as all forms of a verb - ‘play’, ‘plays’, ‘played’, ‘playing’) or a fixed lexical form (only ‘played’). These issues that were revisited during the analysis stage reinforced the need for a strict and rigid set of criteria to follow.

CHAPTER 5

CONCLUSION

5.1 Conclusion

The central aim of this study was to examine formulaic language use in Hong Kong (HK) primary school English textbooks by addressing the following two research questions:

1. Are formulaic sequences represented to the same extent in Hong Kong primary school English textbooks as in a textbook intended for native English speakers in the United Kingdom?

2. Are the formulaic sequences used in Hong Kong primary school English textbooks appropriate and useful, based on corpus linguistic analysis of collocational strength and usage?

After building a corpus of 54 texts, containing over 10,000 words, formulaic sequences from three HK primary school English textbooks and a textbook used by native English speakers in the United Kingdom (UK) were identified using the identification criteria established for this study. Based on these criteria, the texts in the three HK textbooks contained an average of 10% of formulaic sequences, compared to 15% in the UKTB, compatible with results from previous research (Rayson, 2008). Taking into consideration the limitations in size and variety of the corpus, these results suggest that Hong Kong English textbooks contain fewer formulaic sequences than textbooks intended for use by native English speakers, and formulaic language is therefore not represented to the same extent. During the formulaic sequence identification stage, the criteria that were used to identify each sequence were noted and from that process an identification criteria (IC) value was placed on each formulaic sequence. It is interesting to note

that the formulaic sequences in the UKTB had an average Identification Criteria (IC) value of 3, meaning it was identified using an average of 3 out of the five identification criteria, compared to an IC value of 2.2 in the HK textbooks.

Closer qualitative analysis of examples of formulaic sequences found in the HK textbooks revealed that although some sequences, such as 'button my lips', were specifically presented as a sequence within the textbooks, corpus analysis showed that they did not collocate in naturally produced native English in corpora. From the perspective of a native English speaker, such sequences may appear to be unusual and 'odd-sounding'. This raises doubt as to the appropriateness and usefulness of the sequences as taught expressions.

5.2 Limitations of research

Corpus analysis was used in order to justify native speaker intuition, especially when determining strong collocation, of two-word sequences and within sequences. However, although this established the level of non-randomness between words, it did not give a clear indication of how frequently sequences occur in naturally produced language. Perhaps comparing normalized frequencies between sequences and between different corpora would have given a better indication of frequency. In addition, the identification criterion of 'strong collocation' could be misleading as an MI score of 3 is not an indication of 'strong' collocation only that there is a certain level of non-randomness to the sequence. On reflection, a label of simply 'collocation' may have been more appropriate.

The BNC was used as the main corpus for analysis because British English has been identified as the standard (Cummings and Wolf, 2011) and because the BNCweb offers a user-friendly interface capable of a variety of searches (Hoffmann et al., 2008). However, compared to other corpora, such as COCA and NOW, the BNC has a limitation in size and is somewhat outdated

(Davies, 2009). For this reason, much of the qualitative stages of analysis was supported by COCA and NOW.

The three HK textbooks selected for analysis were compared with a UK textbook to establish whether formulaic language was represented to the same extent in the HK textbooks. The age group selected for both the HK textbooks and the UK textbook was aged 10-11. After analysis, it was evident that comparison between textbooks intended for native speaker children would not be a fair comparison for textbooks intended for learners of English. Although a UK textbook intended for a lower age would have perhaps been more equal in language level, establishing which level would be comparable would have been a difficult task. However, it would be interesting to compare textbooks intended for lower ages of native speaker children with the HK textbooks to determine if the levels of formulaic language were still considerably higher.

5.3 Implications of research

The research carried out for this thesis has highlighted some of the difficult issues associated with identification in formulaic language research. Any study which addresses formulaic language identification can hopefully help to form a clearer understanding of the issues. For this research, one of the aims was to quantify formulaic language within texts and, therefore, a set of criteria was adapted from previous studies. The criteria that were applied offered justification for identification of the formulaic sequences and could be utilised in further formulaic language identification research. The results of the research suggest that HK English textbooks may contain less formulaic language than is found in textbooks intended for the UK, and this may potentially apply to English as a foreign language (EFL) textbooks used in other regions around the world. Although the central focus of identification was on the HK textbooks,

formulaic sequences were also counted in the UKTB. There has been limited research conducted on quantifying how much formulaic language there is in spoken or written native English and results from research that does exist is inconsistent, with figures ranging from 15% (Rayson, 2008) to over 50% (Erman and Warren, 2000). These figures suggest that more research is needed in quantifying formulaic language and trying to establish a reliable mean figure. The formulaic language identified in the UKTB plays a small role in that equation with the figure of 15% sitting at the lower end of the range.

The recording of which criteria were used for the identification of formulaic sequences in this research led to the creation and application of the IC value. The IC value shows how many criteria could be applied in the identification of each formulaic sequence. There is a logical hypothesis to suggest that the more criteria applied to a sequence, the more recognisable the sequence is as formulaic. Whether this means that the sequence is more formulaic, and what consequences that may have for formulaic language research, remains unknown without further investigation. However, the IC value could potentially be used as a tool for quantifying degrees of formulaicity, and selecting sequences for teaching and learning English as a foreign language (EFL). It does not necessarily follow, however, that sequences with a higher score would be better suited for such a purpose. The development and use of IC values for this research may give an indication of how identifiable a sequence is as formulaic language, but to establish usefulness, especially in EFL, it would benefit from being used in combination with frequency measures.

Despite recognition by the Education Bureau of the Government of the Hong Kong Special Administrative Region (EDB) that there is a need for authentic texts, most of the texts in the HK textbooks have been created to suit the teaching point and formulaic expressions of each unit. It is a difficult balance between highlighting a phrase for learning and using an authentic text. Unnaturally repeating formulaic expressions in a text could hinder a student's

understanding of the context in which the formulaic expression should be used. On the other hand, authentic texts would not be able to highlight the formulaic expression. The texts in the UKTB are authentic, but the purpose of the texts is different from those in an EFL context. They are included to teach comprehension, not explicitly teach formulaic language. There perhaps needs to be a combination of both varieties of texts in the HK textbooks: artificial texts for highlighting a specific formulaic sequence and authentic texts for discovering formulaicity in a natural context. It is possible that this already occurs at higher levels of language learning, but it could also be applied at primary school level. Selection of texts, however, would be difficult. Texts intended for native speaker 10-11-year olds may be too advanced for language learners of the same age, and although texts intended for younger native speakers may contain the appropriate level of vocabulary for 10-11-year-old learners, the context may be too childish to maintain interest.

5.4 Potential for further research

Due to limitations in size and scale of this research, the formulaic language in the textbooks was identified and described from the perspective of the collocation of lexical units rather than focus on the grammatical pattern or construction of the whole sequence. Pattern grammar and construction grammar, which consider the fixed grammatical patterns and constructions of a sequence, would complement a lexical approach to research into identification and description of formulaic sequences This triangulation of formulaic language, pattern grammar and construction grammar could provide an opportunity for a more complete analysis and understanding of the sequences.

Further exploration of IC values could help to gain insight into the usefulness, if any, of applying an IC value to formulaic sequences. Applying IC values to a larger set of formulaic

sequences would help to contextualise different varieties. Research utilising methods such as questionnaires and interviews could be used to verify perceptions of sequences which could then be compared to IC values.

By using other approaches to formulaic language identification (such as pattern grammar and construction grammar) and perhaps utilising the application of IC values, it may be possible to establish which formulaic sequences have the potential to be the most appropriate and useful to learners of English. These in turn could be applied to teaching and learning strategies and used within an EFL context.

5.5 Reflections on research

It should become apparent very quickly to any researcher who engages in a project which aims to identify formulaic language, that the prototypical formulaic sequence is extremely elusive. Before any data collection can begin, the process of defining formulaic language must take place and this is a difficult process because what constitutes a formulaic sequence is very much open to interpretation. Even with a defined set of criteria, it is conceivable that different judges can interpret the criteria in different ways. This issue makes comparison of results between studies problematic, and it may be the case that formulaic language needs to be viewed with this in mind and from a flexible point of view. Nevertheless, formulaic language research is an exciting area of linguistics and there remains much to be discovered in many different contexts, especially in the area of language acquisition.

I started this thesis, and the research that it describes, from the perspective of an EFL teacher, looking for a way to improve the effectiveness and efficiency of teaching and learning EFL. While that position has remained unchanged, I have become more aware of the importance of English as a global language, not only how it is spoken in regions where English is the majority

language, but also in regions where it is not. What defines a 'type' of English? Singapore English is widely accepted as a legitimate variety, with Singapore's Ministry of Education claiming that English is used as the home language by 50% of children (Deterding, 2007, p.4). In Hong Kong, although English is an official language, competency levels of English have been described as low (Bacon-Shone et al., 2015), but does it have its own variety of English? It is tempting when discussing the advantage of formulaic language in second language acquisition to use terms such as 'native-like' proficiency or fluency, and my experience as an EFL teacher suggests that this is the goal that most students, who actively seek out English tuition, are aiming for. Searches in corpora compiled of different Englishes, reveal the many variations in single lexical units and formulaic sequences. 'Sweep ice-cream men of the streets' may be a phrase in English which has been influenced by the Chinese translation, but who is to say that it is not a valid Hong Kong English sequence? What makes a native English variety more relevant than a Hong Kong English variety or any other English variety? Indeed, in a global context, it may be the case that using formulaic phrases from native British English regions such as Newcastle, Glasgow, Liverpool or Birmingham may cause more communication problems than not using them. It is likely that many non-native and native English speakers around the world, and even in the UK, would have difficulty understanding the following spoken English from the West Midlands, UK.

'Go if yam a going, doh stop.'

'Yam' is sometimes used in West Midlands dialect to mean 'you are' and 'doh' means 'don't' (Clark and Asprey, 2013, p.138).

It is a difficult compromise between advocates of global Englishes or translanguaging and the expectations of language learners to attain a native-like proficiency in their target language. Formulaic language is clearly a relevant part of language and language learning, but the sequence 'yam a goin' would not be very well understood outside of the West Midlands. What

is odd-sounding to one group of English speakers will be relatable and communicative to another. The gold standard variety of English appears to be as elusive as the formulaic sequences contained within it.

APPENDIX 1

COLLOCATIONAL STRENGTH AND FREQUENCY OF FORMULAIC SEQUENCES (BNC)

Hong Kong Textbook 1 (HKTB 1)

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 1				
wants to	want (to)	4.4956		46.39
wants to	wants (to)	4.5065		46.39
grow up	grow (up)	5.87		5.83
lots of	lots (of)	4.94		37.26
all day	(all) day	3.6528		18.7
spend all day	(spend) all day	7.828		0.26
Text 2				
think of	think (of)	1.4771	x	67.56
so that	so (that)	3.2577		230.62
had to	had (to)	1.4498	x	271.5
need to	need (to)	4.106		223.14
at school	(at) school	3.854		24.85
wants to	wants (to)	4.5065		46.39
Text 3				
lots of	lots (of)	4.94		37.26
it‘s [just] like	it’s (like) it’s just (like)	3.2031 4.6818		22.66
need to	need (to)	4.106		223.14
so that	so (that)	3.2577		230.62
going to	going (to)	4.4906		334.08
brush [his] teeth	brush (x teeth)	8.541		0.25
not surprisingly	(not) surprisingly	6.5544		8.54
brothers and sister	brothers (x sister)	5.8853		0.08
Text 4				
wants to	wants (to)	4.5065		46.39
set off	set (off)	5.9048		16.24
moved to tears	moved (to) moved to (tears)	2.9443 7.9205		0.21
found out	found (out)	3.9171		12.69
as [po] would say	(would) say (as x) would say	3.8775 2.8082		0.77
Text 5				
looking for	looking (for)	4.7898		55.76
there are	there (are)	4.9293		406.4
lots of	lots (of)	4.94		37.26

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
you won't believe your eyes	believe (x eyes) (won't) believe your eyes (you) won't believe your eyes	3.7319 - -		0
got into trouble	(got x) trouble (into) trouble	3.4439 5.0259		0.47
as usual	(as) usual	4.9827		13.68
making sacrifices	making (sacrifices)	4.5361		0.02
giving up	giving (up)	4.4237		4.98
Text 6				
finds out	finds (out)	4.3788		1.15
a long time	a long (time) (long) time	7.6991 5.8186		36.95
Text 7				
one day	(one) day	4.7618		42.89
hears about [it]	hears (about)	4.3683		0.16
finds out	finds (out)	4.3788		1.15
giving up	giving (up)	4.4237		4.98
in the end	(the) end (in) the end in (x end)	3.4154 2.7267 1.8471		31.66
Text 8				
got off	got (off)	3.0215		4.48
as much as	(as) much much (as)	3.9438 3.216		37.32
Text 9				
blew away	blew (away)	5.7756		0.29
sit down	sit (down)	8.1254		19.37
good luck	good (luck)	7.8775		5.19
Text 10				
there is	there (is)	4.3757		592.89
there are	there (are)	4.9293		406.4
queue up	queue (up)	4.6983		0.48
Text 11				
last night	last (night)	8.512		86.07
there are	there (are)	4.9293		406.4
in need of	(in) need need (of)	0.2509 -0.1355	x	7.87
Text 12				
raise money	raise (money)	7.74		4.34
decided to	decided (to)	4.2532		65.39
'd rather	('d) rather	5.3387		9.06
raise [much] money	raise (x money)	6.006		1.3
something else	something (else)	7.7924		19.83
run a stall	run (x stall) run (x x stall)	2.7702 3.9852		0.01

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
would [you] rather	(would x) rather	1.4331	x	2.51
burnt down	burnt (down)	6.522		0.96
sorry, i'm late	sorry (i) sorry (x'm) sorry (x x late) sorry i (x late)	2.8524 4.2307 4.744 8.1217		0.25
turned on	turned (on)	2.4938	x	8.54
Text 13				
raised [more than] £200,000	raised (x x £) raised (x £)	0 5.1141		
more than	more (than)	6.9371		335.93
decided to	decided (to)	4.2532		65.39
hoped to	hoped (to)	3.1888		10.4
make £500	make (£) make (money)	0 3.6296	x	
want to	want (to)			290.86
loads of	loads (of)	4.4551		8.71
lost their lives	lost (x lives)	5.7084		0.82
make money	make (money)	3.6296		3.16
burst into tears	burst (into) burst (x tears)	7.4663 11.384		2.22
make money	make (money)	3.6296		3.16
proud of	proud (of)	3.9154		12.61
wants to	wants (to)	4.5065		46.39
raise money	raise (money)	7.74		4.34
Text 14				
help out	help (out)	1.7589	x	2.22
all over the world	(all x x) world (x over x) world all over the (world)	2.6604 3.6077 9.0336		7.32
clean up	clean (up)	5.1743		4.21
needs of	needs (of)	-2.440	x	33.22
around the world	around (the) around (x world)	2.2745 5.6504		10.73
would [you] rather	(would x) rather	1.4331	x	2.51
Text 15				
around the world	around (the) around (x world)	2.2745 5.6504		10.73
there are	there (are)	4.9293		406.4
one another	(one) another	4.3032		27.13
full of	full (of)	2.8585	x	56.3
Text 16				
a lot of	(a) lot lot (of)	5.4049 4.3891		148.83

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
hundreds of years ago	hundreds (of) hundreds (x years) hundreds (x x ago)	4.9109 5.9508 4.9456		0.2
there is	there (is)	4.3757		592.89
Text 17				
spend time	spend (time)	4.5238		2.32
washed away	washed (away)	7.1189		1.21
a clean start	clean (start)	2.6431	x	0.05
set free	set (free)	3.4504		0.9
Text 18				
all over the world	(all x x) world (x over x) world all over the (world)	2.6604 3.6077 9.0336		7.32
next to	next (to)	1.4791	x	29.44
my dear	my (dear)	7.3784		14.73
have a drink	have (x drink)	3.6168		2.96
Text 19				
there was	there (was)	4.3067		501.44
looked around	looked (around)	5.6905		6.53
hardly any	hardly (any)	5.317		3.69
run out	run (out)	4.5996		9.28
the rest of the	(the) rest rest (of) rest (x the)	3.8071 4.3812 2.8941		52.34
have to	have (to)	2.0131	x	439.05
decided to	decided (to)	4.2532		65.39
there was	there (was)	4.3067		501.44
oh no	oh (no)	4.3752		28.11
worried about	worried (about)	7.7294		16.06
come back	come (back)	6.1089		40.29
had a look	(had x) look	1.2997	x	1.82
next to	next (to)	1.4791	x	29.44
went back	went (back)	5.5531		18.83
tried to	tried (to)	4.8496		96.5

APPENDIX 2

COLLOCATIONAL STRENGTH AND FREQUENCY OF FORMULAIC SEQUENCES (BNC)

Hong Kong Textbook 2 (HKTB 2)

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 21				
tell a joke	(tell) joke	3.2909		0.08
hang on	hang (on)	6.0639		13.98
think about	think (about)	4.613		37.41
get it	get (it)	2.0798	x	38.73
i see	(i) see	2.5675	x	52
get carried away	(get) carried carried (away)	2.8777 5.868		0.76
Text 23				
once upon a time	once (upon) upon (x time)	4.4721		1.63
there was	there (was)	4.3067		501.44
a full head of hair	(full) head head (of) (of) hair	0.4394 2.5366 0.2911	x	0.02
decided to	decided (to)	4.2532		65.39
go for a walk	(go x x) walk	3.8795		0.9
walk on	walk (on)	2.3242	x	3.31
in front of	front (in) front (of)	3.3037 3.7501		62.19
wanted to	wanted (to)	4.5728		123.37
want to	want (to)	4.4956		290.86
there were	there (were)	4.5741		215.59
just right	(just) right	1.5163	x	2.54
take a nap	(take x) nap	6.1313		0.07
lay down	lay (down)	6.3333		6.41
fell asleep	fell (asleep)	10.386		2.77
looking forward to	looking (forward) forward (to)	8.4068 3.2505		10.55
stamped [his] foot	stamped (x feet)	8.6598		
noticed that	noticed (that)	4.379		10.35
messed up	messed (up)	7.6967		0.62
run for [her] life	run (for) run (x x life)	1.8349 0.9505	x	0.1
from that [day] on	(that) day Day (on)	1.609 0.6536	x	0.15

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 24				
endangered species	endangered (species)	12.429		2.57
home to	home (to)	1.0171	x	23.82
loss of	loss (of)	4.0977		53.68
lack of	lack (of)	5.0091		87.68
[grow] up to	grow (up) grow (x to)	5.87 0.0318		0.61
spends [the] winter	spends (x winter)	5.4234		0.01
make it [hard] for	make (it) make (x hard) make (x x for)	3.114 1.7886 1.7144		5.53
spends [most of the] time	spends (time)	5.5855		0.1
Text 25				
ladies and gentlemen	ladies (x gentlemen)	12.949		2.75
welcome to	welcome (to)	2.4548	x	8.15
no entry	(no) entry	2.4023	x	0.53
at all times	(all) times (at x) times	3.3269 4.1089		6.86
want to	want (to)	4.4956		290.86
Text 26				
at first	(at) first	3.2787		52.66
after a while	after (x while)	3.8311		7.64
felt sick	felt (sick)	7.0514		1.35
throw up	throw (up)	4.7464		1.67
get off	get (off)	3.8964		8.75
at last	(at) last	3.7008		43.61
there were	there (were)	4.5741		215.59
have the barbecue	(have x) barbecue	1.5872	x	0.04
having a barbecue	(having x) barbecue	3.3402		0.01
having a picnic	(having x) picnic	6.0826		0.11
oh dear	oh (dear)	8.505		18.22
in tears	(in) tears	2.2003	x	3.01
out of	out (of)	3.1616		483.12
burst out laughing	burst (out) burst (x laughing)	6.3094 11.240		1.24
never mind	never (mind)	6.5235		11.99
that's [very] kind of you	kind (of) kind (of x)	4.6454 0.8		0.07
there was	there (was)	4.3067		501.44
soaked to the skin	soaked (to) soaked (x x skin)	1.1002 9.7103		0.21
too much	too (much)	7.0743		73.21
dry off	dry (off)	2.5149	x	0.22
after a while	(after x) while	4.108		7.64
joined in	joined (in)	2.1707	x	5.6

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
on [our] way home	way (home) (our) way	3.7751 3.2669		3.86
fell asleep	fell (asleep)	10.386		2.77
gives me the giggles	(gives x x) giggles	0		0
Text 27				
are left with no	left (with)	1.8987	x	0.02
under the sea	under (x sea)	2.8636	x	0.51
last night	last (night)	8.512		86.07
heavy rain	heavy (rain)	8.8008		2.29
affected areas	affected (areas)	5.3806		0.52
shut down	shut (down)	6.6017		3.86
in force	(in) force	1.5426	x	7.89
lack of	lack (of)	5.0091		87.68
food and water	food (x water) (water x) food	4.6838 2.4015		0.84
Text 28				
hanging out	hanging (out)	4.779		1.41
in the morning	(in x) morning	3.5987		37.42
nice and cool	(nice x) cool cool (x nice)	4.6834 0		0.08
hope so	hope (so)	3.0616		3.1
starts to let up	let (up) (starts x) let	-0.0216 0.5452		0
want to	want (to)	4.4956		290.86
pay attention to	(pay) attention attention (to)	6.537 3.738		1.62
it’s time	(it x) time	0.8761	x	7.25
turn on	turn (on)	1.4414	x	3.65
good evening	good (evening)	6.1847		6.99
strong winds	strong (winds)	9.3175		1.46
heavy rainfall	heavy (rainfall)	9.2569		0.2
this evening	this (evening)	4.3627		11.02
oh dear	oh (dear)	8.505		18.22
need to	need (to)	4.106		223.14
make plans	make (plans)	2.6552	x	0.51
Text 29				
a set of	set (of)	2.4105	x	23.96
in alphabetical order	alphabetical (order)	10.294		0.68
a period of time	period (of) period (x time)	3.1922 4.9642		2.89
heavy rainfall	heavy (rainfall)	9.2569		0.2
strong winds	(strong) winds	9.3406		1.46
heavy rain	heavy (rain)	8.8008		2.29

APPENDIX 3

COLLOCATIONAL STRENGTH AND FREQUENCY OF FORMULAIC SEQUENCES (BNC)

Hong Kong Textbook 3 (HKTB 3)

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 31				
spent hours	spent (hours)	5.9468		0.93
excited about	excited (about)	5.8519		1.8
in the morning	(in x) morning	3.5987		37.42
taking photos	taking (photos)	6.0232		0.1
would rather	would (rather)	2.5832	x	5.56
running around	running (around)	5.1732		1.91
spend time	spend (time)	4.5238		2.32
a long time	(long) time	5.8186		36.95
so much	so (much)	5.705		100.75
Text 32				
summer holidays	summer (holidays)	9.0626		1.51
went swimming	went (swimming)	4.7212		0.29
‘d rather	‘d (rather)	5.2821		9.06
next time	next (time)	4.3445		12.61
want to	want (to)	4.4956		290.86
Text 33				
want to	want (to)	4.4956		290.86
tell [you] about	tell (x about)	4.9541		
went fishing	went (fishing)	3.7229		0.18
it was getting dark	getting (dark)	4.9095		0.34
at night	(at) night	4.2319		30.81
sleep well	sleep (well)	3.083		0.78
Text 34				
once upon a time	once (upon) once (x x time)	4.4721 1.9751		1.63
lost his money	lost (x money) lost (money)	2.5534 3.5014		0.04
one day	(one) day	4.7618		42.89
would like to	(would) like would like (to)	4.1832 4.9912		41.85
got lost	got (lost)	3.1416		1.34
on [his] way	(on x) way	3.1167		53.11
fell asleep	fell (asleep)	10.386		2.77
live with [him]	live (with)	3.7248		13.22
arrived home	arrived (home)	5.1726		1.38
told [his daughter] about	told (x about)	4.1252		10.5
one day	(one) day	4.7618		42.89

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
on the ground	(on x) ground	4.8603		24.68
turned into	turned (into)	5.255		12.51
Text 35				
were held	were (held)	3.7374		10.11
stand for	stand (for)	2.7469	x	5.73
are held	are (held)	2.5806	x	6.72
Text 36				
were held	were (held)	3.7374		10.11
took part in	took (part) took (x in)	5.2864 1.6697		5.51
drawn attention from	drawn (attention) drawn attention (from) drawn attention (to)	7.3655 0 5.2961	x	0
stand for	stand (for)	2.7469	x	5.73
Text 38				
fill in	fill (in)	3.5462		8.07
worry about	worry (about)	7.6824		18.73
Text 39				
want to	want (to)	4.4956		290.86
do [my] homework	(do x) homework	5.372		0.74
take a shower	(take x) shower	4.743		0.22
Text 41				
around the world	around (the) around (x world)	2.2745 5.6504		10.73
near and far	near (x far)	2.9257	x	0.27
made money	made (money)	0.5651	x	0.45
a lot of	lot (of)	4.3891		148.83
one day	(one) day	4.7618		42.89
had a /car accident/	(had x) accident (car) accident	2.6016 6.4554		0.01
took [his] life	took (x life)	0.0695	x	0.19
Text 42				
i'm doing fine	doing (fine) ('m) doing (i x) doing	4.1942 5.4203 2.7969		0.5
last week	last (week)	7.9184		51.7
after school	after (school)	2.6974	x	2.46
at night	(at) night	4.2319		30.81
Text 43				
want to	want (to)	4.4956		290.86
around the corner	(around x) corner	6.7997		2.65
in the old days	old (days) (in x) old (in x) old days	5.2409 1.121 4.8542		2.82

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
lots of	lots (of)	4.94		37.26
in front of	(in) front front (of) in front (of)	4.274 3.7501 5.0246		62.19
from [eleven o'clock] to [five o'clock]	from (x to)	1.2549	x	235.32
lunch breaks	lunch (breaks)	6.9189		0.09
after school	(after) school	2.7166	x	2.46
a piece of	piece (of) (a) piece	4.5988 4.11		25.78
a [carton] of	carton (of) (a) carton	3.2135 3.3929		0.21
long for	long (for)	-0.5268	x	3.08

APPENDIX 4

COLLOCATIONAL STRENGTH AND FREQUENCY OF FORMULAIC SEQUENCES (BNC)

United Kingdom Textbook (UKTB)

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 46				
fair's fair	fair (x fair)	5.1931		0.11
got his living	(got x) living (made x) living	0.433 1.9366	x	0 0.1
running errands	running (errands)	10.139		0.16
a bit of	bit (of)	3.3205		51.12
on the side	(on x) side	4.4198		9.71
nothing better to do	(nothing) better nothing better (to) nothing better (x do)	4.1619		0.5
might have	might (have)	5.103		85.12
for sure	for (sure)	1.2833	x	4.5
one day	(one) day	4.7618		42.89
never seen or heard of again	(never) seen seen (x heard) never seen or heard (x again) heard (x again)	6.4158 4.2162 0 2.27		0
Text 47				
run away	run (away)	6.0273		6.06
not bothered	(not) bothered	5.108		2.11
worry about	worry (about)	7.6824		18.73
keep it in check	(keep x x) check	4.7885		0.03
fell round	fell (round)	-0.4807	x	0.02
lined up	lined (up)	7.5565		4.79
nothing much	nothing (much)	3.3296		2.63
came out	came (out)	4.9225		24.18
caught sight of	caught (sight) caught (x of) sight (of)	9.1647 0.9008 3.7302		2.68
oh no	oh (no)	4.3752		28.11
today was the day	today (x x day)	1.9084	x	0.04
terrified of	terrified (of)	2.8521	x	2.07
Text 48				
cut off	cut (off)	6.7873		11.53
your majesty x 4	your (majesty)	6.4476		1.13
said she	said (she)	1.7373	x	20.76
stood still	stood (still)	4.5857		1.9

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
once and for all	once (x for) once and for (all)	0.8232 8.64		2.78
lose my patience	lose (x patience)	5.6186		0.01
come to	come (to)	2.8719	x	114.41
come in	come (in)	1.865	x	42.53
found myself here	found (myself) found (x here) found myself (here)	6.4933 -0.1436 0		0
full in the face	(full x x) face (in x) face	-0.1436 2.3151	x	0.27
going to	going (to)	4.4906		334.08
unable to	unable (to)	5.3812		60.15
just as	just (as)	3.4318		80.32
gave himself up for lost	gave (himself) gave (x up) gave (x x x lost) gave (up)	3.6518 1.8942 -1.795 4.3571		0
appeared to	appeared (to)	3.7176		31.49
change her mind	change (x mind)	6.0565		0.45
how cold you look	cold (x look)	-0.1989	x	0
Text 49				
one foot in front of the other	(one) foot foot (x front) (in) front front (of) front (x x other)	4.0827 3.5679 4.274 3.7501 0.9604		0.09
holding on	holding (on)	2.1145	x	2.28
oh no, not [he]	oh (no) oh no (not)	4.3752 0.6273		0
one step at a time	(one) step step (x x time) one step (x x time)	4.1353 2.5979 6.0714		0.29
every other	every (other)	3.8388		7.8
a good deal more	good (deal) good (x more)	6.7133 0.0487		1.13
be made [harder] still	made (harder) made (x still)	1.6385 -1.7313		0
something told [him]	something (told) something (x him)	0.6775 1.4359		0.25
Text 50				
grew [cold]	grew (cold)	4.7294		0.12
darkness fell	darkness (fell)	7.2778		0.45
far too [big]	far (too)	6.1501		

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
was nowhere to be found	(was) nowhere nowhere (to) nowhere (x be) nowhere (x x found)	3.6305 3.0124 3.142 4.9304		0.11
ran off	ran (off)	5.9745		3.16
all day long	(all) day day (long) all day (long)	3.6528 2.4773 7.3179		1.46
made [her] way	made (x way)	3.7153		10.05
the smell hung on the air	smell (hung) smell (x x x air)	6.0462 5.2145		0.07
drew [her knees] up	drew (x knees) drew (x x up)	5.3302 1.5695		0.01
go home	go (home)	5.1383		13.92
had to	had (to)	1.4498	x	271.5
it seemed as if	(it) seemed seemed (as) seemed (x if)	4.3871 1.3871 1.8722	x	1.42
stretched out	stretched (out)	6.847		3.99
went out	went (out)	4.8464		23.41
Text 51				
dear sir	dear (sir)	6.5657		1.29
parked cars	parked (cars)	9.8589		0.83
in spite of	(in) spite spite (of)	5.7579 5.1017		27.53
pick up	pick (up)	7.7946		26.11
of course	of (course)	4.4963		301.49
it is only a matter of time before	(it x x x) matter (x is x x) matter (only x) matter matter (of) matter (x time) matter (x x before)	2.662 2.2467 3.2632 3.2439 3.1922 3.7942		0.18
yours faithfully	yours (faithfully)	13.343		1.55
Text 52				
stamp out	stamp (out)	4.9532		0.77
gathering momentum	gathering (momentum)	10.718		0.26
send round	send (round)	1.4899	x	0.06
in addition to	(in) addition addition (to)	5.5263 4.1211		35.07
high street	high (street)	7.519		11.93
make [your views] known	(make x) known (make x x) known	1.7659 2.3538	x	0.59
of course	of (course)	4.4963		301.49
making a difference	make (x difference)	6.3377		0.09

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 53				
tell [you] about	tell (about)	4.9541		15.2
get home	get (home)	3.5523		5.13
by the way	(by x) way	1.9031		15.68
Text 54				
i am writing to complain about	writing (x complain) writing (x x about)	4.6348 2.0695		0.02
appeared from nowhere	appeared (x nowhere)	6.2519		0.15
knocked [her] over	knocked (x over)	3.4218		0.27
take a look	(take x) look	4.0682		4.08
yours sincerely	yours (sincerely)	14.269		8.04
Text 55				
to whom it may concern	(to) whom whom (it) whom (x may) whom (x x concern)	2.5354 1.6941 2.2734 3.1475		0.08
scribbling down	scribbling (down)	5.7651		0.05
beyond [your] wildest dreams	(beyond x) wildest wildest (dreams)	10.772 14.079		0.17
of course	(of) course	4.4963		301.49
made it easy	(made x) easy	2.6325	x	0.83
jumbled up	jumbled (up)	6.2176		0.15
all will become clear	(all x) become become (clear)	-2.0938 4.9283	x	0.01
make sure	make (sure)	8.1171		46.06
in front of	(in) front front (of) in front (of)	4.274 3.7501 5.0246		62.19
be sure not to	(be) sure sure (not) sure (x to)	3.8516 -2.1099 -1.1977	x	0.1
get in touch	get (x) touch (in) touch	5.9324 3.7666		3.93
look at	look (at)	5.9427		150.15
that's another story	(that x) another (is) another another (story)	0.3513 2.2632 4.2387	x	0.29
looking for	looking (for)	4.7898		55.76
let [me] down	let (down)	4.4319		4.53
fall into the wrong hands	(fall x x) wrong wrong (hands)	2.8489 4.0754	x	0.09
good luck	good (luck)	7.8775		5.19
end of message	end (x message)	0.5924	x	0.04

Formulaic Sequence	Collocation Node (collocate/s)	MI	Lacking Strong collocation	Frequency per million words
Text 56				
is there	is (there)	0.7238	x	46.63
leaned over	leaned (over)	6.5595		2.24
a host of	host (of)	3.2711		6.24
in the quiet of the	(in x) quiet	1.8481	x	0.09
kept [my] word	kept (x word)	3.0151		0.18
Text 57				
up to speed	(up x) speed	1.6743		0.23
a copy of	copy (of)	4.0488		17.9
limbered up	limbered (up)	8.4939		0.02
make sure	make (sure)	8.1171		46.06
know your way around	(know x) way way (around)	0.9499 3.3081		0.03
get distracted	get (distracted)	4.6711		0.1
tempted to	tempted (to)	4.8716		7.97
a host of	host (of)	3.2711		6.24
under way	under (way)	4.2017		9.67
one of the	one (of) one (x the)	3.0036 1.5454	x	361.22
name that tune	name (x tune)	1.7196	x	0.01
be sure to	(be) sure sure (to)	3.8516 0.2431	x	2.52
fully air-conditioned	(fully) air-conditioned	10.814		0.23
wall-to-wall	wall (x wall)	4.9363		0.37
a number of	number (of)	4.7149		153.96
quite a few	quite (x few)	5.7179		8.1
value for money	value (x money)	7.2063		7.42
plan on	plan (on)	-0.0747	x	0.99
a lot of	lot (of)	4.3891		148.83
night ‘on the town’	(night x x) town (on x) town	2.3499 0.7119	x	0.15

APPENDIX 5

IDENTIFICATION CRITERIA (IC) VALUE OF FORMULAIC SEQUENCES

Hong Kong Textbook 1 (HKTB 1)

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 1		
1.want to	S	1
2.grow up	A, C, S	3
3.want to	S	1
4. lots of	S	1
5.want to	S	1
6. all day	A, B, S	3
7.wants to	S	1
8.wants to	S	1
9.wants to	S	1
10. spend [all day]	A, C, S	3
		16÷10=1.6
Text 2		
1. think of (opinion)	C	1
2. so that	D, S	2
3. had to	A, C	2
4. so that	D, S	2
5. need to	S	1
6. so that	D, S	2
7.at school	A, B, S	3
8.wants to	S	1
		14÷8=1.8
Text 3		
1. lots of	S	1
2. it‘s [just] like	A, C, D, S	4
3. need to	S	1
4. so that	D, S	2
5. going to	C, D, S	3
6. brush [his] teeth	S	1
7. so that	D, S	2
8. not surprisingly	B, S	2
9. brothers and sister	A, S	2
		18÷9=2
Text 4		
1. wants to	S	1
2. set off	C, S	2
3. moved to tears	A, C, S	3
4. found out	C, S	2
5. as [po] would say	D, S	2
		10÷5=2

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 5		
1. looking for	C, S	2
2. there are	A, D, S	3
3. lots of	S	1
4. won't believe your eyes	A, C, S	3
5. got into trouble	A, C, S	3
6. as usual	S	1
7. as usual	S	1
8. making sacrifices	A, C, S	3
9. giving up	A, C, S	3
		20÷9=2.2
Text 6		
1. finds out	A, C, S	3
2. a long time	D, S	2
		5÷2=2.5
Text 7		
1. one day	A, C, D, S	4
2. hears about [it]	A, S	2
3. finds out	A, C, S	3
4. giving up	A, C, S	3
5. in the end	D, S	2
		14÷5=2.8
Text 8		
1. got off	A, C, S	3
2. as much as	A, S	2
		5÷2=2.5
Text 9		
1. blew away	A, S	2
2. sit down	A, C, S	3
3. good luck	A, D, S	3
		8÷3=2.6
Text 10		
1. there is	A, D, S	3
2. there are	A, D, S	3
3. queue up	A, C, S	3
		9÷3=3
Text 11		
1. last night	A, S	2
2. there are	A, D, S	3
3. in need of	A, C	2
4. in need of	A, C	2
5. there are	A, D, S	3
6. there are	A, D, S	3
		15÷6=2.5
Text 12		
1. raise money	A, C, S	3

Formulaic Sequence	Identification Criteria (IC)	IC Value
2. decided to	S	1
3. ‘d rather	A, D, S	3
4. something else	A, B, C, S	4
5. raise [much] money	A, C, S	3
6. run a stall	C, S	2
7. would [you] rather	A, D	2
8. burnt down	A, C, S	3
9. sorry, i’m late	A, D, S	3
10.turned on	C	1
		25÷10=2.5
Text 13		
1. raised [more than] £200,000	A, C, S	3
2. more than	C, S	2
3. decided to	S	1
4. hoped to	A, S	2
5. make £500	A	1
6. want to	S	1
7. loads of	S	1
8. lost their lives	A, S	2
9. want to	S	1
10. make [some] money	A, C, S	3
11.burst into tears	A, C, S	3
12.make [so much] money	A, C, S	3
13.proud of	A, C, S	3
14.wants to	S	1
15.raise money	A, S	2
		29÷15=1.9
Text 14		
1. help out	A, C	2
2. all over the world	C, S	2
3. clean up	C, A, S	3
4. needs of	A	1
5. around the world	C, S	2
6. would [you] rather	A, D	2
		12÷6=2
Text 15		
1. around the world	C, S	2
2. there are	A, D, S	3
3. one another	B, S	2
4. full of	A	1
		8÷4=2
Text 16		
1. a lot of	C, S	2
2. hundreds of years ago	D, S	2
3. there is	A, D, S	3

Formulaic Sequence	Identification Criteria (IC)	IC Value
		7÷3=2.3
Text 17		
1. spend time	A, C, S	3
2. washed away	A, C, S	3
3. a clean start	C	1
4. set free	A, C, S	3
		10÷4=2.5
Text 18		
1. all over the world	C, S	2
2. next to	A, B	2
3. my dear	A, D, S	3
4. have a drink	A, S	2
5. have a drink	A, S	2
		11÷5=2.2
Text 19		
1. there was	A, D, S	3
2. looked around	A, S	2
3. hardly any	A, S	2
4. run out	A, C, S	3
5. the rest of the [crew]	C, S	2
6. have to	A, C	2
7. decided to	S	1
8. there was	A, D, S	3
9. oh no	C, D, S	3
10.worried about	S	1
11.decided to	S	1
12.come back	S	1
13.had a look	A	1
14.next to	A, B	2
15.went back	S	1
16.tried to	S	1
		29÷ 16=1.8
Text 20		
1. is about	A, S	2
2. sets off	C, S	2
3. look for	C, S	2
4. is about	A, S	2
5. finds out	A, C, S	3
6. is about	A, S	3
7. makes up	A, C, S	3
8. make [the king] laugh	A, B, C, S	4
9. is about	A, S	2
10.want to	S	1
11.get away	A, C, S	3
12.cause trouble	S	1
13.a lot of	C, S	2

Formulaic Sequence	Identification Criteria (IC)	IC Value
14.is about	A, S	2
15.gets into trouble	A, C S	3
16.is about	A, S	2
17.break down	A, C, S	3
		40÷ 17=2.4

APPENDIX 6

IDENTIFICATION CRITERIA (IC) VALUE OF FORMULAIC SEQUENCES

Hong Kong Textbook 2 (HKTB 2)

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 21		
1. tell (you) a joke	A, S	2
2. hang on	A, C, S	3
3. think about [thought]	A, S	2
4. get it	C	1
5. get it	C	1
6. i see	C	1
7. get carried away	A, C, S	3
		13÷ 7=1.9
Text 22		
1. tongue twister	A, C, S	3
		3÷ 1=3
Text 23		
1. once upon a time	A, B, C, D, S	5
2. there was	A, D, S	3
3. a full head of hair	A, C	2
4. decided to	S	1
5. go for a [walk]	S	1
6. walk on	A, C	2
7. in front of	A, B, S	3
8. wanted to	S	1
9. want to	S	1
10.there were	A, D, S	3
11. decided to	S	1
12. just right	A, C	2
13. decided to	S	1
14. take a nap	C, S	2
15. lay down	A, S	2
16. fell asleep	A, C, S	3
17. looking forward to	A, C, S	3
18. stamped [his] foot	A, S	2
19. noticed that	S	1
20. messed up	A, C, S	3
21. woke up	A, C, S	3
22. out of bed	B, S	2
23. run for [her] life	A, C	2
24. from that [day] on	B, C	2
		51÷ 24=2.1
Text 24		
1. endangered species	A, S	2
2. home to	A, C	2

Formulaic Sequence	Identification Criteria (IC)	IC Value
3. loss of habitat	A, S	2
4. lack of	A, S	2
5. all over	A, C, S	3
6. [grow] up to	A, B, C, S	4
7. spends [the] winter	A, C, S	2
8. make it [hard] for	A, C, S	3
9. out of	S	1
10.[grow] up to	A, B, C, S	4
11.make it [hard] for	A, C, S	3
12.spends time	A, C, S	3
13.most of the	A, C, S	3
		34÷ 13=2.6
Text 25		
1. ladies and gentlemen	A, D, S	3
2. welcome to	A, D	2
3. at all times	A, D, S	3
4. no entry	A, D	2
5. want to	S	1
		11÷ 5=2.2
Text 26		
1. at first	A, B, C, S	4
2. after a while	A, S	2
3. felt sick	A, C, S	3
4. throw up	A, C, S	3
5. get off	A, C, S	3
6. at last	A, B, C, S	4
7. there were	A, D, S	3
8. there were	A, D, S	3
9. have the barbecue	A	1
10.having a barbecue	A, S	2
11.having a picnic	A, S	2
12.oh dear	C, D, S	3
13.in tears	A, C	2
14.out of	S	1
15.burst out laughing	A, C, S	3
16.never mind	A, C, D, S	4
17.that’s [very] kind of you	A, S	2
18.there was	A, D, S	3
19.oh dear	C, D, S	3
20.soaked to the skin	C, S	2
21.too much	A, S	2
22.dry off	A, C	2
23.after a while	A, S	2
24.joined in	A, C	2
25.on [our] way home	A, B, C, S	4
26.fell asleep	A, C, S	3

Formulaic Sequence	Identification Criteria (IC)	IC Value
27.gives me the giggles	A, C, S	3
		70÷ 27=2.6
Text 27		
1. are left with no	C	1
2. get back to normal	C, S	2
3. under the sea	A, C	2
4. last night	A, C, S	3
5. heavy rain	A, C, S	3
6. affected areas	A, S	2
7. shut down	A, C, S	3
8. in force	A, C	2
9. lack of	A, S	2
10.food and water	A, S	2
		22÷ 10=2.2
Text 28		
1. hanging out	A, C, S	3
2. in the morning	A, S	2
3. nice and cool	A, S	2
4. hope so	A, B, C, S	4
5. starts to let up	A, C, S	3
6. want to	S	1
7. want to	S	1
8. pay attention to	A, C, S	3
9. it's time	A, C, D	3
10. turn on	C	1
11.good evening	A, D, S	3
12. strong winds	A, S	2
13. heavy rainfall	A, C, D, S	4
14. this evening	A, S	2
15. strong winds	A, S	2
16. heavy rainfall	A, C, D, S	4
17. heavy rainfall	A, C, D, S	4
18. oh dear	C, D, S	3
19. need to	S	2
20.make plans	A	1
		48÷ 20=2.4
Text 29		
1. a set of	A	1
2. in alphabetical order	A, S	2
3. a period of time	A, S	2
4. heavy rainfall	A, C, D, S	4
5. strong winds	A, S	2
6. heavy rain	A, C, S	3
		14÷ 6=2.3

APPENDIX 7

IDENTIFICATION CRITERIA (IC) VALUE OF FORMULAIC SEQUENCES

Hong Kong Textbook 3 (HKTB 3)

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 31		
1. spent [many] hours	C, S	2
2. excited about	A, C, S	3
3. in the morning	A, S	2
4. taking photos	A, C, S	3
5. would rather	A, D	2
6. running around	C, S	2
7. spend a long time	C, S	2
8. so much	A, S	2
		18÷ 8=2.3
Text 32		
1. summer holidays	A, S	2
2. summer holidays	A, S	2
3. went swimming	A, S	2
4. ‘d rather	A, D, S	3
5. next time	A, C, S	3
6. want to	S	1
		13÷ 6=2.2
Text 33		
1. want to	S	1
2. tell [you] about	A, C, S	3
3. went fishing	A, S	2
4. it was getting dark	C, S	2
5. at night	A, S	2
6. sleep well	S	1
		11÷ 6=1.8
Text 34		
1. once upon a time	A, B, C, D, S	5
2. lost [all his] money	A, C, S	3
3. one day	A, C, D, S	4
4. got lost	C, S	2
5. on [his] way	C, S	2
6. fell asleep	A, C, S	3
7. live with [him]	A, S	2
8. arrived home	B, S	2
9.told [his daughter] about	A, S	2
10.one day	A, C, D, S	4
11.on the ground	A, S	2
12.turned into	A, C, S	3
		34÷ 12=2.8

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 35		
1. were held	A, C, S	3
2. stand for	A, C	2
3. are held	A, C	2
		7÷ 3=2.3
Text 36		
1. were held	A, C, S	3
2. took part in	C, S	2
3. drawn [much] attention from	A, C	2
4. stand for	A, C	2
		9÷ 4=2.3
Text 37		
NO SEQUENCES		
Text 38		
1. fill in	A, C, S	3
2. worry about	A, S	2
		5÷ 2=2.5
Text 39		
1. want to	S	1
2. do [my] homework	A, S	2
3. want to	S	1
4. take a shower	C, S	2
		6÷ 4=1.5
Text 40		
NO SEQUENCES		
Text 41		
1. around the world	C, S	2
2. near and far	A, D	2
3. made money	A, C	2
4. one day	A, C, D, S	4
5. had a [car] accident	A, S	2
6. took [his] life	A, C	2
		14÷ 6=2.3
Text 42		
1. i'm doing fine	B, C, S	3
2. last week	A, S	2
3. after school	A	1
4. at night	A, S	2
		8÷ 4=2
Text 43		
1. want to	S	1
2. around the corner	A, S	2
3. in the old days	C, D, S	3
4. lots of	S	1

5. in front of	A, B, S	3
6. from [eleven o'clock] to [five o'clock]	A, D	2
7. lunch breaks	A, S	2
8. after school	A	1
9. a piece of	S	1
10. a [carton] of	A, S	2
11. long for	A, C	2
		20÷ 11=1.8

APPENDIX 8

IDENTIFICATION CRITERIA (IC) VALUE OF FORMULAIC SEQUENCES

United Kingdom Textbook (UKTB)

Formulaic Sequence	Identification Criteria (IC)	IC Value
Text 46		
1. fair's fair	A, B, C, D, S	5
2. got his living	B, C	2
3. running errands	A, C, S	3
4. a bit of	A, S	2
5. on the side	A, C, S	3
6. nothing better to do	A, B, D, S	4
7. might have	D, S	2
8. for sure	B, C	2
9. one day	A, C, D, S	4
10.never seen or heard of again	A, C, D, S	4
		31÷10=3.1
Text 47		
1. run away	C, S	2
2. not bothered	A, C, D, S	4
3. worry about	A, S	2
4. keep it in check	A, B, C, S	4
5. fell round	A, C	2
6. lined up	A, C, S	3
7. nothing much	A, B, C, S	4
8. came out	S	1
9. caught sight of	A, C, S	3
10. oh no	C, D, S	3
11.today was the day	A, B, D	3
		31÷11=2.8
Text 48		
1. cut off	A, C, S	3
2. your majesty	A, C, D, S	4
3. said she	A, B, D	3
4. stood still	A, C, S	4
5. once and for all	A, C, B, S	4
6. lose [my] patience	A, C, S	3
7. your majesty	A, C, D, S	4
8. come to	A, C	2
9. your majesty	A, C, D, S	4
10. came in	A	1
11. found myself here	A, C, S	3
12. your majesty	A, C, D, S	4
13. full in the face	A, B, C	3
14. going to	C, D, S	3
15. unable to	S	1

Formulaic Sequence	Identification Criteria (IC)	IC Value
16. just as	A, B, C, S	4
17. gave himself up for lost	A, B, C, S	4
18. appeared to	S	1
19. change her mind	A, C, S	3
20. how cold you look	A, B	2
		60÷20=3
Text 49		
1. one foot in front of the other	A, C, S	3
2. holding on	A, C	2
3. oh no, not [he]	B, C, D, S	4
4. one step at a time	A, C S	3
5. every other	A, B, C S	4
6. a good deal more	A, B, C, S	4
7. be made [harder] still	A, B, C	3
8. something told [him]	A, C	2
9. sort of	B, C S	3
		28÷9=3.1
Text 50		
1. grew [cold]	A, C, S	3
2. darkness fell	A, C, S	3
3. far too [big]	B, C, S	3
4. was nowhere to be found	A, B, C, S	4
5. ran off	C, S	2
6. all day long	A, B, C, S	4
7. made her [weary] way	A, C, S	3
8. the smell hung on the air	C, S	2
9. drew [her knees] up	C, S	2
10. go home	A, B, S	3
11. had to	A, C	2
12.it seemed as if	B, C	2
13. stretched out	A, C, S	3
14. went out	A, C, S	3
		39÷ 14=2.8
Text 51		
1. dear sir	A, D, S	3
2. parked cars	A, S	2
3. in spite of	A, B, C, S	4
4. pick up	A, C, S	3
5. of course	A, B, C, S	4
6. it is only a matter of time before	D, S	2
7. yours faithfully	A, B, C, D, S	5
		23÷ 7=3.3
Text 52		
1. stamp out	A, C, S	3
2. gathering momentum	C, S	2
3. send round	C	1

Formulaic Sequence	Identification Criteria (IC)	IC Value
4. in addition to	A, B, C, S	4
5. high street	A, C, D, S	4
6. make [your views] known	A, C, D	3
7. of course	A, B, C S	4
8. making a difference	A, C, S	3
		24÷ 8=3
Text 53		
1. tell [you] about	A, C, S	3
2. get home	A, B, C, S	4
3. by the way	A, C, S	3
		10÷ 3=3
Text 54		
1. i am writing to complain about	A, D, S	3
2. appeared from nowhere	C, S	2
3. knocked [her] over	A, C, S	3
4. take a look	C, A, S	3
5. yours sincerely	A, B, C, D, S	5
		15÷ 5=3
Text 55		
1. to whom it may concern	D, S	2
2. scribbling down	C, S	2
3. beyond [your] wildest dreams	A, C, S	3
4. of course	A, B, C, S	4
5. made it easy	A, C	2
6. jumbled up	C, S	2
7. all will become clear	C	1
8. make sure	A, B, C, S	4
9. in front of	A, B, S	3
10. be sure not to	D	1
11. get in touch	A, B, C, D, S	5
12. look at	A, S	2
13. that's another story	A, C, D	3
14. looking for	A, C, S	3
15. let me down	A, C, S	3
16. fall into the wrong hands	A, C	2
17. good luck	A, C, D, S	4
18. end of message	A, D	2
		48÷ 18=2.7
Text 56		
1. is there	A, D, S	3
2. is there	A, D, S	3
3. leaned over	A, C, S	3
4. a host of	A, C, S	3
5. in the quiet of the	A, B, C, D	4
6. kept [my] word	C, S	2

Formulaic Sequence	Identification Criteria (IC)	IC Value
		18÷ 6=3
Text 57		
1. up to speed	A, B, C, S	4
2. a copy of	A, S	2
3. limbered up	A, C, S	3
4.make sure	A, B, C, S	4
5. know your way around	B, C, S	3
6. get distracted	A, B, C, S	4
7. tempted to	A, C, S	3
8. a host of	A, C, S	3
9. under way	A, B, C, S	4
10. one of the	A, S	2
11. name that tune	A, C, D	3
12. be sure to	B, C, D	3
13. fully air-conditioned	A, D, S	3
14. a number of	C, S	2
15. quite a few	A, B, C, S	4
16. value for money	A, B, C, S	4
17. plan on	C	1
18. a lot of	C, S	2
19. night ‘on the town’	A, B, C	3
		58÷ 19=3.1

APPENDIX 9

IDENTIFICATION CRITERIA (IC) VALUE CATEGORIES

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
HKTB 1					
Text 1	7		3		
Text 2	3	4	1		
Text 3	3	4	1	1	
Text 4	1	3	1		
Text 5	3	1	5		
Text 6		1	1		
Text 7		2	2	1	
Text 8		1	1		
Text 9		2	1		
Text 10			3		
Text 11		3	3		
Text 12	2	2	5	1	
Text 13	6	4	5		
Text 14	1	4	1		
Text 15	1	2	1		
Text 16		2	1		
Text 17	1		3		
Text 18		4	1		
Text 19	7	5	4		
Text 20	2	8	6	1	
Total (out of 142 sequences)	37	52	49	4	
%	26%	37%	35%	3%	0%

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
HKTB 2					
Text 21	3	3	2		
Text 22			1		
Text 23	7	9	7	0	1
Text 24	1	5	5	2	
Text 25	1	2	2		
Text 26	2	10	11	4	
Text 27	1	6	3		
Text 28	4	6	6	4	
Text 29	1	3	1	1	
Total (out of 114 sequences)	20	44	38	11	1
%	18%	39%	33%	10%	1%

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
HKTB 3					
Text 31		6	2		
Text 32	1	3	2		
Text 33	2	3	1		
Text 34		6	3	2	1
Text 35		2	1		
Text 36		3	1		
Text 37			-		
Text 38		1	1		
Text 39	2	2			
Text 40			-		
Text 41		5	0	1	
Text 42	1	2	1		
Text 43	4	5	2		
Total (out of 66 sequences)	10	38	14	3	1
%	15%	58%	21%	5%	3%

Formulaic Sequence	Identification Criteria (IC) value				
	1	2	3	4	5
UKTB					
Text 46		4	2	3	1
Text 47	1	3	4	3	
Text 48	3	2	7	8	
Text 49		2	4	3	
Text 50		5	7	2	
Text 51		2	2	2	1
Text 52	1	1	3	3	
Text 53			2	1	
Text 54		1	3		1
Text 55	2	7	5	3	1
Text 56		1	4	1	
Text 57	1	4	8	6	
Total (out of 130)	8	32	51	35	4
%	6%	25%	39%	27%	3%

BIBLIOGRAPHY

Ackermann, K. and Chen, Y.H. (2013) Developing the academic collocation list (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12 (4): 235-247

Arnold, W. (2008) *New magic 6A.* Hong Kong: Oxford University Press (China) Ltd.

Arnold, W. (2008) *New magic 6A.* Hong Kong: Oxford University Press (China) Ltd.

Bacon-Shone, J., Bolton, K. and Luke, K.K. (2015) *Language use, proficiency and attitudes in Hong Kong*. Hong Kong: The University of Hong Kong.

Berry, R. (2015) *From words to grammar: discovering English usage*. London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999) *Longman grammar of spoken and written English*. London: Longman.

BNC Consortium (2007) *The British National Corpus, Version 3 (BNC XML Edition)* [online]. Distributed by Bodleian Libraries, University of Oxford. Available from: http://www.natcorp.ox.ac.uk/

Boulter, C., Gregson, H., Muller, A., Poynton, H. and Sharrock, J. (2012) *English: The 11+ practice book: ages 9-10*. Newcastle upon Tyne: CGP.

Boulter, C., Gregson, H., Poynton, H. and Sharrock, J. (2012) *English: The 11+ practice book: Ages 10-11*. Newcastle upon Tyne: CGP.

Broadbent, P. and Head, A. (2012) *KS2 learn and practise: maths and English*. London: Letts Educational.

Bullen, E. and Gray, J. (2010) *English to enjoy 6A.* Hong Kong: Educational Publishing House Ltd.

Bullen, E. and Gray, J. (2010) *English to enjoy 6B.* Hong Kong: Educational Publishing House Ltd.

Cacciari, C. (2014) Processing multiword idiomatic strings: many words in one? *The Mental Lexicon*, 9 (2): 267-293

Carrol, G. and Conklin, K. (2014) Getting your wires crossed: evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, 17(4): 784-797

Carrol, G. and Conklin, K. (2015) Cross language lexical priming extends to formulaic units: evidence from eye-tracking suggests that this idea 'has legs'. *Bilingualism: Language and Cognition*, 20 (2): 299-317

Carrol, G., Conklin, K. and Gyllstad, H. (2016) Found in translation: the influence of the L1 on the reading of idioms in an L2. *Studies in Second Language Acquisition* [online]. Available from: http://dx.doi.org/10.1017/S0272263115000492 [Accessed 20 October 2016]

Clark, U. and Asprey, E. (2013) *West Midlands English: Birmingham and the Black Country*. Edinburgh: Edinburgh University Press.

Collins COBUILD English language dictionary. (1987). London: HarperCollins.

Collins COBUILD advanced learner's English dictionary. 5th ed. (2006). Glasgow: HarperCollins.

Conklin, K. and Pellicer-Sánchez, A. (2016) Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32 (3): 453-467

Cortazzi, M. and Jin, L. (1996) English teaching and learning in China. *Language Teaching*, 29 (2): 61-80

Creese, A. and Blackledge, A. (2010) Translanguaging in the bilingual classroom: a pedagogy for learning and teaching? *The Modern Language Journal*, 94 (1): 103-115

Cummings, P.J. and Wolf, H. (2011) *A dictionary of Hong Kong English: words from the fragrant harbor*. Hong Kong: Hong Kong University Press.

Cutler, A. (1982) Idioms: the colder the older. *Linguistic Inquiry*, 13 (2): 317-320

Dallas, D. and Pelham, L. (2005) *Longman welcome to English 6A*. Hong Kong: Pearson Hong Kong.

Dallas, D. and Pelham, L. (2005) *Longman welcome to English 6B*. Hong Kong: Pearson Hong Kong.

Davies, M (2008) *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present* [online]. Available from: https://corpus.byu.edu/coca/

Davies, M. (2009) The 385+ million word Corpus of Contemporary American English (1990-2008+): design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14 (2): 159-190

Davies, Mark. (2013) *Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day* [online]. Available from: https://corpus.byu.edu/now/

Deterding, D. (2007) *Singapore English*. Edinburgh University Press: Edinburgh.

Downie, M., Gray, D. and Jimenez, J.M. (2014) *Lighthouse for Hong Kong book 11*. Hong Kong: Educational Publishing House Ltd.

Downie, M., Gray, D. and Jimenez, J.M. (2014) *Lighthouse for Hong Kong book 12*. Hong Kong: Educational Publishing House Ltd.

Durrant,P. and Mathews-Aydınlı, J. (2011) A function first approach to identifying formulaic language. *English for Specific Purposes*, 30 (1): 58-72

Education and Manpower Bureau (EDB) (2004) *English language curriculum guide* [online]. Available from: http://www.edb.gov.hk/attachment/en/curriculum-development/kla/eng-edu/primary%201_6.pdf [Accessed 26 May 2017]

Education and Manpower Bureau (EDB) (2016a) *A parent's guide to textbook matters* [online]. Available from: http://www.edb.gov.hk/en/curriculum-development/resource-support/textbook-info/parent-guide.html [Accessed 13 May 2017]

Education and Manpower Bureau (EDB) (2016b) *Guiding principles for quality textbooks* [online]. Available from: http://www.edb.gov.hk/en/curriculum-development/resource-support/textbook-info/guidingprinciples/index.html [Accessed 13 May 2017]

Education and Manpower Bureau (EDB) (2017) *Recommended textbook list* [online]. Available from: https://cd.edb.gov.hk/rtl/search.asp [Accessed 13 May 2017]

Ellis, N. C., Simpson-Vlach, R. and Maynard, C. (2008) Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42 (3): 375-396

Erman, B. and Warren, B. (2000) The idiom principle and the open choice principle. *Text*, 20 (1): 29-62

Evans, S. and Green, C. (2007) Why EAP is necessary: a survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, 6 (1): 3-17

Firth, J.R. (1968) "A synopsis of linguistic theory, 1930-55." In Palmer, F.R. (ed.) *Selected papers of J.R. Firth 1952-59.* London: Longmans, Green and Co Ltd. pp. 168-205

Flege, J.E., Yeni-Komshian, G.H. and Liu, S. (1999) Age Constraints on Second Language Acquisition. *Journal of Memory and Language*, 41 (1): 78-104

Foster, P. (2001). "Rules and routines: a consideration of their role in the task based language production of native and non-native speakers." In Bygate, M., Skehan, P and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* Harlow: Longman. pp. 75–95

Foster-Cohen, S.H. (1999) *An introduction to child language development*. London: Longman.

Francis, G., Hunston, S. and Manning, E. (1996) *Collins COBUILD grammar patterns 1: verbs.* London: HarperCollins.

Francis, G., Hunston, S. and Manning, E. (1998) *Collins COBUILD grammar patterns 2: nouns and adjectives.* London: HarperCollins.

Garcia, O. and Wei, L. (2014) *Translanguaging: language, bilingualism and education*. Hampshire: Palgrave Macmillan.

Gardner, D. and Davies, M. (2007) Pointing out frequent phrasal verbs: a corpus-based analysis. *TESOL Quarterly*, 41 (2): 339–359

Garnier, M. and Schmitt, N. (2015) The PHaVE list: a pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19 (6): 645-666

Gibb, N. (2014) “Foreword.” In Oates, T. *Why textbooks count: a policy paper*. Cambridge: University of Cambridge.

Gibbs, R.W., Jr., (2007) “Idioms and formulaic language.” In Gerraerts, D. and Cuyckens, H. (eds.) *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press. pp. 697-725

Glenwright, P. (2005) Grammar error strike hard: language proficiency testing of Hong Kong teachers and the four "noes". *Journal of Language, Identity, and Education*, 4 (3): 201-226

Goldberg, A. (2003) Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7 (5): 219-224

Goldberg, A. (2005) “Argument realization: the role of constructions, lexical semantics and discourse factors.” In Östman, J. and Fried, M. (eds.) *Construction grammars: cognitive grounding and theoretical extensions*. Amsterdam: John Benjamins. pp. 17-43

Gray, C., Jones, R., Hall, E. and Gordon, T. (2009*) Primary Longman elect 6A*. Hong Kong: Longman Hong Kong Education.

Gray, C., Jones, R., Hall, E. and Gordon, T. (2009) *Primary Longman elect 6B*. Hong Kong: Longman Hong Kong Education.

Gray, C. and Jones, R. (2015) *Primary Longman express 6A*. 2nd ed. Hong Kong: Pearson Hong Kong.

Gray, C. and Jones, R. (2015) *Primary Longman Express 6B*. 2nd ed. Hong Kong: Pearson Hong Kong.

Gries, S.Th. (2008) “Phraseology and linguistic theory: a brief survey.” In Granger, S. and Meunier, F. (eds.) *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins. pp. 3-25

Grosjean, F. (2010) *Bilingual: life and reality*. [online]. Massachusetts: Harvard University Press. Available from: ebrary Subscription Collection. http://site.ebrary.com/lib/bham/reader.action?docID=10568024&ppg=1 [Accessed 5 August 2017]

Gyllstad, H. and Wolter, B. (2016) Collocational processing in light of the phraseological continuum model: does semantic transparency matter? *Language Learning*, 66 (2): 296–323

Hakuta, K. (1974) Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, 24 (2): 287–297

Hakuta, K. (1976) A case study of a Japanese child learning English as a second language. *Language Learning*, 26 (2): 321-351

Hatami, S. (2015) Teaching formulaic sequences in the ESL classroom. *TESOL Journal*, 6 (1): 112-129

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund Prytz, Y. (2008) *Corpus linguistics with BNCweb - a practical guide*. Frankfurt am Main: Peter Lang.

Holmes, J. (2008) *An introduction to sociolinguistics*. 3rd ed. Essex: Pearson Education Limited.

Hornby, A.S. (1954) *A guide to patterns and usage in English*. London: Oxford University Press.

Howarth P. (1998) Phraseology and second language proficiency. *Applied Linguistics*, 19 (1): 24-44

Huddleston, R. and Pullum, G. K. (2005) *A student's introduction to English grammar*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. (1999) *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Irujo, S. (1986) Don't put your leg in your mouth: transfer in the acquisition of idioms in a second language. *TESOL Quarterly,* 20 (2): 287-304

Ji, F. (2004) *Linguistic engineering: language and politics in Mao's China*. Honolulu, HI: University of Hawai'i Press.

Joseph, J.E. (1997) "English in Hong Kong: emergence and decline." In Wright, S and Holmes, H.K. (eds.) *One country, two systems, three languages*. Clevedon, England: Multilingual Matters. pp. 60-73

Kachru, B.B. (1985) "Standards, codification and sociolinguistic realism: the English language in the outer circle." In Quirk, R. and Widdowson, H.G. (eds.) *English in the world: teaching and learning the language and literatures*. Cambridge: Cambridge University Press. pp. 11-30

Kecskes, I. (2000) Conceptual fluency and the use of situation-bound utterances in L2. *Links & letters,* 7: 145-161

Kurtyka, A. (2001) "Teaching English phrasal verbs: a cognitive approach." In Pütz, M., Niemeier, S. and Dirven, R. (eds.) *Applied cognitive linguistics*. Berlin: Mouton De Gruyter.

Lenneberg, E.H. (1967) *Biological foundations of language*. New York: Wiley.

Lewis, M. (1993) *The lexical approach: the state of ELT and a way forward*. London: Thomson.

Li, D.C.S. (1999) The functions and status of English in Hong Kong: a post-1997 update. *English World-Wide,* 20 (1): 67-110

Littlemore, J. (2009) *Applying cognitive linguistics to second language learning and teaching*. Basingstoke: Palgrave Macmillan.

Littlemore, J., Chen, P.T., Koester, A. and Barnden, J. (2011) Difficulties in metaphor comprehension faced by international students whose first language is not English. *Applied Linguistics*, 32 (4): 408-429

Liu, D. (2011) The most frequently used English phrasal verbs in American and British English: a multicorpus examination. *TESOL Quarterly*, 45 (4): 661–688

Loadman, A., Greaves, S. and Harrop, M. (2004) *Key stage 2 national test revision English*. London: HarperCollins.

Lum, C. C. and Ellis, A. W. (1994). Is "nonpropositional" speech preserved in aphasia? *Brain Language*, 46: 368–391

Mahmoud, A. (2002) Transfer of idioms by Arab students of EFL. *The Internet TESL Journal* [online], 8 (12). Available from: http://iteslj.org/Articles/Mahmoud-Idioms.html [Accessed 20 October 2016]

Marinova-Todd, S.H., Marshall, D.B. and Snow, C.E. (2000) Three misconceptions about age. *TESOL Quarterly*, 34 (1): 9-34

Martin, M., Mullis, I., Foy, P. and Stanco, G. (2011) *TIMSS 2011 international results in science*. Boston College, USA: TIMSS & PIRLS International Study Center.

Martinez, R. and Schmitt, N. (2012) A phrasal expressions list. *Applied Linguistics*, 33 (3): 299-320

Mastering 11+: multiple choice comprehension practice book 1. 2nd ed. (2014) United Kingdom: Ashkraft Educational.

McCarthy, M. (1998) *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Meyerhoff , M. (2006) *Introducing sociolinguistics*. Abingdon, Oxon: Routledge.

Millar, N. (2011) The processing of malformed formulaic language. *Applied Linguistics*, 32 (2): 129-148

Mitchell, H.Q. (2006) *Pop up now*. Hong Kong: Stanford House Publications (HK) Ltd.

Moon, R. (1998) *Fixed expressions and idioms in English*. Oxford: Clarendon Press.

Moon, R. (2009) *Sinclair, lexicography, and the COBUILD project: the application of theory*. Amsterdam: John Benjamins. pp. 1-22

Mullis, I., Martin, M., Foy, P. and Arora, A. (2012) *TIMSS 2011 international results in mathematics.* Boston College, USA: TIMSS & PIRLS International Study Center.

Myers-Scotton, C. (1995) *Social motivations for codeswitching: evidence from Africa*. Oxford: Oxford University Press.

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

Oates, T. (2014) *Why textbooks count: a policy paper*. Cambridge: University of Cambridge.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Oppenheim, N. (2000) "The importance of recurrent sequences for nonnative speaker fluency and cognition." In Riggenbach, H. (ed.) *Perspectives on fluency.* Ann Arbor: University of Michigan Press. pp. 220-240

Orwell, G. (1946) Politics and the English language. *Horizon*, 13 (76): 252-265

Oxford advanced learner's English-Chinese dictionary. 7th ed. (2005). Hong Kong: Oxford University Press (China) Ltd.

Paquot, M. and Granger, S. (2012) Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32: 130-149

Pawley, A. and Syder, F. H. (1983) "Two puzzles for linguistic theory: nativelike selection and nativelike fluency." In Richards, J.C. and Schmidt, R.W. (eds.) *Language and communication*. London: Longman. pp.191–226

Pellicer-Sánchez, A. and Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: do things fall apart? *Reading in a Foreign Language*, 22 (1): 31–55

Pellicer-Sánchez, A. (2017) Learning L2 collocations incidentally from reading. *Language Teaching Research*, 21 (3): 381-402

Peters, A.M. (1983) *The units of language acquisition*. Cambridge: Cambridge University Press.

Rayson, P. (2008) *Software demonstration: identification of multiword expressions with Wmatrix*. Paper presented at the Formulaic Language Research Network (FLaRN) Conference, University of Nottingham, Nottingham, UK.

Read, J. and Nation, P. (2004) "Measurement of formulaic sequences." In Schmitt, N. (ed.) *Formulaic sequences: acquisition, processing and use*. Amsterdam: John Benjamins. pp. 23-35

Reuterskiöld, C. and Van Lancker Sidtis, D. (2012) Retention of idioms following one-time exposure. *Child Language Teaching and Therapy*, 29 (2): 219-231

Schmitt, N. and Carter, R. (2004) "Formulaic sequences in action: an introduction." In Schmitt, N. (ed.) *Formulaic sequences: acquisition, processing and use*. Amsterdam: John Benjamins. pp. 1-22

Setter, J., Wong, C.S.P. and Chan, B.H.S. (2010) *Hong Kong English*. Edinburgh: Edinburgh University Press.

Shin, D. and Nation, P. (2008) Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, 62 (4): 339-348

Simpson-Vlach, R. and Ellis, N. C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31 (4): 487-512

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

Singleton, D. (2005) The critical period hypothesis: a coat of many colours. *International Review of Applied Linguistics in Language Teaching*, 43 (4): 269-285

Siyanova-Chanturia, A., Conklin, K. and Schmitt, N. (2011) Adding more fuel to the fire: an eye-tracking study of idiom processing by native and non-native speakers. *Second Language research*, 27 (2): 251-272

Smith, A. and Ling, J. (2005) *My pals are here! English for Hong Kong 6A*. Hong Kong: Educational Publishing House Ltd.

Smith, A. and Ling, J. (2005) *My pals are here! English for Hong Kong 6B*. Hong Kong: Educational Publishing House Ltd.

Stahl, B. and Van-Lancker sidtis, D. (2015) Tapping into neural resources of communication: formulaic language in aphasia therapy. *Frontiers in Psychology*, 6: 1-5

Tognini-Bonelli, E. (2001) *Corpus linguistics at work.* Amsterdam: John Benjamins.

Trudgill, P. (2016) *Dialect matters: respecting vernacular language*. Cambridge: Cambridge University Press.

Underwood, G., Schmitt, N. and Galpin, A. (2004) "The eyes have it: an eye-movement study into the processing of formulaic sequences." In Schmitt, N. (ed.) *Formulaic sequences: acquisition, processing and use*. Amsterdam: John Benjamins. pp. 153-172

United Kingdom (UK) Government (2017) *Prove your knowledge of English for citizenship and settling* [online]. Available from: https://www.gov.uk/english-language/exemptions [Accessed 12 June 2017]

Van Lancker Sidtis, D. and Rallon, G. (2004) Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language and Communication*, 24 (3): 207-240

Van Lancker Sidtis, D. and Postman, W. A. (2006) Formulaic expressions in spontaneous speech of left- and right-hemisphere-damaged subjects. *Aphasiology*, 20: 411–426

Van Lancker Sidtis, D. (2012) "Two-track mind: formulaic and novel language support a dual-process model." In Faust, M. (ed.) *The handbook of the neuropsychology of language, volume 1*. Sussex: Blackwell Publishing Ltd.

Webb, S., Newton, J., and Chang, A. (2013) Incidental learning of collocation. *Language Learning*, 63 (1): 91–120

Weinreich, U. (1968) *Languages in contact: findings and problems*. [online]. The Hague: Mouton. Available from: ebrary Subscription Collection. http://site.ebrary.com/lib/bham/detail.action?docID=10585523 [Accessed 5 August 2017]

Willis, J. (1988) *Collins Cobuild English course: student's book 1*. London: Collins.

Willis, J. (1988) *Collins Cobuild English course: student's book 2*. London: Collins.

Willis, J. (1989) *Collins Cobuild English course: student's book 3*. London: Collins.

Willis, D. (1990) *The lexical syllabus: a new approach to language teaching*. London: HarperCollins.

Wilson, A., Clarke, K. and Hall, E. (2013) *Primary Longman express 6A*. Hong Kong: Pearson Hong Kong.

Wilson, A., Clarke, K. and Hall, E. (2013) *Primary Longman express 6B*. Hong Kong: Pearson Hong Kong.

Wong Fillmore, L. (1976) *The second time around: cognitive and social strategies in second language acquisition*. Unpublished PhD thesis, Stanford University.

Wood, D. (2010) *Formulaic language and second language speech fluency: background, evidence and classroom applications*. London: Continuum.

Wood, D. (2015) *Fundamentals of formulaic language: an introduction*. London: Bloomsbury.

Wray, A. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008) *Formulaic language: pushing the boundaries*. Oxford: Oxford University Press.

Wray, A. (2009) "Identifying formulaic language: persistent challenges and new opportunities." In Corrigan, R., Moravcsik, E.A., Ouali, H. and Wheatley, K.M. (eds.) *Formulaic language, volume 1, distribution and historical change.* Amsterdam: John Benjamins. pp. 27-51

Wray, A. (2012) What do we (think we) know about formulaic language? an evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32: 231-254

Wray, A. and Perkins, M.R. (2000) The functions of formulaic language: an integrated model. *Language and Communication*, 20 (1): 1-28

Wray, A. and Namba, K. (2003) Use of formulaic language by a Japanese-English bilingual child: a practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism* 9 (1): 24-51

Yu, V. and McNeill, A. (2005) *Step up 6A*. Hong Kong: Educational Publishing House Ltd.

Yu, V. and McNeill, A. (2005) *Step up 6A*. Hong Kong: Educational Publishing House Ltd.