A corpus-based study of academic-collocation use and patterns in postgraduate Computer Science students' writing

Afnan Saleh. Farooqui

A thesis submitted for the degree of Doctor of Philosophy

Department of Language and Linguistics

University of Essex

February2016

Dedication

To my dear parents, caring husband and my three lovely children.

Acknowledgments

First and foremost, I should express my thanks to God (Allah) for His guidance and assistance, which enabled me to overcome difficulties throughout the past four years.

Then, my gratitude should be expressed to my supervisor, Dr.Sophia Skoufaki, for her constant encouragement, precious guidance, and inspiring remarks in all stages of research design and writing. She was always available to discuss my work and to sort out any doubts. I owe her a debt of gratitude for all she has done for me and will never be able to repay her. Moreover, I am deeply indebted to Dr. Nigel Harwood, who was my second supervisor, for his guidance and encouragement throughout the PhD. I am also grateful to Mr. Phil Scholphield whose expertise in statistical methods was an invaluable resource throughout the PhD period.

Furthermore, I am extremely grateful to the Computer Science departments at the University of Essex, Sheffield University, and Leicester University for their help in providing me with postgraduate MSc dissertations. Thanks are also due to the Saudi Ministry of Higher Education and Umm Al-Qura University who generously funded this thesis through a scholarship.

I also owe a special debt to my friends who gave me their support throughout the four years. My greatest gratitude should be expressed to my faithful friend, Alawyah Al-Sharif, whose support was invaluable. She was constantly by my side and was always happy to listen to me in my weakest moments. I would also like to thank my friends Heba Gazzaz, Nadien Halabi, and Suhad Sonbol for their encouragement.

Finally yet importantly, I would like to express my deepest gratitude to my family for their support throughout my postgraduate studies. Thanks in particular to my parents, Saleh Farooqui and Aminah Masoumi, for their sincere prayers. My very special thanks are directed to my husband Waleed Haddad, my sons Khaled and Mohammed, and my daughter Ayah. Without their genuine attention, my experience at Essex would have been much more difficult and far less rewarding.

Abstract

Collocation has been considered a problematic area for L2 learners. Various studies have been conducted to investigate native speakers' (NS) and non-native speakers' (NNS) use of different types of collocations (e.g., Durrant and Schmitt, 2009; Laufer and Waldman, 2011).These studies have indicated that, unlike NS, NNS rely on a limited set of collocations and tend to overuse them. This raises the question: if NNS tend to overuse a limited set of collocations in their academic writing, would their use of academic collocations in a specific discipline (Computer Science in this study) vary from that of NS and expert writers?

This study has three main aims. First, it investigates the use of lexical academic collocations in NNS and NS Computer Science students' MSc dissertations and compares their uses with those by expert writers in their writing of published research articles. Second, it explores the factors behind the over/underuse of the 24shared lexical collocations among corpora. Third, it develops awareness-raising activities that could be used to help non-expert NNS students with collocation over/underuse problems.

For this purpose, a corpus of 600,000 words was compiled from 55 dissertations (26 written by NS and 29 by NNS). For comparison purposes, a reference corpus of 600,269 words was compiled from 63 research articles from prestigious high impact factor Computer Science

academic journals. The Academic Word List (AWL) (Coxhead, 2000) was used to develop lists of the most frequent academic words in the student corpora, whose collocations were examined. Quantitative analysis was then carried out by comparing the 100 most frequent noun and verb collocations from each of the student corpora with the reference corpus. The results reveal that both NNS (52%) and NS (78%) students overuse noun collocations compared to the expert writers in the reference corpus. They underuse only a small number of noun collocations (8%). Surprisingly, neither NNS nor NS students significantly over/underused verb collocations compared to the reference corpus.

In order to achieve the second aim, mixed methods approach was adopted. First, the variant patterns of the 24 shared noun collocations between NNS and NS corpora were identified to determine whether over/underuse of these collocations could be explained by their differences in the number of patterns used. Approximately half of the 24 collocations identified for their patterns were using more patterns including (Noun + preposition +Noun and Noun + adjective +Noun) that were rarely located in the writing of experts. Second, a categorisation judgement task and semi-structured interviews were carried out with three Computer Scientists to elicit their views on the various factors likely influencing noun collocation choices by the writers across the corpora. Results demonstrate that three main factors could explain the variation: sub-discipline, topic, and genre.

To achieve the third pedagogical aim, a sample of awareness-raising activities was designed for the problematic over/underuse of some noun collocations. Using the corpus-based Data Driven Learning (DDL)approach (Johns,1991), three types of awareness-raising activities were developed: noticing collocation, noticing and identifying different patterns of the same collocation, and comparing and contrasting patterns between NNS students' corpora and the reference corpus.

Results of this study suggest that academic collocation use in an ESP context (Computer Science) is related to other factors than students' lack of knowledge of collocations. Expertness, genre variation, topic and discipline-specific collocations are proved important factors to be considered in ESP. Thus, ESP teachers have to alert their students to the effect of these factors in academic collocation use in subject specific disciplines. This has tangible implications for Applied Linguistics and for teaching practices.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACL	Academic Collocations List
AFL	Academic Formulas List
AI	Artificial Intelligence
AVL	Academic Vocabulary List
AWL	Academic Word List
BNC	British National Corpus
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
CJT	Categorisation Judgement Task
CLEC	Chinese Learner English Corpus
COCA	Corpus of Contemporary American English
CS	Computer Science
CSMWL	Computer Science Multi Word List
CSWL	Computer Science Word List
DDL	Data Driven Learning
EAP	English for Academic Purposes
EGAP	English for General Academic Purposes
ESAP	English for Specific Academic Purposes
EOP/EVP/EPP	English for Occupational /Vocational /Professional Purposes
ESP	English for Specific Purposes
FLOB	Freiburg-LOB Corpus
GAC	General Academic Collocation
GCSC	General Computer Science Collocation
GSL	General Service List
ICLE	International Corpus of Learner English
IS	Information Systems

KSA	Kingdom of Saudi Arabia
MAWL	Medical Academic World List
MICASE	Michigan Corpus of Academic Spoken English
NS	Native Speaker
NNS	Non-Native Speaker
PICAE	Pearson International Corpus of Academic English
POS	Part Of Speech
RC	Reference Corpus
SAWL	Specific Academic Word List
SCSC	Specific Computer Science Collocation
SE	Software Engineering
UWL	University Word List

Chapter1 Introduction

1.1 Research Background

Corpus linguistics has made a highly influential contribution to descriptions of language use. In the last three decades, it has greatly increased our understanding of grammar, vocabulary, and lexico-grammar in general English as well as in English for Specific Purposes (ESP) (McEnery and Wilson, 1996; McEnery et al., 2006; O'Keeffe and McCarthy, 2010; Boulton et al., 2012). Many studies have pointed out the usefulness of corpus analysis in investigating different features of language (Meunier and Granger, 2008; Lindquist, 2009; O'Keeffe and McCarthy, 2010) and in its applications to language teaching and learning (Gavioli, 2005; O'Keeffe and McCarthy, 2010; Boulton et al., 2012).

The corpus-based approach has become recognised to be a particularly suitable approach for ESP research and teaching for two main reasons. First, the lexico-grammar of ESP discourse, which is distinguished by its "selective use of certain structures, the prevalence of domain-specific, often highly conventionalized, phraseologies (collocations, lexical bundles), and the extremely rapid evolution of ESP terminology and lexis to keep pace with technical and professional developments" (Boulton et al., 2012: 2) can best be investigated by the corpus-based approach. A corpus-based approach is indispensable: "information about the frequency of use of certain structures, and about specialised phraseologies and patterns, can only be obtained from corpora, not from textbooks or grammar books, while traditional dictionaries cannot compete with web-based corpora where lexical and terminological evolution is concerned"(Boulton et al., 2012: 2).

Second, ESP learners, who are usually adult and advanced learners, face difficulty in recognising the specific register conventions of the ESP variety they need to use (Hutchinson and Waters, 1987; Dudley-Evans and St. John, 1998). Genre analysis has become one of the fruitful approaches to ESP and can be investigated by the corpus-based approach, which enables the distinctive features to be reliably identified (e.g., Bhatia, 1993; Swales, 1981, 1990; Paltridge, 2001). Moreover, ESP learners today are encouraged to be autonomous and independent in their learning, thus, the corpus-based approach, which can be incorporated into the learner-centred approach, is considered particularly relevant to promote autonomous and individualised learning (Hutchinson and Waters, 1987; Dudley-Evans and St. John, 1998; Hyland, 2006).

More specifically, corpus linguistics has contributed widely to the research and teaching of the vocabulary of ESP registers (see studies in Boulton et al., 2012). Investigating the phraseology of ESP registers has been one of the main contributions. A number of studies have been conducted to investigate semantic prosody (e.g., Sinclair, 1991; Stubbs, 1995), lexical bundles (e.g., Biber et al., 2004; Cortes, 2004) and collocations (e.g., Gledhill, 2000a; Marco, 2000).

Collocations, which are considered a particular kind of formulaic sequences and prefabricated patterns (Foster, 2001; Howarth, 1998b; Nattinger and DeCarrico, 1992; Wray, 1999, 2000), are viewed as a necessary component of second language (L2) lexical competence. They are therefore considered an important unit for second language learners' improvement in their spoken and written production. Gledhill (2000b: 1) notes, "It is impossible for a writer to be fluent without a thorough knowledge of the phraseology of the particular field he or she is writing in". This is partly because a good deal of the procedural vocabulary of academic

disciplines consists of such predicate structures as *make a claim, reach a conclusion, adopt an approach,* and *set out criteria* (Howarth, 1998a). Conversely, lack of this knowledge may impede the comprehensibility of learners' expression (Laufer and Waldman, 2011: 647-48).

A large and growing body of literature has investigated learners' knowledge and use of collocations. Several corpus-based studies have been conducted to investigate native speakers (NS) and non- native speakers (NNS) learners' use of different types of collocations (Siyanova and Schmitt, 2008; Durrant and Schmitt, 2009; Laufer and Waldman, 2011) in English for Academic Purposes (EAP), but few studies have investigated the use of collocations in English for Specific Purposes (ESP). Gledhill (2000a) investigated the collocational framework of medical research articles to identify the phraseology used in specific sub-disciplines of medical science. Marco (2000) also investigated the grammatical collocations used in a corpus of medical research articles to reveal the most frequent grammatical collocations used in that field.

Although such studies have provided us with some good insights about the use of grammatical collocations in research articles, what is still needed are studies, which investigate learners' use of academic collocations in an ESP context. To my knowledge, no studies have been conducted to investigate non-experts' (NNS and NS learners) and experts' use of academic collocations in ESP and no studies have been conducted to find out the factors underlying over/underuse of academic collocations in ESP specifically in Computer Science (CS). Moreover, one of the main findings from Farooqui's (2010) investigation of the difficulty of academic vocabulary for CS undergraduate students in KSA was their misunderstanding of the concept of collocations as well as their misuse of collocations in their writing. Thus, there is a need for designing materials to raise CS students' awareness of the use of academic collocations in their writing. To fill these

gaps, this thesis aims to investigate the use of academic collocations in discipline specific writing (CS) by non-expert writers (both NNS and NS), to explore the factors underlying the non-experts' over / underuse of such collocations and to develop teaching materials that could be used to help non-expert NNS students with collocation over/underuse problems.

To achieve these aims, this study has three main stages. First, a written corpus of MSc dissertations written by CS students at three UK universities was compiled to locate the most frequent Academic Word List (AWL) (Coxhead, 2000) words and their collocations in this corpus; these were then compared with the reference corpus (RC). Second, the located most frequent academic collocations were further investigated to explore the factors underlying the students' over/ underuse of such collocations. CS experts were interviewed and filled in a categorisation judgement task (CJT). Finally, a sample of awareness-raising activities for NNS learners was designed considering the problematic over/underuse of the academic collocations located in the previous two stages.

1.2 Organisation of the Thesis

The thesis contains not one study but a series of three studies corresponding to the three aims mentioned above: the use of academic collocations in ESP by non-expert NNS and NS students (presented in Chapter 4), patterns of these academic collocations and other factors underlying their over/underuse by non-expert students (presented in Chapter 5), and awareness-raising activities for teaching academic collocations to NNS students (presented in Chapter 6).

An overview of the literature is provided in Chapter 2. It will review the main research related to collocations and corpus–based collocation studies conducted in an EAP and an ESP context. It will also review the specific academic wordlists developed for ESP disciplines (e.g., Chen and Ge, 2007; Martínez et al., 2009), to throw light on the usefulness of Coxhead's (2000) AWL in developing specific academic wordlists for a number of disciplines. Chapter 3 will then present the broad methodology applied in this thesis as well as the three corpora compiled for this study.

The following three main Chapters (4-6) will each report on one of the thesis's three studies, each with its own literature review (except Chapter 4), method, and results sections. Chapter 4 will present the first study in this thesis, which investigates the most frequent academic collocations used by non-expert NNS and NS CS students in their writing of dissertations. It will also investigate the over/under use of the most frequent lexical collocations as compared with expert writers' use. Chapter 5 then presents the second study investigating the factors behind over/underuse of the most frequent lexical collocations. Chapter 6 will present the third study, which aims at designing a sample of corpus-based activities to raise NNS students' awareness of problematic collocations' use and patterns. Finally, Chapter 7 will highlight major findings of this thesis and will conclude with pedagogical implications, limitations, and suggestions for future research.

1.3 Definitions of Key Terms used in the Thesis

1.3.1 Corpora

A corpus is defined by McEnery and Wilson (1996:87) as "a body of text which is carefully sampled to be maximally representative of a language or language variety". This means that any

principled collection of recorded instances of spoken or written language can be compiled as a corpus. With the rapid development of computer technology, a number of different types of corpora have been constructed in the form of machine-readable language data (Biber et al., 1998; Gabrielatos, 2005; Sinclair, 1991). Corpora come in many shapes and sizes to serve different purposes. In general, corpora can be classified into two kinds with reference to size and design. A general corpus (e.g. the British National Corpus or BNC) is designed for general descriptive linguistic purposes, so it is usually much larger than a specialised corpus. In contrast, a specialised corpus is designed from particular types of texts for specific research or teaching purposes (e.g. the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), the Michigan Corpus of Academic Spoken English (MICASE)).

1.3.2 Concordance

A concordance is defined as an "exhaustive list of the occurrences of the word in context" (Biber et al., 1998:15). Concordances are usually presented in the KWIC (Key Word in Context) format where the key word is placed in the middle of each line, for example:

1 …JADE framework is an open ***source*** project distributed by…

2 …is to implement an open ***source*** Real-Time Operating System…

3 …look at a couple of open ***source*** operating systems is followed…

That is, concordances or concordance lines are examples of a word or a phrase with some context on its left and right sides. As Sinclair (1991:170) argues, since the concordance gives access to many important language patterns in texts, it is considered central to corpus linguistics.

1.3.3 Concordancer

A concordancer is the tool most often used in corpus linguistics to investigate corpora. It is a search engine that can be used for retrieving, displaying, counting, and in certain ways analysing language in a corpus. A general concordancer makes it possible to enter a word or phrase and search for and quantify multiple examples of how that word or phrase is used in texts. Typical concordancers generate displays of concordance lines and more sophisticated concordancers can also provide a range of textual information and perform different levels of analysis, relating to frequency information and collocation patterns.

1.3.4 ESP/EAP

A number of researchers (Jordan, 1997; Dudley-Evans and St. John, 1998) define English for Specific Purposes (ESP) as two main strands: English for Occupational/Vocational/Professional Purposes (EOP/EVP/EPP) and English for Academic Purposes (EAP). EOP is the language needed in the real working environment (e.g., a doctor conversation with his nurse) whereas EAP is the language used in academic contexts. EAP is sub-divided into two sub-strands: English for Specific Academic Purposes (ESAP) and English for General Academic Purposes (EGAP) (Blue, 1988a cited in Jordan, 1997). ESAP is the language required for a particular academic subject, e.g. Medicine and CS, while EGAP is the general academic language used in any academic subject. Since this study is focused on academic collocations in subject specific ESP, it is related to ESAP rather than EGAP. In this thesis, ESP is used to refer to ESAP and EAP is used to refer to EGAP.

Chapter 2 Literature Review

2.1 Introduction

This Chapter reviews the relevant literature of the central linguistic phenomena, which forms the central focus of this thesis: collocations. Knowledge of a word's collocation restrictions is one aspect of word knowledge. Moreover, collocations are one type of formulaic sequences. To situate collocations within the broader context of the literature, the Chapter starts with a short review of the nature of vocabulary knowledge (Section 2.2) and formulaic language (Section 2.3). Next, a detailed coverage of collocations and corpus-based studies of their use by both NS and NNS and by experts in EAP and ESP. The Chapter closes with a detailed description of Coxhead's (2000) AWL and its applications and teaching implications in the literature.

2.2 The Nature of Vocabulary Knowledge

Different language acquisition researchers have focused on different dimensions of vocabulary knowledge. First, Anderson and Freebody (1981:92-93) distinguished two main dimensions of vocabulary knowledge: breadth (size or quantity of vocabulary) and depth (quality or what kinds of information are known about the vocabulary). In addition to breadth and depth, Meara (1984) focused on a third dimension, lexical organisation, which concerns how vocabulary knowledge is interconnected in a person's mind. Vocabulary has also been categorised as receptive/passive (related to word recognition) or productive/active (which is related to word production) (Nation,

2001). Fourthly, Segalowitz and his colleagues (e.g., Segalowitz and Segalowitz, 1993) identified automaticity as an important dimension of vocabulary knowledge stressing its central role in vocabulary development and use.

The depth dimension is particularly important for my purposes as it represents an important step away from the traditional view of vocabulary knowledge as consisting simply of form-meaning mappings. There are two main approaches to defining vocabulary depth: the *developmental approach* (Wesche and Paribakht, 1996), based on the assumption that words undergo a number of stages in a learner's mind, from zero knowledge to full mastery, and the *dimensions approach* (Nation, 2001),which details the various aspects of word knowledge.

The dimensions approach has attracted more attention than the first; during the last twenty years, several attempts were made by different vocabulary acquisition researchers (Carter, 1987; Laufer, 1997; McCarthy, 1990) to analyse vocabulary knowledge into different components. The most comprehensive of these was Nation's (1990) framework, which was revised 10 years later (Nation, 2001; Table 2-1). As can be seen in Table 2-1, various aspects of word knowledge are divided into three major categories: form, meaning, and use. Under each category, features can be either related to receptive mastery (R) or productive mastery (P).

Table 2-1: What is involved in knowing a word? (Nation, 2001: 27)(*Note*: in column 3, R = receptive, P = productive).

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?

	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express this meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use (Register, frequency…)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

In the light of previous discussion about receptive and productive distinction of these aspects, Nation (2001) has pointed out that receptive knowledge is easier for learners to acquire than productive. Learners tend to need to know only few distinctive features for a word to be understood while they need more knowledge of a word to produce it. Moreover, concerning the developmental approach, the learning burden of vocabulary increases if learners are exposed to all aspects of knowing the word together (Laufer, 1997). In fact, at any given time, learners tend to have mastered some of these aspects, but not all (Nation, 1990). For example, collocations are not typically the first aspect of a word to be learned.

2.3 Formulaicity: Pervasiveness and Significance

In recent years, formulaicity has emerged as a promising area of research. Corpus linguistic data revealed formulaicity as a pervasive phenomenon in language (Foster, 2001; Howarth, 1998a, 1998b; Wray, 2002). Bolinger (1976) was arguably the first to highlight that speakers use a large number of memorised 'prefabs' (i.e. formulae). Similarly, Pawley and Syder (1983) stressed that sounding native is not only related to knowledge of grammatical rules (e.g. Chomsky's (1965) generative grammar) but also entails knowledge of which sequences that follow the rules are also acceptable. Thus, Sinclair (1991) proposed two principles to explain how meaning is conveyed in texts: the *open-choice principle* (creative language use based on complex choices within grammatical rules) and the *idiom principle* (fixed expressions based on co-occurrence restrictions). These principles will be explained in more detail in section 5.2.1.

Likewise, Moon (1997:41) contrasted the traditional, syntactic model of language that monitors well-formedness, with what she called the "collocationist model" observing co-occurrence restrictions. It is well accepted now that in order to reach native-like fluency, learners need to be in control of formulaic sequences in the second language as well as grammatical rules. As Ellis put it: "speaking natively is speaking idiomatically using frequent and familiar collocations" (1997:129). Adequate use of formulaic word strings (e.g. collocations) has been shown to help L2 learners present as proficient in speech (Boers, Eyckmans, Kappel, Stengers, and Demecheleer, 2006) and in writing (Dai and Ding, 2010).

According to Howarth (1998b), compliance with collocational restrictions is not only a matter of stylistic elegance; rather, it is essential for effective communication and problems in using collocations can lead to serious communication problems. Lacking the appropriate native-like knowledge of formulaic sequences might make the learner come across as arrogant or disrespectful (Moon, 1997; Wray, 2002).

Indeed, from the above, I can see that formulaic sequences have two main functions in language: saving processing effort (so enhancing fluency) when they are memorised as wholes, and achieving interactional, communicational effectiveness or acceptability (Wray, 2000). Formulaic sequences, if stored and retrieved as wholes, help users to process language more efficiently compared to processing on a word-by-word basis. Schmitt and Carter (2004) claim that it is this processing advantage of formulaic sequences that can account for why they are used to realise as many interactional functions as they do.

But what are formulaic sequences? Simply put, any string of words that is processed as a holistic unit without any recourse to its constituent parts can be described as formulaic (Wray, 2002). However, things are not that straightforward, mainly because formulaicity is a phenomenon that can take so many forms (Schmitt and Carter, 2004). There has been disagreement not only on defining formulaic sequences but also on coining terms to refer to them (Wray, 2008, 2009). Wray (2000, 2002) listed fifty different terms used in the literature to refer to the phenomenon, such as *chunks, collocations, frozen metaphors, idioms, lexicalised sentence stems, multiword items/units, ready-made expressions, binominals (compound nouns)* and *routine formulae*. Finally, she presented her definition of what she calls a 'formulaic sequence':

> A sequence, continuous or discontinuous, of words or other 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, 2002: 9).

This definition seems to be based on Moon's (1997) three gradable, non-absolute criteria of formulaic language that distinguishes them from other strings. These are institutionalisation (holistic status in a language), fixedness (sequence frozenness), and non-compositionality (unitary, non-analysable, meaning). Howarth (1998a, 1998b) maintained that the different categories of formulaic sequences (mentioned above) have different degrees of 'institutionalisation', 'fixedness', and 'non-compositionality'. Thus, it is misleading to treat them all as belonging to one single class.

Collocations, which are considered a category of formulaic language, have been categorised in a scale of ten criteria (Nation, 2001) including Moon's (1997) three criteria. Collocations can be memorised as holistic units or as chunks (Sinclair, 1991). Some collocations are fixed

‘'unchangeable’' while others are flexible ‘'allowing substitution in all or one part’' (Handl, 2009). Some collocations have only one meaning while others have several meanings with ‘'related meaning’' as the mid-point (Nation, 2001).

The focus of the present thesis is on one category of formulaic language, that is, collocations. The next section will present collocations in detail, regarding their definition, approaches to identifying collocations, difficulties of collocations for L2 learners, and corpus-based studies related to collocation use in both EAP and ESP context.

2.4 Collocations

2.4.1 Definition of Collocation

Collocations have been defined variously throughout the literature, and not usually by unity of storage and retrieval. This is evident in the different conceptualisations of the term by different scholars:

1- “The relationship a lexical item has with items that appear with greater than random probability in its (textual) context” (Hoey, 1991:7).

2- “The occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991: 170).

3- “A composite unit which permits the substitutability of items for at least one of its constituent elements (the sense of the other element, or elements, remaining constant)” (Cowie, 1981: 224).

4- “Fixed, identifiable, non-idiomatic phrases and constructions” (Benson, Benson, and Ilson, 1997: xv).

These various definitions represent the two approaches to defining collocations (see Firth, 1986; Cowie, 1981; Mel’çuk, 1998; Nesselhauf, 2003, 2005; Barfield and Gyllstad, 2009): frequency-based (Definitions 1 and 2) and phraseological (Definitions 3 and 4). Various terms have been used in the literature to refer to this contrast: quantitative versus qualitative (Bartsch, 2004) and empirical versus theoretical approaches (Evert, 2008), respectively. The first approach is related to how often the words are seen together in a corpus and uses statistical measures of strength of association as a criterion for collocation extraction (regardless of the phraseological criteria specified above: ‘institutionalisation’, ‘fixedness’, and ‘non-compositionality’).

Conversely, the second approach treats collocations as sequences of words that meet certain phraseological criteria to some degree, and uses native speakers’ intuitions as a validation of their status. The next two sub-sections will describe each approach in detail with an account of how collocations are operationalised under each. The final sub-section will explain how the term ‘collocation’ will be used in the present thesis combining both approaches.

2.4.1.1 The Frequency-based Approach

The frequency-based approach is related to the statistical definition of collocation in which “words are collocates if, in a given sample of language, they are found together more often than their individual frequencies would predict” (Jones and Sinclair 1974: 19). Words that stand in such a relationship can be said to ‘predict’ one another because the presence of one makes the presence of the other more likely than it would otherwise be (Sinclair, 1966: 417-418). It was

first originated in Firth's slogan "You shall know a word by the company it keeps!" (Firth, 1957/1968: 179). It has been applied in corpus linguistics to locate collocations (e.g., Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008). Thus, collocations have been distinguished by their frequency into frequent and infrequent collocations.

In fact, looking at Sinclair's definition above (Definition 2), it might be claimed that, in this approach, any occurrence of a pair in a corpus might be considered a collocation. He (Sinclair, 1966:418), however, distinguished between 'casual', accidental collocations with only very few occurrences and 'significant', typical collocations manifested by above-chance frequency. Similarly, Stubbs (1995) observed that frequency of co-occurrence is not enough in identifying collocations and hence the additional need for measures of association strength (e.g., MI and *t*-score). These measures will be discussed in detail in section 4.2.3.

Halliday (1966) introduced *'node'*, '*collocate*', and '*span*' as three fundamental terms in the operationalisation of the frequency-based approach of collocation. The word under investigation is '*the node'*, the co-occurring word is '*the collocate'*, and the specified environment in which the node and the collocate may co-occur is '*the span'* (Halliday, 1966: 156). Jones and Sinclair (1974) expanded the investigation of probability of co-occurrence of collocates to a span of +/- 4 words from the node. That is, four words to the left and to the right of the node are considered the optimal environment in which 95% of that node's collocational influence occurs.

A recent extension to the frequency-based approach to collocation is termed 'lexical bundle analysis' in which researchers instruct the computer to search for identical occurrences of n-word sequences (e.g., 2-word,3-word,4-word bundles) in a specific registers(Biber and Conrad, 1999;

Biber and Barbieri, 2007). Hyland (2008) refers to lexical bundles as 'extended collocations' since they are seen as having pre-fabricated or formulaic status.

On the other hand, it should be noted that compounds (which consists of two words, written as one or more words, or joined by a hyphen. e.g., *travel agent, dark-haired and bathroom*) are different from collocations. Compounds are "much less compositional, since meanings of the bases alone is not sufficient to predict the meaning of the compound as a whole" (Shopen, 1985: 22). In addition, compounds do not allow variation of the word order. That is, *travel agent* cannot be changed to *agent travel*. It will be meaningless. On the other hand, collocations allow for variation of the word order, such as *source code* can also be *code source*, this seem in line with Sinclair's (1991) idiom principle feature in which he confirms that some phrases allow for variation of word order (see section 5.2.1 for detailed information about Sinclair's idiom principle features).

2.4.1.2 The Phraseological Approach

Unlike the previous statistical approach, which employs corpus frequency as an identification criterion, the phraseological approach uses either native speakers' intuitions (Greenbaum, 1988; Hasselgren, 1994), collocational dictionaries (Laufer and Waldman, 2011), or a combination of both (Nesselhauf, 2003) in identifying collocations. Collocations are identified in "a scalar analysis, ranging in the form of a continuum from transparent, freely recombinable collocations at one end to unmotivated and formally invariable idioms at the other" (Barfield and Gyllstad, 2009: 6).

Nesselhauf (2005:21) specified three main linguistic criteria according to which these categories have been identified in the literature: syntactic characteristics (constituents' part of speech),

semantic characteristics (sense restrictions), and commutability of elements (substitution of one or both elements). Based on the final two criteria, she (Nesselhauf, 2003:226) distinguished between three types of combinations:

Free combinations (e.g. *want a car*)

The senses in which the verb and the noun are used are both unrestricted, so they can be freely combined according to these senses, and freely substituted (e.g. *want a drink, buy a car*).

Collocations (e.g. *take a picture*)

The sense in which the noun is used is unrestricted, but the sense of the verb is restricted, so that the verb in the sense in which it is used can only be combined with certain nouns (*take a picture/a photograph*, but e.g. * *take a film/movie*).

Idioms (e.g. *sweeten the pill*)

Both the verb and the noun are used in a restricted sense, so substitution is either not possible at all or only possible to an extremely limited degree.

The phraseological approach has proven that collocations are in most cases lexically variable and characterised by arbitrary limitations of one or more features. However, Stubbs (1995) pointed out limitations of the phraseological approach claiming that natives' intuitions, while interesting, are not a reliable source of evidence on collocational restrictions (natives can give some examples of collocations but cannot give accurate frequency estimates). This highlights the need to combine both approaches when identifying collocations.

2.4.1.3 Collocations in the Present Thesis: A Complementary Approach

The two approaches to defining collocations outlined above are not in opposition but should rather be viewed as complementary. As Nation (2001: 317) suggested, from the perspective of language learning, collocations should be considered as "items which frequently occur together and have some degree of semantic unpredictability". Evert (2008) also stressed the close connection between the two approaches. Many collocations identified through corpus analysis have phraseological significance. Conversely, many collocations that have phraseological significance will stand out in corpus analysis. The approach taken in the present thesis is for the combination of both approaches. The term 'collocation' is operationalised here as:

A non-idiomatic pair comprising two open class lemmas which occurs in a corpus (within a window of ±3) above chance (*f* >5, MI >3 and *t*-score>2) and which exhibits specific usage restrictions.

This definition employs both statistical and phraseological criteria. On the statistical side, the following criteria are applied:

1- Collocations are two-word combinations (combinations with more than two open class words are not considered in this thesis).

2- The unit of analysis is not the word form but rather the lemma. Lemmas were used so that all possible forms of a given collocation (e.g., *heavy rain, heavier rain,* and *heaviest rain*) can be included together in the frequency count (see Fitschen and Gupta, 2008 for more on benefits of lemmatisation when extracting collocations).

3-Only lexical collocations, with two open class words (as opposed to grammatical collocations), are considered here.

4- The collocation span is set to three words to the left and right of the node word.

5- The pair should have a minimum of five occurrences in the RC to be considered a collocation.

6- MI is used as a measure of the strength of association with a minimum score of 3.The *t*-score is used as a measure of the significance of collocations with a minimum score of 2.

The above criteria are combined with two other phraseological ones (which will be applied in the identification of collocational patterns in Chapter 5):

1- The pair has a transparent, non-idiomatic meaning that is clearly deducible from the senses of the individual words.

2- Two dictionaries of collocations are used in checking whether the collocations located are specific CS terms.

Thus, the frequency-based approach was applied first in locating significant collocations in the corpora (presented in Chapter 4) and then combined with the phraseological approach to investigate the factors that underlie over/underuse of the shared set of collocations among corpora (presented in Chapter 5).

2.4.2 The Importance of Collocation in L2 Learning

Collocations, viewed as types of formulaic sequences (see section 2.3 above), are seen as a necessary component of second language (L2) lexical competence. They are considered as

especially important for second language learners to acquire in order to improve their spoken and written language.

Collocational knowledge, which is stored in chunks (Wray, 2002, 2009), is held by some to be the foundation of language learning, use, and knowledge. Three major types of evidence support this claim. First, the intuitive feeling of NS that certain phrase seem to act as units (Pawley and Syder, 1983; Nattinger and DeCarrico, 1992). Second, collocation corpus-based studies have proved the frequent co-occurrence of certain groups of words (e.g., Durrant and Schmitt, 2009). Third, studies of learning and knowledge show that "language users make use of unanalysed collocations, and that analysed collocations are used with greater speed than would be possible if they were recreated each time they were used, and that there are errors which demonstrate that collocations are being used as lexicalised units" (Nation, 2001:335).

Collocation has been considered an important element to be fluent and proficient as native speakers(Pawley and Syder, 1983; Boers et al., 2006) and to distinguish between advanced L2 learners and intermediate ones(Thornbury, 2002).Thus, for L2 learners, collocations have come to be considered the gateway to higher levels of English and gaining native-like competence(Henriksen and Stæhr, 2009).

Collocation also plays an important role in taking on or rejecting a group identity (Wray, 2002). This is clearly confirmed in academic writing, where a writer from a particular discipline such as CS needs to demonstrate his knowledge of the [collocations] used in the field he or she is writing in to be a member of that field(Gledhill, 2000a). Thus, collocation not only plays an important role in the development of L2 learners' fluency and native-like competence but also constitutes a

vital means for the writer to become an 'insider' (Durrant and Aydınlı, 2011) in a specific group of users of an academic community.

2.4.3 Difficulties of Collocations for L2 Learners

Findings of previous studies on collocation have proved that collocation is problematic and difficult for L2 learners. Its difficulty is related to the correct use of native-like collocation rather than to its comprehension (recognition) (Nesselhauf, 2003; Bahns and Eldaw, 1993). NNS learners strive for achieving native-like production by trying to be idiomatic in their production of language. In order to achieve this native-like production, they use different strategies. They tend to rely on creativity and make "over liberal assumptions about the collocational equivalence of semantically similar items" (Wray, 2002: 201). That is, they assume that synonymous words such as *surgery* and *operation* have similar collocations. If the word *surgery* collocates with *plastic*, then it can collocate with *operation*. Thus, an atypical collocation is produced (Siyanova and Schmitt, 2008: 430). Moreover, they tend to use grammatical sentences that are not used by NS. This results in producing unconventional combinations of words (Pawley and Syder, 1983).Therefore, Skehan (1998) and Foster (2001) propose that NNS construct their language from rules rather than from lexicalised routines.

Comparing NS and NNS use of collocation reveals that NNS use collocation in their writing but not to the same extent as NS do (Foster, 2001; Granger, 1998; Howarth, 1998a; Durrant and Schmitt, 2009). Researchers have found that NNS rely on a limited set of collocations in their productive language use. In some cases, they tend to overuse a certain set of collocations. For example, collocations constructed with core verbs (*be, have, make,* etc.) or particular amplifiers (*very, completely, highly, strongly*), whereas they do not use other native-like collocations

(Granger, 1998). An interesting explanation has been provided by Siyanova and Schmitt (2008), who found that NS and NNS have different levels of familiarity with adjective noun collocations. The NNS rated the infrequent collocations as more familiar than NS did and they rated frequent collocations as less familiar than NS did.

Cobb (2003) commented that the overuse of a small set of collocations makes learners sound odd. Other researchers (De Cock et al., 1998; Foster, 2001; Granger, 1998; Kaszubski, 2000; Nesselhauf, 2005) pointed out the overuse of this collocation set may indicate that these collocations are cognate with their L1. In contrast, if L2 collocations are incongruous with L1 collocations, a negative L1 transfer for these L1-incongruous collocations will be produced (Wolter and Gyllstad, 2013).The difficulty of NNS using collocations appropriately is not restricted to beginners; even advanced learners face this difficulty.

McCarthy (1990:13) noticed, "Even very advanced learners often make inappropriate or unacceptable collocations". An exception to this finding is Siyanova and Schmitt (2008) who found that their NNS advanced learners of English produced as many adjective-noun collocations as native speakers. However, 25% of their produced collocations were considered atypical since they do not appear in the BNC (For more details about this study, see section 2.4.5).

Other collocation studies have shown that advanced L2 learners not only overuse certain collocation phrases and underuse others, but also make numerous collocation errors (Altenberg and Granger, 2001; Hasselgren, 1994; Nesselhauf, 2003, 2005). An explanation provided by Hasselgren (1994) is that the infelicitous collocations result from overdependence on the familiar

ones, that is, structures that learners learned early, used widely, and with which they felt comfortable.

Nesselhauf (2005) found from her investigation of learners' writing development that the number of collocational errors was not different when they wrote with or without the use of a dictionary. This may suggest that either the dictionary did not provide the necessary information about the use of collocations or that learner do not seek it since they are not aware of its importance. Another important factor that has been investigated is time pressure. She found that writing with or without time pressure had no great effect on the use of collocations on learners' writing. This suggests that learners' use of collocations demonstrates a lack of knowledge rather than a lack of control.

2.4.4 Approaches to Collocation Research and Identification

Two main approaches to collocation research are found in the current literature: the experimental approach and the corpus-based approach.

The experimental approach entails the use of experiments and tests with focus on the processing and acquisition of collocations. Thus, time constraints are considered a vital factor in this approach and the results reflect cognitive processes involved in listening or reading collocations (e.g., Siyanova and Schmitt, 2008; Durrant and Schmitt, 2009). In this approach, collocations are identified for inclusion in the tests or other stimuli used in the study by the researcher using criteria relevant to what is being studied. That might include phraseological criteria (e.g. Gyllstad, 2007), and/or criteria based on frequency of the items elsewhere (e.g., Siyanova and Schmitt, 2008, study 2).

The corpus-based approach, on the other hand, focuses on studying collocations used in learners' spoken and written language by generating frequency lists and concordances. The results obtained do not reflect the immediate cognitive processes involved as no time constraints are involved. In this approach the collocations come initially from the participants (learners or NS) rather than the researcher, but the researcher still makes the decisions regarding what to count as a collocation, which again may involve phraseological criteria as well as frequency criteria (both in our study: see section 2.4.1.3).

Numbers of researchers have applied both approaches in their investigation of EAP learners' use: psychological processing and automatisation of collocations in learners' cognition (e.g., Siyanova and Schmitt, 2008; Durrant and Schmitt, 2009). The corpus-based approach is typically applied first in the identification of collocations in both NNS and NS corpora and then the set of collocations located is used further in experiments to test learners' processing and automatisation of these collocations.

Other researchers have applied only corpus-based research to identify and study different categories of collocations in parallel NNS and NS corpora (e.g. Nesselhauf, 2003; Granger, 1998) or in a single discipline-specific corpus(e.g., Gledhill, 2000a, 2000b; Ward, 2007). In this approach, collocations have been identified by applying frequency criteria, i.e. statistical association measures (*t*-score and MI) to identify which pairs were strong collocations(these measures are discussed in detail in Chapter 4). For example, Durrant and Schmitt (2009) conducted a study to investigate the use of adjective noun collocations in NNS and NS academic

writing. They adopted the frequency approach in their identification of collocations since it is considered the most reliable approach.

"The particular strength of computerised corpora is that they offer the researcher the potential to check whether something observed in everyday language is a one-off occurrence or a feature that is widespread across a broad sample of speakers"(McCarthy, 1998:151). It has been noted that quantitative techniques are essential for corpus-based studies investigating actual patterns in language use. That is, it is necessary to make quantitative measurements to gain insights into patterns of language and their use because typical patterns tend to occur more frequently.

However, quantitative and statistical techniques applied to corpora need to be combined with qualitative methods to provide full explanations of language use and prevalent patterns. This is especially true in the case of ESP collocation studies (Gledhill, 2000a; Marco, 2000). Both quantitative and qualitative methods are used as a complementary approach, with the use of concordance lines or statistical measures combined with context and genre-sensitive approaches. As a result, genre norms and specific discourse features may be clearly highlighted in the ESP context (Boulton et al., 2012).

In the current thesis, the corpus-based approach will be employed since the main aim of the study concerns NNS and NS learners' collocation use in their writing of MSc dissertations, with no concern for the cognitive process of acquiring or processing the collocations. Thus, only the relevant corpus-based collocation studies will be reviewed in detail in the following section.

2.4.5 Previous Corpus-based Collocation Studies

In the following sections, studies investigating learners' collocation use in their writing will be reviewed in detail since the primary focus of the current thesis is on investigating collocation use in learners' written data rather than in their speech (section 2.4.5.2). The most current studies will be reviewed to ascertain the methods applied and to determine which types of collocations have been investigated and what was found. Next, collocation studies in ESP will also be reviewed in detail (section 2.4.5.3); finally, academic collocation studies across disciplines will be reviewed (section 2.4.5.4).

2.4.5.1 Criteria for Collocation Identification in Corpus-based Studies

Various criteria have been used by researchers to identify collocations on some scale of strength in corpus-based studies. Kjellmer (1984 cited in Nation, 2001) uses six criteria to measure distinctiveness, or degree of lexicalisation, of collocations: absolute frequency, relative frequency, length of sequence (number of collocates in collocation), distribution over texts (range), distribution over text categories, and structural complexity.

The most obvious scale is the frequency of co-occurrence, which ranges from 'frequently occurring together' to 'infrequently occurring together' items. It is considered an important criterion, is measured by counting, and can be expressed in absolute or relative terms. "The absolute frequency refers to the actual number of times a collocation occurs in a corpus, while the relative frequency compares actual frequency of occurrence with an expected number of occurrences"(Kjellmer, 1984: 166-168 cited in Nation, 2001:329). A number of researchers have applied these criteria in their identification of collocations in their corpora (e.g., Durrant and

Schmitt, 2009; Siyanova and Schmitt, 2008). Both criteria are used in this thesis as well (for more details see Chapter 4).

Adjacency of members of a collocation is another criterion for classifying collocations: this ranges from 'next to each other' to 'separated by several items' (Nation, 2001). Collocates can be located in a span of four to five words to the right and to the left of the node words. The space to the left and to the right of the node words included in the search is called the 'window' (Lindquist, 2009:73). Sinclair et al. (2004: xxvii) pointed out that "the wider the span, the lower is the significance in general". Thus, a span of three words to the right and to the left of the node words was used in locating collocates in this thesis.

On the other hand, Shin and Nation (2008) applied six different criteria from Kjellmer in their identification of the 1,000 most frequent spoken collocations in their corpus of 10 million words. The first criterion was related to the node words that were counted as word types rather than word families (e.g. *take, takes* and *taken* are word types of the word family *take*).They claimed that "different types of the same word family have different collocates" (2008:341).Another criterion (2) focused on locating lexical collocates of the node word, that is, the collocate should be a content word like nouns, adjectives, verbs, and adverbs. The different senses of the same word type were counted separately as another criterion (6).

Two other criteria (3 and 4) were concerned with the frequency of the occurrence of collocations. Collocations had to occur at least thirty times in their 10 million words corpus and had to be in the 1,000 most frequent content words of English according to the spoken word frequency list by Leech, Rayson, and Wilson (2001). The criterion (5) 'grammatical well-formedness' related to "the ability of collocations to stand as a comprehensible unit often as a

part of a sentence" (Shin and Nation, 2008:341).Three criteria were applied in this thesis: criterion 1 and 2 were applied in locating academic collocations in the first study reported in Chapter4. Criterion 5 was applied in checking the erroneous located collocations, which is reported in detail in Chapter 5.

Nation (2001: 329-332) employed eight criteria other than frequency and adjacency for classifying collocations. Four of these criteria focused on grammatical issues: how far the collocations were grammatically connected, grammatically structured, or exhibited grammatical uniqueness, and grammatical fossilisation. The first two criteria are related to the structure and connection of collocations within the same sentence. The grammatical uniqueness of some collocations is related to some collocations that are grammatically unique e.g. *hell for leather*, while other collocations follow regular patterns. The grammatical fossilisation refers to some collocations, which do not allow any change in word order or in part of speech e.g. *kick the bucket*. This criterion applies to idioms rather than collocations and it is related to another criterion 'semantic opaqueness' when the meaning of idioms cannot be deduced from its parts. Hence, it is not relevant to collocation identification.

The other criteria are related to lexical issues: lexical fossilisation, uniqueness of meaning, and collocational specialisation. Lexical fossilisation concerns some collocations that are unchangeable, e.g. *a bird's eye view,* and some collocations that contain words that can be replaced by other words of related meanings, e.g. *entertain a belief, entertain a desire*. Uniqueness of meaning is related to the meanings of collocations: some collocations have only one meaning e.g. *keep a secret*, while others have several meanings, e.g. *kick the bucket.* The collocational specialisation criterion refers to some collocations whose component words are

fixed and never or rarely occur without each other e.g. *hocus pocus.* None of these criteria has been applied in the current thesis.

Other researchers have categorised collocations into two main types: grammatical collocations and lexical collocations (Benson et al., 1997). Grammatical collocation refers to "a phrase consisting of a dominant word (noun, adjective, and verb) and a preposition or grammatical structure such as an infinitive or clause" (Benson et al., 1997, p.xv as cited in Barnbrook et al., 2013). In contrast, lexical collocations "do not contain prepositions, infinitives or clauses; [they] consist of nouns, verbs, adjectives and adverbs" (Benson et al., 1997:p.xxx as cited in Barnbrook et al., 2013).

A great number of collocation studies have been conducted to investigate lexical collocations' use in both NNS and NS writing in EAP (e.g., Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008), while few studies have investigated grammatical collocation use in the writing of experts in ESP (Gledhill, 2000a; Ward, 2007). To my knowledge, no studies have been conducted to investigate lexical collocations' use in the writing of NNS and NS in ESP. Since the focus of the current study is on comparing both NNS and NS CS postgraduate students' use of lexical collocations with CS experts, the following section will review corpus-based studies on lexical collocations only.

2.4.5.2 L2 Learners' Use of Collocation in EAP

A number of studies on lexical collocation use have been conducted throughout the literature (Chi et al., 1994; Granger, 1998; Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008). Most

of the studies compared NS and NNS lexical collocation use in their written production, as this current study also does.

Chi et al. (1994) conducted their study to investigate the inappropriate collocation use of delexical verbs in a corpus of 1 million words compiled from first year university students' writing. Using *Microconcord* (Scott and Johns, 1993), the concordance lines of the selected five delexical verbs (*do, have, take, make* and *get*) were extracted to locate the faulty occurrences. Then, they were double-checked with BBI and three other collocation dictionaries.

In case of ambiguous occurrences, two other procedures were applied. First, wider context of the concordance verbs was considered. Second, NS were asked to verify whether ambiguous items were erroneous collocations or not. Although, native speakers' evaluation can help in deciding which collocations are typical and which are not, it might not provide quite accurate results since they sometimes have different views. One main criticism to this study is that they do not compare their learners' corpus with a RC.

The results showed that there were two main reasons for delexicalised verb-noun collocation errors. First, learners may confuse the delexicalised verbs. For example, they may replace the verb *make* with *have* to collocate with *progress*. Second, they may confuse the use of these delexicalised verbs with other verbs. For example, they use the noun *interview* with *take* instead of *make*. L1 interference was thought to be the main factor behind these errors. Although this study is distant from my study in many ways, its use of dictionaries and NS (in our case, CS experts) to check collocations was a feature I adopted. However, it did not use a RC, which I consider valuable.

Granger (1998) extracted amplifier-adjective collocations from the French ICLE (International Corpus of Learner English) sub-corpus and a small native corpus to explore the collocational behaviour of French EFL learners in comparison to natives. French learners were found to use fewer amplifiers with adjectives than natives. Among these, French learners also used two main amplifiers (*completely* and *totally*) as "safe-bets" (148). Granger's study was an important first step in collocational corpus-based research but was extremely limited in that the collocations extracted were not matched against external norms (e.g., collocational dictionaries or natives' intuitions). Hence, in our study I will take care to use all these checks.

Durrant and Schmitt (2009) investigated the high frequency adjective-noun collocations used in both NS and NNS academic writing. A corpus of approximately75, 000 words was compiled. A total of 96 texts from both NS and NNS writing were collected consisting of both long and short texts. These texts were divided equally into 24 texts of each type. The long NNS corpus consisted of pre-sessional projects and undergraduate argumentative essays, while the NS corpus consisted of MA assignments from the Applied Linguistics department and from *The Prospect* magazine. The short NNS corpus consisted of pre-sessional short essays and a segment of ICLE whereas the NS short corpus consisted of essays from LOCNESS and opinion articles from *The Guardian* and *The Observer*.

Even though their corpus consisted of both NNS and NS writing compiled from texts in different genres, no concern was devoted to distinguish between experts' and non-experts' writing. Their NS corpus consisted of a mixture of expert writing (newspapers and magazines) and non-expert writing (students' writing in LOCNESS or MAs). This issue was considered in compiling our NNS and NS corpora as they were both compiled from MSc dissertations written by non-experts.

To identify adjective-noun collocations in their corpus, only adjacent pairs were extracted first manually. Only direct adjacent pairs were used since admitting collocations at a wider range of distances ran the risk of making association measures non-comparable between collocations. Proper nouns, acronyms, pronouns, possessives, semi-determiners and numbers/ordinals were all excluded. Quotations were also excluded since they are not considered part of writers' performance. A total of 10,839 word collocations from the 96 texts were retrieved.

The collocations identified were first filtered in terms of their frequency in the BNC, the largest RC available for general British English. Then, strength of association measures (MI and *t*-score) were applied. It has been suggested that a *t*-score ≥2 and a MI score ≥3 may be indicative of a significant collocation (Hunston, 2002a; Stubbs, 1995), but at this stage they used these measures to grade collocations rather than to divide them into collocates vs. non-collocates. Thus, a scale of seven bands of *t*-score and a scale of eight bands of MI were applied.

Moreover, they recorded results individually for each text and compared the four groups of texts using standard inferential statistics, taking each text as an individual case. The results showed that NS writers used more low-frequency collocations whereas NNS writers used more high-frequency collocations. Interestingly, both NNS and NS used collocations with very high *t*-scores similarly. On the other hand, NNS significantly underused collocations with high MI scores in comparison to NS. This study is valuable since its applied method is quite clear in identifying collocations by their frequency compared to the BNC that was used as a RC and in ranking collocations using scales of both *t*-score and MI. However, their approach of comparing both students' corpora as wholes and individuals is somewhat different from the well-known approaches of collocation identifications.

Similarly, Siyanova and Schmitt (2008) investigated the use of adjacent adjective-noun collocations in the writing of Russian L1 learners of English. A corpus of 31 essays written by Russian university students, selected from the ICLE consisting of 25,000 words was compiled.. A comparable corpus of NS writing was selected from LOCNESS. They first extracted adjacent adjective-noun collocations from both corpora manually. This procedure retrieved 810 adjective-noun collocations from the NNS corpus and 806 adjacent adjective-noun collocations from the NS corpus. To determine the frequency of both NS and NNS collocations as well as their MI, the BNC was consulted as a RC. Using the BNC frequency information, the collocations were split into five frequency bands: 0 (failed to appear in the BNC), 1–5, 6–20, 21–100, and>100 occurrences.

As a result, half of the learners' collocations occurred in the BNC and thus considered as native-like collocations in their uses, one quarter of learners' collocations did not occur in the BNC at all, and another quarter of these collocations were less frequent in the BNC. Thus, around half of the learners' collocations were either atypical or, at least, infrequent in the BNC. Around 45% of the collocations met the native-like threshold of frequency ≥ 6 and MI threshold ≥3.

Unlike Durrant and Schmitt's (2009) use of association measures in grading collocations, Siyanova and Schmitt (2008) use only MI ≥3 and frequency≥6 as indication of strong collocations. This study is valuable since it highlights the importance of having a comparable NS and NNS corpus in both genre and size to have valid results. However, their results contradict previous studies' findings, by suggesting that NNS can master the use of adjective-noun collocations in their writing to resemble the NS students' use.

Unlike Durrant and Schmitt (2009) and Siyanova and Schmitt(2008) who investigated adjective-noun collocation use in NS and NNS academic corpora, Laufer and Waldman (2011) investigated the use and errors of verb-noun collocations by L2 Israeli learners from three proficiency levels: basic, intermediate, and advanced. Their NNS corpus was compiled from 759 assignments consisting of 324,304 words, collected from schools and universities while the NS student corpus was represented by LOCNESS.

In their analysis, they began by locating the most frequent nouns in the NS corpus to be used as the baseline. They selected the most frequent nouns that occur 20 times or more in the NS corpus. These nouns were further investigated for their verb collocations. After they had listed all verb-noun collocations from the NS corpus, they checked these collocations in two dictionaries: *The BBI Dictionary of English Word Combinations* (Benson et al., 1997) and *The LTP Dictionary of Selected Collocations* (Hill and Morgan, 1997). If the verb-noun combination was listed as a collocation in either one of the dictionaries, it was noted as a collocation. A similar procedure of verification of collocations was used by Nesselhauf (2005), and Wang and Shaw (2008). In all, 2,527 verb-noun collocations were extracted from the NS corpus.

The 220 most frequent nouns identified in NS corpus were then located in the learner corpus. Then, their verb collocates were extracted. The well-formed collocations were verified in collocation dictionaries, which results in 1,082 verb-noun collocations in the learner corpus. After that, these well-formed verb-noun collocations were checked separately in each of the sub-corpora of various learner proficiency levels. Advanced learners produced 852collocations altogether involving 13,805 noun tokens, intermediate learners produced 162 collocations using 3,057 noun tokens, and the basic learners produced 68 collocations with 553 noun tokens.

The results confirmed the previous finding that NNS produced fewer collocations in their writing than NS do. The descriptive data shows that in each of three learner groups, there were fewer collocations than in the NS corpus: 4.3% in the basic sub-corpus, 5.3% in the intermediate sub-corpus, and 6.2% in the advanced sub-corpus, as opposed to 10.2% verb-noun collocations in the NS corpus. These results are in agreement with the findings of the aforementioned studies of collocation use, with the exception of Siyanova and Schmitt (2008).Even though this study is distinct from my study in a number of ways, its use of collocation dictionaries to verify the existence of their located verb noun collocations was adopted with different purpose.

It can be concluded from the previous studies that NNS learners find difficulty in their use of different types of lexical collocations in an EAP context, with except of Siyanova and Schmitt (2008) who found that their advanced NNS learners use adjective-noun collocations as native-like learners.

2.4.5.3 Expert Writers' Use of Collocations in ESP

Collocation has been proved essential language knowledge not only for EAP learners but also for ESP learners. Based on Hyland and Tse's (2007) attention to discipline-specific collocations, a number of studies have been conducted to investigate either the grammatical or the lexical collocations in ESP contexts (Gledhill, 2000a; Marco, 2000; Ward, 2007; Williams, 1998; Yang, 1986). However, most of these studies focused on locating collocations in expert writing. To my knowledge, no previous studies have compared that with the use of collocations in non-expert learners' corpora in an ESP context.

Researchers have often emphasised the topic- and genre-specificity of collocations in ESP (Marco, 2000; Gledhill, 2000a). Gledhill (2000a:116) claims the usefulness of the use of corpus and genre-based approach in identifying collocations in an ESP context by commenting that "the attraction of a combined approach to both genre and corpus analysis lies in the potential for a corpus to reveal recurrent patterns across a representative sample of texts. The genre approach in turn allows us to nuance the often monolithic descriptions that may emerge from corpus work, by offering a contextual, ethnographic basis for the construction of a textual corpus as well as a view of text as a series of choices, ebbing from one style to the next".

In the Medical discipline, studies conducted by Marco (2000) and Gledhill (2000a) focused on research articles. The former study investigated collocations in Medical research articles while the latter compared the grammatical collocations between different sections of research articles. Both studies confirm the pedagogical importance of grammatical collocations for ESP learners.

From his analysis of 150 cancer research articles, Gledhill (2000a) tried to examine the fixedness and idiosyncratic nature of scientific phraseology. He emphasised the importance of having a representative and specialised corpus of the research articles and a contextual approach to corpus work that is appropriate to the teaching of languages for specific purposes. The results confirmed the importance of collocation patterns in the discourse analysis of Medical research articles. This seems to be in agreement with Halliday and Martin's (1993) view of the central importance of lexico-grammatical patterns in the way discourse is constructed.

Marco (2000) obtained similar results when the three most frequent collocational frameworks were examined in context in his corpus of 100 Medical research articles of 298,457 words. He found that each of these three collocational frameworks was used differently. The first

framework *'the …of'* was used for nominalisation and the second framework '*a… of'* was used for quantifying or categorising. The final framework '*be…to'* was used for lexical items, which indicate relational processes expressing cause or similarity e.g., *be related to* and lexical items that realise modality e.g., *be thought to* (2000: 77). As a result, he pointed out the importance of collocational frameworks in presenting sub-technical items not on their own but together with the syntactic structures where they occur, thus highlighting the integrated relation between lexicon and structure.

Moreover, ESP discourse includes scientific discourse, which has been seen as a specific discourse unlike other types of written discourse. Biber (1988) described science discourse by specific characteristics: the frequent occurrences of nouns, long words, prepositions, conjunctions, agentless and by-passives, the use of past participial adverbial clauses and markedly infrequent occurrences of private verbs, and the use of contractions and that-deletions. The presence of complex technical nouns can be seen as the most difficult characteristic to handle. Halliday (1998) refers to this characteristic as nominalisation. For example, the term *network traffic* nominalises the amount/type of traffic that travels round a network. Nominalisations are important in Science because they allow complex phenomena to be summarised in a few words.

However, researchers such as Yang (1986) and Ward (2007) have used lexical collocations to define these complex technical nouns. None of the previous studies, to my knowledge, has investigated CS academic collocations in experts and non-experts' (both NNS and NS students) writing. To fill this gap, this study aims to locate the most frequent academic collocations used by non-expert postgraduate students of CS in its first stage and then to investigate the factors underlying their over/underuse, compared with experts, of their most frequent academic

collocations. To achieve this aim, the most frequent academic words used by CS postgraduate students will be first located using the AWL as its main point of departure and then will be compared with expert writing in the RC.

2.4.5.4 Academic Collocations across Disciplines

Most of the existing work on collocations in ESP has concerned itself only with the academic collocations found in specific disciplines (e.g., Gledhill, 2000a; Marco, 2000; Ward, 2007; Williams, 1998; Yang, 1986), and few studies focused on academic collocations across disciplines. Durrant (2009: 159) argued that even though Hyland and Tse's (2007) argument for disciplinary divergence in the "use of academic collocations indicates that discipline-specific collocations do exist, it does not indicate that there are not also sufficient across-disciplinary regularities for an EAP collocation list to be of use".

As a result, a number of studies have been conducted to investigate the most frequent academic collocations across disciplines (Biber, Conrad, and Cortes, 2004; Ellis, Simpson-Vlach, and Maynard, 2008; Durrant, 2009; Peacock, 2012; Ackermann and Chen, 2013). Different lists of academic collocations have been developed as different disciplines and different methods were applied in identifying these lists. Academic collocations were defined as "those pairs which appear significantly more frequently in academic than in non-academic texts"(Durrant, 2009:162). However, in this thesis, academic collocations have been defined as those that have Coxhead's (2000) AWL words as node words.

Durrant (2009) compiled his corpus of approximately 25 million words from five academic disciplines: Arts and Humanities, Life Sciences, Science and Engineering, Social-administrative, and Social-physiological to create a list of positionally variable academic collocations. Each sub-

corpus includes approximately 5 million words constructed mainly from research articles. Research articles were chosen as the main genre for academic writing since this text type is considered the central type of academic writing as Hyland (2008) notes that research articles are often "the target of good writing which students are encouraged to emulate" (47).

To identify academic collocations, *WordSmith Tools* (Scott, 1996) was used to calculate the frequency of academic collocations in the academic corpus by comparison with a sub-section of approximately 85 million non-academic texts from the BNC. No criterion was set for the inclusion of collocations, but *log likelihood* was used to produce a ranked list of collocations that were considered the most important to academic writing. The 1,000 most frequent collocations were then identified using both frequency and MI within the academic corpus.

Following these criteria, separate lists of collocations were generated for each of the sub-corpora. This was achieved using the Word List function in *WordSmith Tools* (Scott, 1996). The collocations common to all groups were identified. For these shared collocations, an overall frequency figure for the academic corpus as a whole was then calculated by summing the frequencies in each sub-corpus. Some collocations were removed manually if they included an acronym or abbreviation, a proper name, or an article. They were also excluded if a collocation corresponded to Latin word or if it occurred outside the main text of the article.

Three main results emerged from the previous analysis. First, the 1,000 academic collocations identified were different from the collocations identified by traditional researchers whose focus was on collocations of lexical words. The academic collocations identified were pairs of one lexical and one grammatical word. Out of these 1,000 academic collocations, 763 were grammatical collocations that were described as 'legitimate learning targets'. Although Durrant's

list might not be considered an effective list in determining lexical collocations, it can be pedagogically beneficial since it highlights the use of both grammatical and lexical collocations in different academic disciplines.

Second, another important result was observed when the1,000 most frequent academic collocations located were compared with Coxhead's AWL(2000).Only 425 of these academic collocations included an item from the AWL, which reveals the usefulness of separate academic collocation studies.

Third, there was a clear difference between the academic collocations used in different disciplines. The results showed that collocations for the Arts and Humanities werelower in their occurrences compared to other disciplines. Durrant (2009) suggests that students in the Arts and Humanities use less academic collocation than students in other disciplines do. Therefore, teachers and researchers should pay more attention in dealing with these disciplines than in others.

Unlike Durrant's (2009) academic collocations lists which were created by focusing on two word pairs using a corpus-based approach, Simpson-Vlach and Ellis (2010) adopted the mixed method approach by combining statistical information and human judgement from EAP instructors in developing the Academic Formulas List(AFL) focusing on 3-, 4-, and 5- lexical bundles. Their aim was to create a pedagogically useful list of formulaic sequences that are most frequently used in academic speech and writing. To achieve this aim, a corpus of academic discourse, which included 2.1 million words each of academic speech and academic writing, was used. The academic speech corpus was comprised from MICASE and BNC files of academic speech. The written corpus consisted of Hyland's (2004) research articles' corpus and BNC files

of academic writing. Two other non-academic corpora were used for comparative purposes: the Switchboard (2006) corpus of 2.9 million words was used for non-academic speech and LOB and Brown corpora was used for non-academic writing.

They first extracted all 3-, 4-, and 5- formulas occurring at least 10 times per million from the target and the comparative corpora. Next, they compared the frequencies of the occurrences of formulas in academic and non-academic corpora using *log-likelihood* (LL) ratio. As a result, 979 items were located in the spoken AFL, 712 items were located in the written AFL, and 207 items were found to be the core AFL.

Since Simpson-Vlach and Ellis were not satisfied with ranking AFL by their frequency, they used another criterion that involved using both MI and frequency to rank AFL and to help EAP instructors in their judgement of the pedagogical usefulness of these AFL. This criterion was called 'formula teaching worth'(FTW)(Simpson-Vlach and Ellis, 2010: 488), which was described by Simpson-Vlach and Ellis(2010: 496) as "methodologically innovative approach to the classification of academic formulas, as it allows for a prioritisation based on statistical and psycholinguistic measures, which a purely frequency-based ordering does not".

Following FTW score, they grouped AFL into three groups – core AFL, spoken AFL, and written AFL – and only the 200 most frequent items from each group were given to 20 EAP instructors to judge whether they should be included as useful pedagogical AFL. Then they further categorised the final AFL lists according to their functions into referential expressions, stance expressions, or discourse organisers.

Two other lists of academic collocations have been developed focusing on lexical collocations only, therefore, closer to the interest of the current thesis. Peacock (2012) investigated the frequency and distribution of the most frequent noun collocations in a corpus of 320 research articles complied from eight disciplines. For this purpose, he first located the 16 most frequent nouns using the wordlist function in *WordSmith Tools*. Words that can be used as adjectives or verbs were excluded manually. Then the most frequent collocates of these nouns were identified using the Concord function in *WordSmith Tools* plus Collocate Clusters and patterns sub-functions.

The results showed that even the 16 most frequent nouns seem to have similar collocations in the selected disciplines; however, *process* and *model*, which were frequent collocates within different disciplines, were found to have discipline-specific collocations. When the context of these collocations was examined, they were found to be standard terminology and thus discipline-specific. For example, *software process* and *user model* were found to be specific to CS, while *memory model* and *cognitive process* were specific to Neuroscience.

Peacock (2012) thus claims that the disciplinary variations would be related to the choice of topics, choice of methodology, and content of discussion of each discipline. Therefore, Peacock (2012: 43) argues that collocations represent "disciplinary norms, and that the different patterns presented are accepted within different disciplines as recognised ways for writers to describe and discuss their research". Peacock's (2012) study thus appears in agreement with Hyland and Tse's (2007) claim (for more details about Hyland and Tse's (2007) claims, see section 2.5.3) that there are discipline-specific collocations.

Ackermann and Chen (2013) developed their academic collocations lists (ACL) by again focusing on lexical collocations that occur across academic disciplines. A corpus of 25.6 million words was derived from the written curricular component of the Pearson International Corpus of Academic English (PICAE). It consisted of research articles and textbook chapters of 28 academic disciplines. From this, applying mixed methods of both corpus-driven and expert judgement, a list of 2,468 academic collocations was compiled.

Ackermann and Chen (2013) developed their ACL in four stages. First, a computational analysis was conducted to locate the most frequent content words in their corpus. Using *Microconcord* (Scott and Johns, 1993) and applying MI, *t*-score, and frequency of five, 130,000 collocations were located. Second, manual refinement of the located lists of collocations based on quantitative parameters and part of speech (POS) tagging was carried out. In this stage, they filtered their collocations to those that follow four quantitative parameters: normed frequency >1 per million, normed frequency >0.2 per million in each field of study, MI score>3, and *t*-score>4. They also added POS tagging to facilitate their extraction of the target lexical collocations.

Only four types of lexical collocations were investigated: verb noun, adjective noun, adverb noun, and adverb verb. After their manual checking, the list contained 6,808 collocations. They then assessed these collocations manually to determine whether they should be included or excluded from further analysis following some rules. Collocations that include geographical reference, collocations with high degree of fixedness, collocations with adverbs referring to time, and hyphenated collocations were all excluded. The list reduced to 4,558 collocations.

Third, the refined list of 4,558 collocations was given to six experts to judge whether the collocations included were pedagogically useful and to select which collocations are important. Experts agreed to include 1,215 collocations (27%). In the final stage, they systemised their lists following experts' suggestion to make the lists more accessible to the learners: for example, changing nouns to singular, changing adjectives to their base form, and other processes. Thus, the final list includes 2,468 collocations.

The results showed that noun collocations were the most frequent (74.3%) in the ACL, comprising nearly three quarters of the total lists, followed by verb collocations with nouns and adjectives (13.8%). The other two types of collocations were few. When these collocations were validated in a sub-corpus of the BNC of the same size, the overall coverage was 0.1%, which suggested that the ACL has a 14-times higher coverage in the academic corpus than in a general corpus. It can be clearly seen from the aforementioned studies that they focused on expert writing in developing their academic collocations' lists rather than focusing on non-expert learners' corpora since expert writing was viewed as the standard (Hyland, 2008).

All academic lists, except Peacock's list (2012) were developed using another non-academic comparative corpus. Both frequency and MI were used in locating collocations in Ackermann and Chen's (2013) ACL and Peacock's list of lexical collocations (2012). However, Simpson-Vlach and Ellis (2010) used them for ranking collocations. On the other hand, *log likelihood* was used in locating the most frequent collocations in both Durrant's (2009) and Simpson-Vlach and Ellis's (2010) studies. Thus, it can be concluded that different criteria were applied in locating collocations and this depends on the researcher purpose.

Another interesting point is the use by Ackermann and Chen (2013) and Simpson-Vlach and Ellis (2010) of expert judgement. They asked experts to rank their lists to include the most pedagogically useful items. Ackermann and Chen (2013) also used experts' views in refining their ACL lists in the final stage to be ready for use by teachers and researchers. Thus, it seems that depending on corpus-driven data alone might not be enough in deciding which collocations are most beneficial in pedagogical setting; experts' judgments are also needed.

2.5 Academic Word List (AWL) and ESP Wordlists

My study is of academic collocations, not just of any collocations that occur in academic text. I have decided to operationalise this concept in part by requiring every academic collocation to contain at least one academic English word. Hence, I will need a list of academic words for reference. For this reason, I next review the available lists and how they have been established.

2.5.1 General Vocabulary, Academic Vocabulary, and Technical Vocabulary

Vocabulary has been classified into three main types throughout the literature: general vocabulary, academic vocabulary, and technical vocabulary. Dresher (1934) was the first scholar who made this distinction, which was then accepted by other researchers (Nation, 2001). Coxhead and Nation (2001) have added low frequency words as a fourth category. Both categories were based on the frequency of words in specific texts as one of their main criteria. Some additionally use the term 'core vocabulary', which refers only to the first 2,000-3,000 words in general use, covering approximately 80% of most texts. A list of these words was originally made in the General Service List (GSL) by West (1953).

Different terms have been given to academic vocabulary by different researchers: sub-technical vocabulary (Cowan, 1974; Yang, 1986) and specialised non-technical lexis (Cohen, Glasman, Rosenbaum-Cohen, Ferrara and Fine, 1988). The term also covers semi-technical vocabulary that can be best defined as "formal, context-independent words with a high frequency and/or wide range of occurrence across scientific disciplines, not usually found in basic general English courses" (Cowan, 1974: 391). This academic vocabulary can be exemplified by the following words: *compound, achieve,* and *proportion*. It covers approximately 8% to 10% of academic texts. Further, these words are mostly used in academic texts but have a considerably higher frequency of occurrence in scientific and technical descriptions and discussions (Dudley-Evans and St. John, 1998).

Moving to the final type, technical vocabulary is considered the most specialised vocabulary compared to the previous mentioned types. It can be defined as "specialised subject related vocabulary" (Nation, 2001; Chung and Nation, 2004; Kennedy and Bolitho, 1984) that "occurs in a specialist domain and part of a system of subject knowledge" (Chung and Nation, 2004: 252).

Throughout the literature, different terms have been given to technical vocabulary: terminological words (Becka, 1972), specialised lexis (Baker, 1988), and specialist vocabulary (Kennedy and Bolitho, 1984): It covers up to 5% of the running words in technical texts. The technicality of these words can be classified into more detail depending on the criteria of relative frequency of form and meaning in the field (Nation, 2001). For example, *pixel* and *modem* are highly technical computing words as they are unique in both form and meaning to the field,

while *program* and *icon* are less technical words as they are more common in both form and (different) meaning outside the field(Nation, 2001: 199).

Turning to academic vocabulary, many studies show that it is problematic for learners, especially EFL learners (Baker, 1988; Kennedy and Bolitho, 1984). According to Cohen et al. (1988), two main reasons cause this difficulty: first, the meaning is unknown to the learners as they are used in both general and technical contexts. Second, learners are not aware of their lexical relations. That is, they cannot recognise related words if they are used in paraphrasing. Baker (1988) also found from her study that academic (semi-technical) vocabulary causes the real difficulty for students, especially in their writing.

Cohen et al. (1988) conducted several studies investigating the effects of technicality levels on L2 vocabulary acquisition. They concluded that academic vocabulary poses more difficulty for EFL learners than technical vocabulary since the latter has fixed meanings that can be learned more easily. In addition, students may not be as familiar with academic vocabulary as they are with their subject-technical vocabulary (Worthington and Nation, 1996; Xue and Nation, 1984). In ESP settings, learners face more difficulty dealing with academic vocabulary than with technical vocabulary (Strevens, 1973) since they have regular access to their discipline-technical vocabulary more than academic vocabulary (Li and Pemberton, 1994; Shaw, 1991; Thurstun and Candlin, 1998).

Although academic vocabulary has been considered a challenge for EFL learners, it plays an important role in constructing the meaning of a text. Its importance is related to “its supportive role in learners’ academic writing such as describing and evaluating empirical studies” (Storch

and Tapper, 2009: 212). A number of studies have clearly demonstrated the importance of the role of academic words in academic texts and the necessity of students' acquisition of this set of words (Shaw, 1991; Li and Permberton, 1994; Santos, 2002). Using *Microconcord* (Scott and Johns, 1993), Thurstun and Candlin (1997) developed comprehensive concordance-based accounts of the rhetorical functions of the AWL in academic texts, which covered a wide range of academic disciplines.

In conclusion, it can be said that the four kinds of vocabulary are not equally difficult at different stages of learning. Some words deserve more attention and effort than others do for different learning purposes. It has been agreed that the first 2,000 most frequent word families represented in the GSL are more important to beginners than to intermediate or advanced learners who may need to shift their concern to academic vocabulary (Nation and Waring1997). Academic vocabulary plays an important role both in general and specific academic settings.

2.5.2 Development of the AWL

A number of researchers have tried to develop general academic vocabulary lists since the1970s. Three different approaches have been applied to develop such a list. First, words translated into learners' first language were collected from their textbooks. Lynn (1973) and Ghadessy (1979) made their word lists by counting the translated words from their learners' textbooks that were identified as difficult words in their reading texts. Second, a corpus of specialist academic English such as from Electronics (Farrell, 1990) or Medicine (Salager, 1983) was analysed to classify the kinds of vocabulary found and to eliminate the general words presented in the GSL. The third approach, mostly used today, is to compile a diverse academic corpus, which covers

various disciplines, then to exclude words from the GSL and to identify the most frequent words in various disciplines.

Following the last approach, Campion and Elley (1971) developed their word list consisting of 500 words by analysing textbooks and research articles from 19 academic disciplines. In their list, they tried to cover the words encountered by university students. Praninskas (1972) compiled the American University Word List extracted from 10 university-level textbooks covering 10 academic disciplines. Xue and Nation (1984) combined the previous four lists (Campion and Elley, 1971; Praninskas, 1972; Lynn, 1973; Ghadessy, 1979) into a University Word List (UWL). However, Coxhead (2000: 214) crictised the UWL as "it lacked consistent selection principles and had many of the weaknesses of the prior work. The corpora on which the studies were based were small and did not contain a wide and balanced range of topics".

Coxhead (2000) therefore tackled the deficiency of the previous word lists (Campion and Elley, 1971; Praninskas, 1972; Lynn, 1973; Ghadessy, 1979; Xue and Nation, 1984) by using a diverse academic corpus of 3.5 million words to develop her AWL. The AWL was carefully designed, taking into consideration the requirements of compiling a representative organised corpus. Following the advice of Sinclair (1991), the father of corpus linguistics, different length texts written by various writers within each discipline were selected to ensure the inclusion of a representative range of lexical types in the corpus(Sutarsyah, Nation, and Kennedy, 1994) and to avoid the bias that may result from the idiosyncratic style of one writer (Sinclair, 1991).

A corpus of 3.5 million running words of academic writing was compiled from different written genres: research articles, university textbooks, and laboratory manuals. Four main discipline

areas were included in the corpus –Law, Commerce, Arts, and Science – each containing approximately 875,000 running words and each sub-divided into seven subject areas.To develop academic word lists, word families were selected for two main reasons: they have proved to be an important unit in the learner's mental lexicon (Nagy et al., 1989) and the inflected and derived members of the word family are not difficult to be learned later (Bauer and Nation, 1993).

Moreover, three criteria were used in developing the AWL. First, specialised occurrence of the words: that is, words should not be included from the first 2,000 of the GSL of West (1953). Second, range: a member of a family should occur 10 times or more in all of the four disciplines and in half or more of the 28 subjects included. Third, frequency: a member of a family has to occur 100 times or more in the compiled corpus (Coxhead, 2000). As a result, a list of 570 word families[1]was developed. These word families were divided according to their frequency into ten sub-lists, each consisting of 60 word families, with the exception of the final sub-list that consisted of only 30 word families.

The results showed that the 570 word families account for 10.05% of the running words of the whole corpus and occurred in a wide range of the subject areas in the academic corpus. In total, 67% of the word families in the AWL occurred in 25 or more of the 28 subject areas and 94% occurred in 20 or more of the 28 subject areas. The coverage of the AWL in the four disciplines chosen was different, with the highest coverage (12 %) occurring in Commerce and the lowest (9%) in Science.

[1] For more details about the AWL sub-lists visit http://www.victoria.ac.nz/lals/staff/Averil-Coxhead/awl/awlinfo.html

2.5.3 Criticisms of the AWL

Although the AWL has made a significant contribution to vocabulary research and has been evaluated by comparing its coverage with another academic corpus (Coxhead, 2000), it has also been criticised, mainly by Hyland and Tse (2007), for some issues related to the AWL words' range, frequency, collocation, and meaning in different disciplines. Their main criticism has questioned the generality of the AWL to the linguistic production of students of various disciplines and queried the assumption of the existence of a single core academic vocabulary valuable to all students irrespective of their field of study.

For this reason, a corpus of 3.3 million running words was compiled, collected from three main fields: Sciences, Engineering, and Social Sciences. Different genres were included from each discipline: research articles, textbook chapters, book reviews, and scientific letters, which were only taken from Biology and Physics. Unlike Coxhead's (2000) academic corpus, Hyland and Tse's (2007) corpus does not only include professional writing but also students' writing. Both postgraduate and undergraduate students' writing was included to represent students' productive use of academic vocabulary.

Their analysis reveals that the AWL has a number of criticisms but also some limitations. First, the coverage of the 570 AWL words was not equally distributed in their multi-disciplinary corpus. They found that approximately 534 (94%) out of the 570 AWL families were different in their distribution across the sub-fields, with the majority of these occurring in one sub-corpus. Approximately 227 (40%) had at least 60% of all occurrences concentrated in one discipline. Only 36 word families were found to be equally distributed across the sub-fields. The Sciences

were the lowest disciplines in coverage by academic vocabulary. They recommended that more valid criteria are needed to determine the most frequent words, using the mean frequency of words in the whole corpus rather than the threshold of 100 occurrences of words applied by Coxhead (2000). By applying their frequency criteria, only 192 word families, covered by approximately a third of the 570 AWL families, were considered frequent.

Second, Hyland and Tse argued, "All disciplines shape words for their own uses" (2007: 240).That is, words have specific meanings that tend to be meaningful to members of specific disciplinary communities. From their investigation of a set of academic vocabulary items, they found that there is a semantic variation across disciplines. For example, the noun *analysis* is often associated with particular types of approach to form a discipline-specific compound noun (technical term) such as *genre analysis* in linguistics or *neutron activation analysis* in Science. The verb form is also used differently since it has a different meaning in Social Sciences of "considering something carefully" while in Engineering it stands for "methods of determining the constituent parts or composition of a substance" (Hyland and Tse, 2007: 244). See examples 1and 2 taken from Hyland and Tse (2007: 245):

*Example 1.*We used a variety of methods to *analyse* fungal spore load, volatiles, and toxins. (Biology article)

*Example 2.*The major objective of this report is to *analyse* developments in political sociology over the last half century. (Sociology article)

To tackle such issues, a new Academic Vocabulary List (AVL) has recently been developed by Gardner and Davies (2014) by identifying word lemmas (inflections only) rather than word families (inflections and derivational forms) from a sub-corpus of 120 million words derived

from the 425 million word Corpus of Contemporary American English (COCA) (Davies, 2012). Their corpus covered nine academic disciplines and included more academic journals compared to Coxhead's corpus (2000) and Hyland and Tse's corpora (2007). To create their core academic vocabulary list, they applied four statistically robust criteria: ratio, range, dispersion, and discipline measure.

Ratio of frequency in the academic texts compared with the non-academic part of COCA (minimum 1.5) was used to exclude general high-frequency words from academic core words. Range (occurrence in at least 7 out of 9 disciplines), Dispersion (minimum 0.8 on a scale with maximum 1), and a Discipline Measure (the word (lemma) cannot occur at more than three times the expected frequency in any domain: e.g., the word *federal* occurs at 3.69 the expected frequency in Law, so is not included as a "core academic" word)(Gardner and Davies, 2014) were used together to exclude technical words and words that occur mainly in only one or two disciplines.

A word list of 3,000 lemmas was created and called the AVL. They were also able to locate three different categories of academic words: core academic words, general high- –frequency words and discipline- specific words. For example, *define* and *definition* were categorised as core academic words as they occur across all academic disciplines, while *definitional* and *indefinable* were categorised as discipline- specific words as they occur in two or more of the academic disciplines.

As a result, 2000 word families were created by grouping the related inflected lemmas of the same POS (e.g. *define, defines, defined,* and *defining* (for the verb *define*) with their frequency

presented in descending order. Thus, the adjective *defining* was identified as a separate lemma since it has a different POS. When the top 570 word families of the AVL was compared with the 570 word families of AWL in COCA and BNC, the AVL coverage of both academic corpora was twice the AWL. This result indicates the fundamental differences in creating these two lists. Thus, Gardner and Davies' AVL (2014) can be considered the most recent and useful list of core academic words. However, it was not yet published at the time the current study was conducted, so I relied instead on the AWL.

In addition, Hyland and Tse (2007) have argued that words may take on additional discipline-specific meanings because of their regular co-occurrence with other items. For example, the word *strategy* has different co-occurrence in different fields; *marketing strategy* in Business and *learning strategy* in Applied Linguistics (2007:246). Thus, they have called for paying more attention to context, co-text, and the use of collocations of these academic vocabulary items.

By contrast, Wang and Nation (2004) in their investigation of the homographic features of the 570 AWL families found that only 60 word families could be considered homographic and that their different senses were not problematic since they met the criteria for frequency and range necessary to remain in the list. Therefore, they recommended that the AWL could be considered as a general academic list since only a small percentage of the 570 word families are homographic.

Another problem identified in the AWL was the decontextualised use of these words. The ignorance of the context and co-text of the words makes AWL seem "a chimera". Hyland and Tse (2007:247) argued, "Different views of knowledge, different research practices, and

different ways of seeing the world are associated with different forms of argument, preferred forms of expression, and most relevantly specialised uses of lexis".

Their results also showed that the AWL families were underrepresented in the Sciences, suggesting the need for a list of specialised scientific vocabulary. Therefore, they have described the AWL as a "cline of technically loaded or specialised words ranging from terms which are only used in a particular discipline to those which share some features of meaning and use with words in other fields"(Hyland and Tse, 2007:249).Thus, they suggested that the AWL could be divided into two lists: the general AWL that is related to common terms that occurred in most academic fields and the specific AWL that consists of more specific terms related to certain academic fields.

In their development of AVL,Gardner and Davies (2014) found that academic vocabulary can be identified into three categories: a)core academic words 'those that appear in the vast majority of the various academic disciplines'; b)general high frequency words'those that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register'; c)discipline-specific words 'those that appear in a narrow range of academic disciplines'(2014:8).

Hyland and Tse's call for developing specialised academic word lists (SAWLs) has been considered an important issue in ESP. A number of researchers (Wang, Liang, and Ge, 2008; Coxhead and Hirsh, 2007; Ward, 2009; Chung, 2009) have conducted studies to discover the SAWL related to a particular discipline. Almost all of the studies including Chen and Ge (2007) have used the AWL as the starting point to determine which of the 570 word families are specific in their fields. Few researchers have made their own lists considering the most frequent academic

words in their own corpus such as the Medical Academic Word List (MAWL) developed by Wang et al. (2008). On the other hand, Minshall (2013)[2] created a specific technical word list for CS as a supplementary list to the AWL and the GSL. The results show that the AWL is significant in making SAWL for certain fields. These studies will be reviewed in more detail in section 2.5.4.

2.5.4 Corpus-based Studies based on the AWL

Although the AWL was developed to meet EAP learners' needs, a number of studies on academic vocabulary using different corpora compiled from different genres have challenged the usefulness of the AWL in ESP courses. Based on Hyland's (2002; 2006) main claim for specificity in ESP, a number of researchers have tried to develop SAWLs for distinct disciplines to meet students' discipline-specific needs (Chen and Ge, 2007; Martínez et al., 2009; Li and Qian, 2010). This section will review the most current studies that contributed to the establishment of SAWLs in different fields. The procedure used in compiling a specific corpus, in identifying the most frequent AWL, and the main findings revealed will be highlighted and compared to discover which procedure will be followed in the current study.

2.5.4.1. The Procedure Applied for Developing SAWLs for ESP

2.5.4.1.1 Using the AWL

[2]Even though the Computer Science Word List (CSWL) developed in Minshall (2013) was technical in nature, it could be used with the AWL to exclude the technical words from this thesis' analysis. Unfortunately, it had not been developed at that time.

The AWL has been used as the base list by a number of researchers in order to develop their SAWLs for different fields (Medicine (Chen and Ge, 2007), Finance (Li and Qian, 2010), Agriculture (Martínez et al., 2009), Business (Konstantakis, 2007), Engineering (Mudarya, 2006; Ward, 2009) and Applied Linguistics (Vongpumivitch et al., 2009)). Although different genres have been used in the previous studies for corpus compilation (mainly research articles) (including Chen and Ge, 2007; Martínez et al., 2009), few have focused on textbooks (Mudarya, 2006; Coxhead et al., 2010). It must be stressed that the procedures these studies followed in identifying their SAWLs were almost similar.

The typical procedure for identifying the specific AWL for various fields was carried out following three main steps. First, after the target corpus was compiled, the most frequent words in the corpus were located using one of the following programs: Range [3](Nation and Heatley, 2007), which was used by Coxhead (2000) in developing the AWL was considered the most applicable program. It was used either as the main software (Ward, 2009) or was complemented by other software such as *Wordsmith Tools* (Scott, 2004) in Li and Qian (2010) or *Corpus Builder* (Hyland and Tse, 2007). *Wordsmith Tools* was also used as the main software in identifying the most frequent words by others (e.g., Mudarya, 2006; Martínez et al., 2009). Other self-designed programs were also used by other researchers such as Chen and Ge (2007).

Different criteria have been applied in order to identify the most frequent words. The three main criteria – specialised occurrence, range, and frequency – used in developing the AWL were applied by Wang et al. (2008) in developing their MAWL and by Vongpumivitch et al. (2009) in

[3] This program is available as a free downloadable zip file at http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx.This software is preloaded with West's (1953) GSL of the most frequent 2,000 English words and the AWL and it shows the frequency of items from each list in any corpus together with its range or the number of different sub-corpora they occurred in (Hyland and Tse, 2007: 239).

their Applied Linguistics AWL. Others have focused on the frequency and the distribution of the words in the corpus as their main criteria for distinguishing the most frequent words in certain specific fields (Martínez et al., 2009; Chen and Ge, 2007).

The frequency criterion was considered questionable by a number of researchers. Some researchers, such as Ward (2009), have followed Coxhead's (2000) frequency criterion that was based on 100 occurrences overall with at least 10 in each of the four corpus-represented disciplinary areas; others such as Martínez et al. (2009) have followed Hyland and Tse's (2007) frequency criteria that considered items as frequent if "they occurred above the mean for all AWL items in the corpus" (240).

A word was defined as a word family in most corpus-based AWL studies including Martínez et al. (2009). The analysis carried out relies on assumptions about the way vocabulary is organised in the mind of the people who provided the samples for the corpus. It has been assumed that most of the research articles used in compiling these corpora (Chen and Ge, 2007; Wang et al., 2008; Martínez et al., 2009) were written either by advanced NS or NNS writers. Therefore, their word knowledge is most likely to be built on word families rather than on its separate inflected and derived members (Coxhead, 2000). In the second stage of developing SAWLs, the identified frequent words are compared to other word lists, mainly the AWL, to locate the specific words from the 570 AWL word families that are related to the specific target field.

Thirdly, qualitative analyses have been applied to check word meanings and collocations by running their concordance. Martínez et al. (2009) found that the word *culture* has a specific meaning in the Agriculture corpus rather than the general meaning. It was revealed that *culture* was used with meanings associated with Agriculture, meaning "cultivation of plants" (e.g.,

blueberry cell cultures, cultures were grown). The word *strategy* also has discipline specific collocational patterns (e.g., *control strategies, management strategies,* and *adaptation strategy*), unlike the word *strategy* in Applied Linguistics (e.g. *learning strategy*) or in Business (e.g. *marketing strategy*). It is clear that the same word has different collocational patterns in different fields. This finding is in agreement with Hyland and Tse's (2007) claim for discipline-specific meanings and collocations.

2.5.4.1.2 Developing SAWL

The AWL has been considered unable to cover all academic vocabulary in specific disciplines such as Medical Science (Chen and Ge, 2007) and Agriculture (Martínez et al., 2009). Moreover, it assumed to be a complex and difficult word list for undergraduate students who have no mastery of the 2,000 GSL. Therefore, other approaches have been employed to tackle these two main issues observed by the previous studies. Both Wang et al. (2008) and Ward (2009) have developed their own SWLs from scratch. These two studies will be reviewed in detail to examine their methodological strengths and weaknesses.

Wang et al. (2008) compiled their Medical corpus from Medical research articles. They selected their target word families to be included in the MAWL; following Coxhead (2000), three main criteria were applied in developing the MAWL. First, the word families included in the GSL were eliminated. Then, the word families that occur in at least 16 or more of the 32 subject areas included in their corpus were selected. From the selected word families, only those that occurred at least 30 times in their corpus were selected for the Medical Word List. In case of uncertainty about the inclusion of some word families, two experienced English professors who have taught and conducted studies on English for Medical Purposes for more than 20 years were consulted.

A corpus of one million words consisting of 623 word families was compiled. The coverage of the word families was satisfactory. Approximately 104 of 623 word families occurred in all 32 subject areas and 321 in 25 or more subject areas. When the MAWL was compared to the AWL, only 342 of the 623 word families overlapped with the 570 word families in the AWL. Their results confirm Hyland's main argument for researching SAWL, which claimed that "different practices and discourses of disciplinary communities require a more restricted discipline-based lexical repertoire, which undermines the usefulness of general academic word lists across different disciplines"(Hyland, 2008:451).

Wang et al.'s (2008) study is considered a preliminary investigation of the MAWL, but the methodology employed in establishing the MAWL is very convincing, since they compiled a large corpus following the same criteria applied in developing the AWL. Additionally, they consulted two experts in the field who had 20 years' experience. Their corpus can be considered representative since it covered a wide range of Medical subjects.

Ward (2009) questioned the usefulness of the AWL to undergraduate students who have no mastery of the 2,000 GSL. He also argued that the AWL could not be seen as relevant to Engineering as a discipline. Coxhead's (2000) corpus has no Engineering section and it, therefore, cannot address the specific needs of Engineering students. Hyland and Tse's (2007) sub-corpus of Engineering, which consisted of only 569,000 words, was confined to Mechanical and Electronic Engineering. In order to identify the vocabulary frequency in a wider range of Engineering sub-disciplines in a specific genre, Ward therefore compiled his own corpus of approximately 250,000 tokens from 25 Engineering textbooks, which were chosen after expert consultation.

Obtaining frequency data for all the words over the five sub-sections representing each of the Engineering sub-disciplines, a 299-word list for foundation Engineering students was compiled. It is considered to be a relatively easy target for learners whose high school education has not equipped them for the linguistic challenges they face in reading English language textbooks. The list is short and non-technical in nature, but gives excellent coverage of a wide variety of Engineering textbook materials. By concentrating on word types rather than lemmas or families, it encourages learning of not only individual words but also their lexico-grammatical environments.

Unlike previous studies, Minshall (2013) developed a Computer Science Word List (CSWL) that was intended to cover the technical words of that field. It was created to serve as a pedagogical list for the NNS CS students who are studying in UK universities. It was also created to supplement the AWL and the GSL. Another aim was to discover if Multi Word Units (MWU) exist in CS. A corpus of 3,661,337 tokens was compiled from 165 journal articles (1.8 tokens) and 243conference proceedings (1.8 tokens) covering the 10 sub-disciplines of CS as defined by the Association for Computing Machinery (ACM).

Coxhead's (2000) three main criteria (see section 2.5.2) were applied in developing the CSWL. First, the specialised occurrences' criterion was applied by eliminating the word families included in the GSL and AWL. Second, the range criterion was applied in a modified version to cover words that are present in at least half of the corpus (Wang et al., 2008; Coxhead and Hirsch, 2007). Thus, words should appear in five of the ten sub-corpora to be included in the CSWL. Third, the minimum frequency of 80 occurrences was applied since the size of the CSC is similar to Coxhead's corpus (2000). A list of 433 words was developed.

The results showed that the CSWL was highly technical and covers 6.0% of the CSC, while the coverage of the AWL was 12.79%. These results are in agreement with the required coverage of the technical words (5%) and the academic words (10%) as mentioned in section 2.5.1. Surprisingly, the coverage of the AWL is higher than the coverage of the previous SAWL studies. This coverage indicates that the AWL is a useful list for identifying academic words in any CS corpus.

To investigate the existence of a Computer Science Multi Word list (CSMWL), Minshall ((2013) applied the same two criteria for developing his CSWL: range and frequency. Using *Antconc*, a list of 23 CSMWL words was developed after locating the hyphenated words in the CSWL and other multi-words outside the CSWL: "The CSMWL showed that whilst multi-word units do exist in CS literature, they are mostly compound nouns with domain specific meaning"(Minshall, 2013:1). Even though the CSMWL seems to be limited in number, it could be used to verify the existence of specific CS collocations in my study from general academic collocations (see section 5.3.2 for more information about these categories of collocations). Unfortunately, this CSMWL was not available at that time.

2.5.4.2 Main Findings from Previous Studies

From the previous studies, a number of findings have been observed. First, the coverage of the AWL in most of the specific corpus-based studies is in agreement with the percentage of AWL in Coxhead's study (2000): "This seems to testify the claim that AWL covers approximately 10% of any academic text" (Coxhead and Byrd, 2007:132). The highest percentage of the AWL coverage of 12.79% occurred in Minshall's (2013) CSC followed by 11.51% occurred in Konstantakis's (2007) Business corpus while the lowest percentage of 1.4 occurred in Coxhead's (2000) Fiction corpus. Other corpus-based studies compiled from research articles in different

disciplines have varied in their AWL coverage. The AWL word families cover 10.46% of Li and Qian's (2010) Financial corpus while it covers 9.06 % of the Agriculture corpus (Martínez et al., 2009).

The frequency of the AWL word families has also varied in these studies. Li and Qian (2010) in their analysis of the Financial corpus found that 162 word families –only approximately 28.42% of the AWL – occur in their corpus. Vongpumivitch et al. (2009) found 475 word families of the AWL in their Applied Linguistics corpus. In their Agriculture corpus, Martínez et al. (2009) found that only 92 families of the AWL occurred in their specific list. This recognised variation of the frequency of word families reinforces the idea that differences are based on discipline-specificity (Hyland and Tse, 2007). By contrast, only two word families of the AWL did not occur in Minshall's (2013) CSC. Minshall (2013) highlights that many of the polysemic words of AWL (Wang and Nation, 2004) that were also a CS bias (e.g., *data*, *process*, *section*, *compute*, and *network*) had very high presentation in the CSC.

Another interesting finding about the occurrence of the most frequent word families in the SAWLs adds extra evidence to Hyland and Tse's (2007) contention that the more specific the corpus the greater the specificity of items and, consequently, the lower the variability. In their Agriculture corpus, Martínez et al. (2009) found that only 26 of the 92 frequent word families occurred in the first sub-list of the AWL.

Chen and Ge (2007) reported that their most frequent words did not occur as frequently as they had in Coxhead's study and vice versa. For example, words such as *legal* and *economy* were ranked as high frequent words in Coxhead's corpus and were listed in sub-list 1 of the AWL, while they were less frequent in their corpus. However, some low frequent words in Coxhead's

study, such as *found*, *detect*, and *induce* were ranked as high frequent words in their Medical corpus.

2.5.5 Gaps in Previous SAWL Corpus-based Studies

Developing SAWLs for various disciplines have been one of the main achievements of AWL research. Throughout the last ten decades, a number of studies were conducted to develop their SAWLs for various fields: Medicine, Agriculture, Finance, Business, and Engineering.

However, no studies have been conducted to investigate the use of the AWL in CS students' writing. Lam (2001) investigated the difficulty of the AWL when reading academic texts. Using both tests and retrospective interviews, she noted that learners face difficulty in understanding the specific meaning of the AWL in their technical computer texts. She also noted that the semantic distinction of the AWL from the same vocabulary when it appeared in general texts is one of the reasons for viewing AWL as difficult. Her recommendation was that such lexical terms should be presented as a glossary of academic vocabulary with information about frequency of occurrences based on a specialised corpus. Thus, a call for developing SAWL for CS is needed. Minshall (2013) created CSWL of technical vocabulary for pedagogical purposes. He excluded the AWL and thus it was not considered a SAWL for CS.

Since the main aim of the current thesis is to locate the most frequent academic collocations used by CS postgraduate students in their writing, the AWL is used first to locate the most frequent academic words in the students' corpora and then to identify their collocations.

2.6 The Focus of the Current Thesis

Having reviewed the literature of corpus-based studies on collocations' use, it appears that most of the studies were about learner use of any collocations (not just academic collocations) in what is mostly general EAP (see section 2.4.5.2 for full details). No studies (to my knowledge) have been conducted to investigate ESP learners' or non-expert NS writers' use of such collocations in their academic writing. Lam (2001) was the first to examine students' use of academic collocations in CS for different purposes. She developed a list of difficult semi-technical (academic) words to aid L2 learners in their understanding of CS texts.

Minshall (2013) developed the CSWL and the CSMWL from research articles and conference proceedings to serve as pedagogical lists for NNS CS students. These two lists could be useful to my selection of the academic collocations: the CSWL could be used with the AWL to exclude the technical words from my analysis while the CSMWL could be used in excluding the specific CS collocations from my selected list of collocations. Unfortunately, they were not available at the time of this current study.

CS has been chosen for this study for two main reasons. First, to my knowledge, there are no collocations studies conducted in this discipline that focus on comparing non-expert students' use of academic collocations in their writing with experts' CS use. Second, one of the main problems I have encountered while teaching English to CS undergraduate students at Umm al Qura University (UQU) in Saudi Arabia is their difficulty in writing good essays, due to their misuse of academic collocations. Moreover, one of the main findings from Farooqui's (2010) investigation of the difficulty of academic vocabulary for CS undergraduate students in UQU

was their misunderstanding of the concept of collocations as well as their misuse of collocations in their writing.

Thus, I am interested in investigating their academic collocation use as well as which factors underlie the over/underuse of the most frequent academic collocations and to discover which types of activities will raise their awareness about the use of academic collocations. At the beginning, five research questions were the focus of this thesis (1, 2, 3, 6, and 7). The other research questions (4 and 5) emerged from the analysis of results of the first study presented in Chapter 4:

RQ1. What are the most common academic collocations used by Computer Science students in their MSc dissertations?

RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus?

RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?

RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora?

RQ5. To what extent do the shared collocations differ in their patterns?

RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views?

RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?

RQ7. What kind of teaching materials are needed to raise NNS students' awareness of the use of academic collocations?

2.7 Summary

In this Chapter, I reviewed the relevant literature to situate and make the case for the first study (presented in Chapter 4). Three areas of literature, namely, formulaic language, collocations, Coxhead's AWL (2000), and other SAWLs, were reviewed. The literature review has shown how corpus analysis was taking the lead in the studies of collocations in EAP (e.g., Durrant and Schmit, 2009); however, only a few corpus-based collocational studies were conducted in ESP (e.g., Gledhill, 2000). Since no studies have been conducted to investigate the use of academic collocations in CS experts and non-experts' writing, the current study initially aims to locate CS academic collocations in experts' academic writing to be compared with their uses in non-expert students' writing.

The review also highlights the effectiveness of Coxhead's AWL (2000) in developing SAWLs for different academic disciplines; thus, it will be used in locating the most frequent academic words in students' corpora to be searched for their collocations. Although Minshall (2013) located the CSWL and the CSMUL, his wordlists are considered more technical than academic. Thus, it was decided that these wordlists would not be used. Two other areas of research relevant to this thesis, collocation patterns and awareness raising materials, – will be reviewed in Chapter 5 and in Chapter 6, respectively.

Chapter 3 Broad Methodology and Corpus Design

3.1 Introduction and Overview of the Methods Applied

This Chapter presents the broad methodology applied in this thesis and explains how each aspect of the research method addresses my research questions. It comprises two main sections: the first section presents a summary of the main studies conducted, the main research questions, and the methodology applied in each study; the second section reviews the related literature about corpus design and then describes the compilation of the three corpora designed for this thesis.

The main aim of this thesis is to discover the most frequent academic collocations used by postgraduate students of CS and to investigate the factors behind their over/underuse of the located academic collocations. To achieve this aim, a corpus-based study was carried out. The most frequent academic collocations were located in NNS and NS students' corpora. Due to the limitations of the corpus-based approach (Stubbs, 2001; Widdowson, 2000)(for more details about these limitations see section 3.5) and due to the researcher's restricted knowledge of CS, further qualitative analysis was undertaken to investigate the factors behind the over/underuse of the most frequent collocations used by postgraduate CS students.

CS experts were asked to verify my analysis and to categorise collocations according to their specific uses in the three selected sub-disciplines of CS (Artificial Intelligence (AI), Information

System (IS), and Software Engineering (SE)). The secondary aim for this thesis is to develop teaching materials that could be used to help non-expert NNS students with collocation over/underuse problems. Table 3-1summarises the main research questions addressed and the methods applied by the studies reported in this thesis.

Table 3-1: A summary of the main studies, research questions, and methods applied in this thesis.

Chapters/studies	**Research questions**	**Methods**
Chapter 4/ The use of academic collocations by non-expert CS postgraduate students	RQ1: What are the most common academic collocations used by Computer Science students in their MSc dissertations? RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus? RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?	Quantitative – frequency-based approach Applying association measures(*t*-score and MI) to locate strong collocations in students' corpora using *ConcGram*(Greaves, 2005)
Chapter 5/ Factors underlying the non-experts' over and underuse of noun collocations	RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora? RQ5. To what extent do the shared collocations differ in their patterns? RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views? RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?	Mixed methods 1-Quantitative – patterns' identification following Hunston (2002b), Hunston and Francis (1996), and Coxhead and Byrd's (2012) procedures. 2-Quantitative – Categorisation judgement task 3-Qualitative – semi-structured interviews

Chapter 6/ Academic Collocations' Awareness-raising activities	RQ7. What kind of teaching materials are needed to raise NNS students' awareness of the use of academic collocations?	Corpus-based activities Based on the literature review

Chapter 4 presents the most frequent academic collocations used by NNS and NS students, after they were verified in the RC applying the frequency-based approach. Then, students' over/underuse of the most frequent collocations was compared with the RC as well as between NNS and NS corpora. Steps of locating collocations and testing their significance are all presented in detail in section 4.5.2

Chapter 5 reports on both patterns' identification and verification of the results. To answer the second research questions regarding the factors behind students' over/underuse of academic collocations, both quantitative and qualitative methods were applied. First, patterns were identified for the 24 shared N collocations among corpora following Hunston (2002b), Hunston and Francis (1996), and Coxhead and Byrd's (2012) procedures, then CS experts were asked to fill in a categorisation judgement task and were interviewed to gain better understanding of the results and to verify our primary analysis. Steps for identifying patterns, design, and results of both categorisation judgement task and semi-structured interviews are all presented in detail in Chapter 5.

Chapter 6 presents a sample of pedagogical awareness-raising activities that were designed using corpus-based approach in teaching collocations: data-driven learning (DDL) (Johns, 1986; 1991a; 1991b). These activities were mainly devoted to raise NNS students' awareness about collocations' use and patterns.

3.2 Corpus Design

In this section, first corpus deign considerations will be reviewed and then design issues related to my corpus design will be presented.

3.2.1 Learner Corpus Design Considerations

The compilation of a corpus is difficult and time consuming. Some people want quick solutions and tend to cut corners when designing and building corpora, but studies based on such corpora may yield results that are not valid or reproducible. Thus, it should be designed carefully. Tono (2003: 801) confirms that "If data is gathered in an opportunistic way without proper control and documentation of learner and task variables, the resulting corpus will be unlikely to be of much use".

Although most learner corpus-based studies vary on their corpus design due to the research aims and aspects of language investigated, in all cases of corpus compilation certain design considerations need to be taken into account. Tono (2003: 800) divides these design considerations into three main categories: language-related, task-related, and learner-related. Each category has further sub-divisions. Table 3-2 (taken from Tono, 2003: 800) offers a complete list of these considerations.

Table 3-2: Corpus Design Considerations according to Tono (2003: 800).

Language-related	**Task-related**	**Learner-related**
Mode (written/spoken)	Method of collection (e.g. cross-sectional/longitudinal)	Internal-cognitive (age/cognitive style)

Genre (e.g. fiction/essay)	Method of elicitation (e.g. spontaneous/prepared)	Internal-affective (motivation/attitude)
Style (e.g. narration/ argumentation)	Use of references (e.g. access to dictionaries, source texts)	L1 background L2 proficiency
Topic	Time limitation (e.g. fixed/free/homework)	L2 environment (ESL/EFL/level of school)

Language-related considerations are important in designing any corpus that aims for identifying certain linguistic issues either lexically or grammatically (Biber and Conrad, 1999; Biber et al., 2004; Biber et al., 1999; Cortes, 2004; Gardner and Davies, 2007; Hyland, 2008; Simpson-Vlach and Ellis, 2010). Whether the corpus will focus on written or spoken language is the main concern of many corpus-based studies (Biber and Conrad, 1999; Byrd and Coxhead, 2010; Carter and McCarthy, 2006; Simpson-Vlach and Ellis, 2010). Genre has also to be specified in corpus-based studies to be able to identify the differences and similarities between different genres (e.g., Swales, 1990; Flowerdew and Peacock, 2001; Hyland, 2004; 2008). Topic and style are also important in designing learners' corpus-based studies (Cortes, 2004; Hyland, 2008).

Unlike Tono (2003), Granger (2004) classified the criteria for designing computerised learner corpora into two major dimensions: learners and task settings. Learners are related to four main variables: learning context, mother tongue, other foreign languages, and level of proficiency. Task settings are also related to four main variables: timing, reference tools, exam, and audience.

Most of the previous learner corpus-based studies were conducted taking learners' level of proficiency and their mother tongue as their two main criteria(e.g., Laufer and Waldman, 2011; Nesselhauf, 2003). Advanced NNS learners were selected as they are considered "close to the end of the interlanguage continuum and are keen to move even closer to the NS norms"

(Granger, 2004: 133). Granger (2004) claims that examining advanced NNS learner corpora can help us to identify their language differences and see what needs to be taught. Cobb (2003:419) described advanced NNS learners as "not defective native speakers cleaning up a smattering of random errors, but rather learners working through identifiable acquisition sequences. The sequences are not the *–ing* endings and third person *–s* we are familiar with, but involve more the areas of lexical expansion, genre diversification, and others yet to be identified".

There are other sets of criteria for designing a learner corpus. Granger (2002) has identified four dichotomies: monolingual/ bilingual, general/ technical, synchronic/ diachronic, and written/ spoken. It appears that designing a monolingual, general, synchronic, and written corpus is easier than compiling a bilingual, technical, diachronic, and spoken corpus.

Other criteria, such as size and variability, are also considered essential in designing a balanced corpus (Biber, 1993; Biber, Conard, and Reppen, 1998; Atkins et al., 1992; Reppen, 2010; Nelson, 2010). The question of corpus size is a difficult one; corpus size is not a case of one size fits all (Carter and McCarthy, 2001). Halliday and Sinclair (1966) proposed a corpus of at least twenty million words, if it is used for exploring features of general English. The BNC, for example, consists of one hundred million words to be used to investigate various features of general English.

On the other hand, other researchers (Ma, 1993; Flowerdew, 1998) called for the use of a small corpus to explore a specific area of the language. Flowerdew (2004:19) notes that there is general agreement that a small corpus should have at least 250,000 words. Thus, the size of the corpus depends on the purpose of the research (Koester, 2010). It has been claimed that small

corpora are not suitable for research on vocabulary and phraseology, but in ESP research, small corpora can be searched for lexico-grammatical or structural features (Flowerdew, 2004, 2005).

Carter and McCarthy (1995) highlighted a number of advantages of working with a small specialised corpus over the large general corpus. First, the data are more manageable and all occurrences of the items under investigation can be examined, unlike the data in a large corpus that are unmanageable and result in the analyst having to work with a smaller sub-sample. Second, contextualised analysis can be easily examined in small corpus and insights into the lexicon-grammatical patterns of language in particular settings can be investigated. Third, with a small corpus, the corpus compiler is often also the analyst and usually has a high understanding of the context. This means that the quantitative findings can be complemented with qualitative findings (Flowerdew, 2004; O'Keeffe, 2007). A specialised corpus is often targeted to reveal contexts of use that are particularly relevant in the field of ESP/EAP.

Although compiling a large learner corpus is a major asset in terms of representativeness of the data and generalisability of the results, it has been argued that the preparation and tailoring of language samples and its subsequent corpus application are more important than the sample size (Biber, 1993; Carter and McCarthy, 2001). Nelson (2010), among others (e.g., Reppen, 2010; Biber et al., 1998), has confirmed that the size and representativeness of the corpus should be related to research questions and this will guide the design of the corpus. Since the research in this thesis focused on investigating the use and patterns of academic collocations in a specific discipline, which is CS, and since data will be analysed not only quantitatively but also qualitatively, the size of the corpora will be compiled to be specialised rather than general.

Having summarised corpus design considerations, it appears that task-related and learner-related criteria are essential (Granger, 2004; Tono, 2003). In summary, my own corpus has considered the following design features: genre, topic, style, tasks, and learner as well as size and representativeness. These features are in line with Tono (2003), Granger (2002; 2004), and Biber's (1993) recommendations. I will move on to reviewing the methodological steps used in compiling my students' corpus and the RC and outline all these individual design features and justify the choices I made.

3.2.2 Design Issues shared among Corpora

A number of issues have been considered in designing the students' and the reference corpora: representativeness, topic, style, size, genre, and learners. Each of these issues will be discussed in detail.

3.2.2.1Mapping the CS Main Domains to CS Degrees

One of the main issues was whether the CS degrees offered at the University of Essex are representative to the CS main domains. To determine this, I first browsed the school of Computer Science and Electronic Engineering at the University of Essex to find out about their taught MSc degrees. Two types of programs are offered by the department: Computer Science MSc degrees and Telecommunication and Data Communications MSc degrees.

There are six main MSc degrees under the Computer Science program: MSc Computer Science, MSc Embedded Systems, MSc Intelligent Systems and Robotics, MSc Advanced Web Engineering, MSc in Computational Intelligence, and MSc in Computer Engineering. The other program has four main degrees: MSc in Electronic Engineering, MSc Telecommunication and

Information Systems, MSc Computer and Information Networks, and MSc Computer Security[4].Then, I looked at the CS main domains in available CS corpora, such as PERC (Professional English Research Consortium Corpus) and Durrant's CS sub-corpora (2009): ten main domains were identified. These are Imaging Science and Photographic Technology, Cybernetics, Information Systems, Artificial Intelligence, Software Engineering, Hardware and Architecture, Interdisciplinary Applications, Theory and Methods, Neuroimaging, and Remote Sensing.

Due to the variations between the CS degrees offered at the University of Essex and the main CS domains identified in PERC and Durant's (2009) sub-corpus of CS, there was a need to map between them. To map between the CS degrees offered at the University of Essex with these main domains, a lecturer from the School of Computer Science and Electronic Engineering was consulted. The ten MSc degrees were mapped to just five main domains. These domains are Artificial Intelligence, Information Systems, Software Engineering, Theory and Methods, and Hardware and Architecture.

Table 3-3: Mapping the CS degrees offered at the University of Essex with PERC CS domains.

CS MSc degrees	**PERC domains**
Computer Science	Information Systems, Artificial Intelligence, Software Engineering, Theory and Methods
Embedded Systems	Artificial Intelligence
Intelligent Systems and Robotics	Artificial Intelligence

[4] For more detailed information about these programmes visit http://www.essex.ac.uk/csee.

Advanced Web Engineering	Software, Graphics, Programming (Software Engineering)
Computational Intelligence	Artificial Intelligence
Computer Engineering	Software, Graphics, Programming (Software Engineering)
Electronic Engineering	Hardware and Architecture, Theory and Methods
Telecommunication and Information Systems	Information Systems
Computer and Information Network	Information Systems
Computer Security	Software, Graphics, Programming (Software Engineering)

Two or more of the MSc degrees fall under each domain. Three of these degrees were categorised as Artificial Intelligence. These are MSc Embedded Systems, MSc Intelligent Systems and Robotics, and MSc Computational Intelligence. Three other degrees were classified under Software Engineering: MSc Computer Engineering, MSc Computer Security, and MSc Advanced Web Engineering. Two other degrees were classified as Information Systems: MSc Telecommunication and Information Systems and MSc Computer and Information Networks. The final two degrees (Computer Science and Electronic Engineering) were problematic as they were considered broad degrees and were classified under more than one domain. Therefore, they were excluded.

Having identified the CS main domains offered at the University of Essex, MSc dissertations were classified according to these domains. Three main CS domains were covered by the MSc dissertations. These are Artificial Intelligence (AI), Software Engineering (SE), and Information System (IS) (detailed information is given in section 3.4.1.1). This classification helps in identifying the main domains covered by the students' corpora. Thus, a comparable RC can be compiled from the same main domains (for more details about this corpus compilation see

section 3.5). The three corpora were comparable in terms of their representativeness of the same CS main domains, thus, topics covered in the dissertations were similar to the topics covered in the RC. Moreover, the style of the three corpora was similar as they were all compiled from written academic genres. Other design issues related to the size and learners will be covered in the next section.

3.3 Computer Science Students' Corpora

The main aim of the study was to locate academic collocations' use in postgraduate students' assignments rather than in dissertations. After interviewing two CS specialists in the field, they recommended conducting my study by investigating dissertations instead of assignments, for two main reasons. First, CS assignments do not contain enough written text. They are full of formulae and programming commands. Second, in order to access these assignments, I would have to wait until the end of each term to gather them and to obtain the students' agreement to participate, which would be time-consuming. Therefore, I decided to compile my students' corpus from previous available dissertations.

The CS student corpus was compiled from postgraduate students' writing containing about 600,000 words collected from 55 dissertations. These dissertations were divided into 29 NNS students' dissertations and 26 NS students' dissertations, having approximately 300,000 words each. No concern was given to collect equal numbers of NNS and NS dissertations. What was more important was to have the same length (number of words) for each student's corpus (Biber et al., 1998).

Taking into consideration Granger's (2004) learners' task variables, only the learning context and mother tongue for NNS students were considered during the compilation of the corpora. All NNS were native speakers of Arabic and second language speakers of English. They completed their Bachelor degrees in their own countries. They can be considered advanced learners of English since they are required to have an IELTS score of (6.0) in order to study MSc degrees offered at the University of Essex.

Regarding the task related criteria, all collected dissertations were written and produced for a specific subject from CS, and "the students are relatively free from time constraints and in most cases are expected to consult and cite data sources" (Nesi, 2008:8).

3.3.1The NNS Corpus

The School of Computer Science and Electronic Engineering at the University of Essex granted me access to three hundred dissertations. These dissertations were collected from 2009, 2010, and 2011. Most of the dissertations were in Word document format, excluding 10 dissertations that were in PDF format. Fifty dissertations were easily identified as NNS dissertations by checking writers' first names and surnames (adopted from Swales' (non-) nativeness test, 1985[5]). For example, most of the Arabic surnames started with (Al-) e.g. Al-Asiri, Al-Hindi. Thus, it was easily identified. Moreover, as I am Arabic, I could recognise the Arabic names from writers'

[5] Swales (1985) has designed a test to determine the (non-) nativeness of a research article's authors by awarding or subtracting points depending on (i) whether the author's last name is Anglo-Saxon or anglicised in some way (+/1); (ii) whether the author is affiliated with an institution in an English-speaking country (+/-3); (iii) whether all of the author's citations are to English language publications (+/-1); (iv) whether the author's first name is anglicised (+/2); (v) whether all of the author's self-citations are to English language publications (+/-2); finally, (vi) whether there is any evidence of (non-) nativeness from the article footnotes or endnotes (+/-3). If the total number of scores were (+5 to +12) it is a native speaker of English, but if it is (-5 to -12) it is a non-native speaker of English. However, two criteria (i and iv) were adopted in identifying both NNS and NS writers.

first names. After that, thirty dissertations were randomly selected from the fifty NNS dissertations using Randomizer[6].

The selected dissertations were pre-processed for analysis first by removing all unwanted parts, that is, tables, figures, formulae, references, and appendices. Quotations were also deleted since they were not considered the writer's own words (Durrant and Schmitt, 2009). Abbreviations were not deleted since they seem to be important parts of the students' writing in CS. They tend to use abbreviations many times for various kinds of names or other software programs. For example, *DB* is an abbreviation of database. An extract below shows how abbreviations are used as an integral part of the text.

The Physical Traveling Salesperson Problem (PTSP) adds a simple twist on the Traveling Salesperson Problem (TSP)…The typical ***TSP*** *is to find the shortest and cheapest way through all the cities and then return to the same city, where in the* ***PTSP*** *adds the additional factor of the salesman having 1kg in mass and moving through the force vectors.*

It is obvious from the extract that the writer mentioned the full name of the problem for the first time only and then used the abbreviations to add more information. Thus, they are integral and essential for understanding the text. After removing unwanted parts from the NNS dissertations, the number of words fell from 451,411 to 316,981. The final number of words was 301,233, after excluding one of the dissertations in the processing stage. This dissertation (NO.9) could not be converted to a text file; therefore, it was excluded.

[6] This software was used to select random samples from the data collected for the research. It is available at http://www.randomizer.org/.

The size of the NNS corpus seemed to be large enough for locating academic collocations' use in the students' writing compared to previous corpus-based studies on collocations. Durant and Schmitt's (2009)corpus of 93,868 words from NS texts and 80,298 words from NNS written texts was considered large enough to reveal the similarities and differences between NNS and NS students' use of collocations. Siyanova and Schmitt(2008) had also compared adjective-noun collocations use in a corpus of 24,500 words from Russian essays taken from ICLE with a sub-corpus of 25,000 words of NS students' use taken form LOCNESS.(For more details about these studies and other corpus-based studies on collocations see section 2.4.5).

The next stage was to map the NNS dissertations to CS degrees. Most of the NNS dissertations were identified according to their degrees; only five of them could not be identified. A friend who had finished her PhD study in CS at the University of Essex was consulted. To be certain about her decision and to be sure about her classification of the degrees of the five dissertations, a PhD student from Nottingham University was also consulted.

After identifying all NNS dissertations, they were grouped according to PERC main domains. Three main domains were covered: AI, SE, and IS. Most of the NNS dissertations were classified under IS (16 dissertations); AI and SE had a similar number of dissertations: seven dissertations were classified as AI and six were SE. Since my NNS corpus covered three main domains of CS, it can be considered a representative corpus of CS taking into consideration Biber's (1993) advice of having enough samples from the target register. Reaching this stage, the NNS corpus was now ready for processing and analysis. Having discussed the methodological issues in compiling the NNS corpus, I now move on to the NS corpus design issues.

3.3.2 The NS Corpus

Unlike the NNS corpus' quick and relatively trouble-free compilation, there were a number of difficulties involved in obtaining the required number of NS dissertations. First, the identification of these dissertations was difficult; they could not be identified as quickly as the NNS Arabic students could, since nationalities were not allowed to be given to researchers, in order to ensure confidentiality.

Thus, I tried to identify students' nationalities following different strategies. I looked at their acknowledgment and their whole dissertations for any nationality clues, but only two of the remaining 245 dissertations mentioned clues about their countries. Another strategy was applied to locate NS by their surnames (adopting one of Swales' 1985 criteria(i) to test nativeness of writers) or to search for native-like names in the names of authors in Google, Facebook, and LinkedIn websites. None of the dissertations were identified this way either. Then, the Alumni office at the University of Essex was asked for help. Though they promised to help, no dissertations were obtained.

Thus, another strategy was used: a native speaker senior lecturer from the International Academy department was consulted. As she was a NS, she could easily identify NS students by their first names and surnames (adopting two criteria (i and iv) of Swales' (1985) nativeness test); however, only eight names were identified in this stage. Since none of the previous strategies was effective enough to confirm the nationalities of the students, I returned to the School of Computer Science and Electronic Engineering for help. After I explained to them the importance of this identification for my research and the difficulties I had faced in identifying them, they

agreed to provide me with the NS students' list of names for the last three years. Unfortunately, only 22 NS dissertations were identified, as most of the postgraduate students in this department were NNS.

To complete the compilation of the rest of the NS corpus, an email asking for MScs in CS dissertations written by NS was sent to three other UK universities: Sheffield University, the University of Leicester, and the University of Nottingham. Two of these universities responded and provided access to NS dissertations. Two NS dissertations were recommended by the head of department from Sheffield University and were downloaded from their website, which had full lists of their students' dissertations from 2001 to2012[7]. The head of department at the University of Leicester supplied another four dissertations, after a confidentiality letter had been signed. As a result, the final number of the NS dissertations was 28: 22 from the University of Essex, two from Sheffield University, and four from the University of Leicester.

Having collected the required number of NS dissertations, the same procedure of identifying the domains of NNS dissertations was followed. Most of the NS dissertations were identified by their topics or their degrees; only seven of them could not be identified. Therefore, a specialist CS from the School of Computer Science and Electronic Engineering was contacted for help. Two of these dissertations were identified under other CS domains other than AI, SE, and IS; thus, they were excluded from the NS corpus. The 26 identified dissertations were all processed following the same procedure in processing the NNS corpus. The NS corpus consists of 294,362

[7] For more information visit http://www.shef.ac.uk/dcs/research/publications/studis.

words. Table 3-4 presents the number of NNS and NS dissertations in the three selected sub-disciplines of CS.

Table 3-4: Number of dissertations and number of words for NNS and NS corpora.

Corpus	**No. of dissertations**	**No.of words**	**AI**	**IS**	**SE**
NNS	29	301233	7	16	6
NS	26	294362	8	6	12

3.4 Reference Corpus Compilation

3.4.1 Selecting Journals for the Reference Corpus

To compare students' use of the most frequent AWL, a RC is needed. My first plan was to use an available specific CS corpus as my RC, as advised by Reppen (2010) and Nelson (2010). Two specific corpora were considered: PERC and Durrant's corpus of academic collocations (2009). However, due to practical reasons, which will be outlined below, I decided to compile my own RC.

First, PERC[8]was considered. It consists of a 17million word corpus of English academic journal texts in 22 subject fields.Each subject has a balanced 1 million corpus of each. Although it has a sub-corpus of a million words compiled from CS research articles, a problem arose after

[8]To find out more about this corpus visit http://www.perc21.org/corpus_project/index.html and http://scn.jkn21.com/~perc04/.

checking the user interface on the PERC website. Their user interface does not allow sub-corpora collocation searching but only in the PERC corpus as a whole. Therefore, Durrant's sub-corpus of CS (2009), which was developed from top research articles in the field, was then considered as an alternative RC. From this large corpus of 25 million words, the CS sub-corpus consists of approximately 600,000 words compiled from 67 research articles selected from the six main CS domains. Due to the restriction of copyright permission, I could not access Durrant's sub-corpus of CS.

Thus, I decided to compile my own reference corpus. Research articles were chosen for this purpose for a number of reasons. First, they are easily accessed and downloaded in electronic formats, unlike textbooks and other written sources that require considerable effort to be scanned and converted to electronic form. Moreover, research articles, as Hyland (2008: 47) notes, "are often the target of good writing that students are encouraged to emulate and are the most comparable to student writing". I can argue that research articles are (for most disciplines) the most prestigious form of academic writing and they are more analogous in their aims and structure to student writing than other forms of professional academic prose (e.g. textbooks).They would seem to provide the best available model of 'target language' for students of EAP (Hyland, 2008).

Therefore, a corpus based on research articles may be more representative of the language students should be aiming to acquire than a more broadly based sample would be. Even though a corpus of distinct dissertations written by NS could be used as a RC, no attempt was given to make such a corpus since it was difficult to access the grades of the students and, if so, too few distinct dissertations might be found.

3.4.2 Building the Reference Corpus

Since identifying collocations requires a large corpus (Halliday, 1966), and as this corpus was to be compiled by a single researcher with limited resources and within the limited time-scale permitted by this thesis, I decided to compile a RC of approximately the same number of words as in the combined students' corpora (approximately 600,000 words). Moreover, since Durrant's CS sub-corpus was approximately 620, 000, a corpus of approximately 600,000 words was considered large enough for my study.

The same three main CS domains located in the students' corpus were used to build my RC. To compile a balanced RC, a sub-corpus of approximately 200,000 words was aimed at for each sub-domain. Nativeness of writers was not considered an important factor in my selection of the articles, since Durrant notes, "Academic language was presumed not to have any native speakers and to exist somewhat independently of national linguistic varieties" (2009: 192). Therefore, no attempt was made to distinguish between writers from different L1 backgrounds or between journals using British, American, or other forms of English.

My plan was to select articles from the top three high impact factor journals for each sub-domain. Using the *ISI Web of Knowledge* database (http://portal.isiknowledge.com/portal.cgi), which provides listings of journals under disciplinary headings ranked according to the journal's contribution to scholarly communication, full lists of the names of the highest impact factor journals were provided. Browsing the three top journals in each of the three selected domains, some journals were excluded. For example, the third high impact factor journal for IS was *IEEE Communication Tutorials and Surveys*, which includes articles of different sections from the sections of research articles in CS. Thus, the fourth high impact factor journal (*Transactions on*

Information Systems) was selected instead. Research articles were downloaded from the final two volumes of the last two years (2011 and 2012). The following criteria were followed in selecting articles:

a- No state of the art and review articles were included.

b- Articles that were replies to other articles and articles that are criticisms to some articles were excluded. These critical and reply articles are too brief and they can be classified under different genre, thus, they were excluded.

c- Articles that focus on the Business, Psychology, or Sociology side of IT were excluded because they were more related to other disciplines than to CS. For example, an article from Vol. 36(2) of *MIS Quarterly (The career paths less (or more) traveled: a sequence analysis of IT career histories, mobility patterns, and career success)* was excluded because it summarised the career path of IT graduates, which is a sociological investigation, as shown by the quotation taken from the abstract.

This paper examines the objective career histories, mobility patterns, and career success of 500 individuals drawn from the National Longitudinal Survey of Youth (NLSY79), who had worked in the information technology workforce… Of the 500 individuals in the IT workforce, 173 individuals pursued IT careers while the remaining 327 individuals left IT for other high-status non-IT professional jobs in PLM or lower-status, non-IT jobs in SLM careers.

d- Articles called ‘Research note’ (instead of ‘Research article’) were excluded because they were briefer than research articles.

First, 30 research articles were downloaded for each sub-domain, 10 from each journal. The next step was to convert them into Word files: Using Nitro PDF 8 software, which was available for a free two-week trial (downloaded from http://www.pdftoword.com), all research articles were

converted to Word documents. Then, the cleaning step began: all unneeded parts such as graphs, tables, formulae, references, and appendices were deleted.

Since different journals have different word limits, I aimed to compile my sub-corpora of approximately the same number of articles (30 articles), but that was unsuccessful since different journals have different words limits. Therefore, I tried to compile them so that they would include approximately the same number of words (approximately 200,000 words), as suggested by Biber et al. (1998). After cleaning 63 articles from the three selected domains, a corpus of 600,269 words was compiled. These articles were divided between the three sub-corpora as follows: 26 AI (200,375), 18 IS (200,838), and 19SE (199,056).Table3-5 presents full information about journals selected for each domain and the number of articles selected from each journal (see Appendix A for a full list of references of research articles selected).

Table 3-5: The three high impact factor journals selected for compiling the RC from the three selected CS sub-disciplines

AI selected journals	**Impact factor**	**SE selected journals**	**Impact factor**	**IS selected journals**	**Impact factor**
1-IEEE Transactions on Pattern Analysis and Machine Intelligence (10)	5.3	1-IEEE Transaction on Visualisation and Computer Graphics(7)	4.8	1-*MIS Quarterly*(7)	5.0
2-International Journal of Intelligent Systems(6)	5.1	2-ACM Transactions on Graphics (TOG) (6)	4.5	2-Enterprise Information Systems (4)	4.3

3-IEEE Transactions on Evolutionary Computation(10)	4.4	3-ACM Transactions on Software Engineering and Methodology (6)	4.2	3-ACM Transactions on Information Systems (7)	4.0
26 articles (200,375 words)		19 articles (199,056 words)		18 articles (200,838 words)	

3.5 Problematising the Study

All methodologies have their weaknesses; the corpus-based approach that I have adopted here is no exception. Widdowson (2000) details these weaknesses:

> "[Corpus linguistics] can only be one aspect of what they do that is captured by such quantitative analysis. For obviously enough, the computer can only cope with the material products of what people do when they use language. It can only analyse the textual traces of the processes whereby meaning is achieved; it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted. It cannot produce ethnographic descriptions of language use. [...] [Corpus analysis] is necessarily only a partial account of real language" (pp.6-7).

To these I can add some other common objections to corpus methodologies outlined by Stubbs (2001), even though he is a corpus enthusiast. He mentions the familiar complaint that corpora are by definition unrepresentative, since they "cannot represent a whole language" and are "merely a collection of what it is convenient to collect" (p.223). Another complaint is that corpora only provide positive data: "a corpus can reveal only what does occur and not what cannot occur" (p.224).

Although the weaknesses Widdowson (2000) and Stubbs (2001) describe are formidable, one of Widdowson's objections can be dismissed immediately, not only with regard to my own study

but with regard to also the vast majority of corpus studies being carried out to date. Widdowson seems to assume that corpus linguists only conduct quantitative enquiry, when in fact the analysis in this thesis, and indeed in all other studies described in Chapter 2, takes a combined quantitative/qualitative approach. The other objections will be addressed in turn in the remainder of this section.

Regarding the representativeness of the corpus, what matters is not the size of the corpus but the representation of a sample of the linguistic issue under investigation. The important question is 'have I made my corpus as representative as can reasonably be expected?' I have tried to answer this question throughout this Chapter, describing how both my student and reference corpora were compiled to represent the academic collocations used in the three selected CS sub-disciplines, taking into account the issue of sub-disciplinarity, addressing the native/non-native speaker issue, and compiling the RC of the same size as the students' corpora. Obviously, I would prefer my corpora, especially the RC, to have been far larger, but a researcher working alone for a limited period cannot achieve so much. Reppen (2010) and Biber et al. (1998) have confirmed that time constraints is an essential factor in building a representative corpus. Moreover, corpus size does not necessarily guarantee representativeness, principle, or suitability.

The final complaint that Stubbs (2001) mentions is that corpora only describe what occurs, not what does not occur. For the purposes of my study, this means the fact that some collocations do not exist or are underrepresented in one or both of the student sub-corpora when compared to the RC does not mean that the students do not know them or do not know how to use them. Perhaps they were not relevant to their dissertation topics, so they did not need to use them. To concede the limitation of the corpus-based approach, other mixed methods were applied: a categorisation

judgement task and specialist interviews were carried out with CS experts to investigate the use of academic collocations and their patterns in details (for more information see section 5.4.).

3.6 Processing the Students' Corpora

All cleaned students' Word files were converted to text files to begin locating the AWL using Antconc[9].First, the word list of each student corpus was developed. To locate the AWL in each wordlist, I followed Laurence Anthony's steps of generating a specific list from the general wordlist[10].This was done by selecting the AWL as the specific wordlist and adding it to the general wordlist. A list of approximately 1,000 academic words was developed. Following Durrant's (2009) procedure of selecting the most frequent words (see section 2.4.5.4), the 100 most frequent AWL were located. (For the full list of the 100 most frequent words from the AWL, see Appendix B).

A problem occurred with identifying the POS of some of the words. A number of words can be identified as noun or verbs e.g. *affect.* The main aim of locating the most frequent academic words is to discover which POS is the most used by CS postgraduate students and, therefore, locate their collocations. Thus, POS tagging was needed to avoid POS misclassification of the words and to save time. Using the free CLAWS tagging facility (available at http://ucrel.lancs.ac.uk/claws/trial.html), all 55 students' text files were tagged. After that, tagged wordlists were developed for each sub-corpus using *ConcGram* (Greaves, 2005), a

[9] Antconc is free concordance software available at http://www.antlab.sci.waseda.ac.jp/software.html.
[10] Steps of locating academic words from general wordlists explained by Anthony Lawrence were followed in his tutorial 8 of Antconc 3.2.4: word list tools: basic features. For more information visit http://www.youtube.com/watch?v=Zb71yaBP_lI&hd=1#!

concordancing software specialising in multiword expression identification (Cheng et al., 2006; Cheng et al., 2009).

A number of words were tagged wrongly; the main reason was the style of the students' writing. Some words were classified as single words when they were actually two words (e.g. *listedamong* that should be *listed among*).Other words were hyphenated and divided into syllables (e.g. *spe-cific* that should be *specific*).Words that contain slashes were also wrongly tagged (e.g. *up/down*). Thus, manual checking of all tagged wordlists was carried out. 550 words were tagged wrongly in the NNS wordlist and 750 words in the NS wordlist. These words were corrected using the following set of rules:

1- Add a space if two words were tagged together as one e.g. *listedamong*. It was corrected to *listed among.*
2- Add two spaces before and after slashed words. E.g., *up/down* is corrected to *up / down*.
3- If the word is hyphenated, the hyphen is deleted to make one word. E.g., *Spe-cific* was corrected to *specific*.
4- If two words had a full stop between them, a space was added after the full stop. E.g., *direction.the* was corrected to *direction. The*.
5- If words were typos, they were not corrected.
6- If words were computer jargon, they were not corrected.
7- If words were hyphenated adjectives, the following steps were followed:

    A) The word was checked in context; B) if it really was an adjective, a dictionary and online browser (Google search) were searched to see which spelling is the most frequent; C) the most frequent spelling was used.

For example, the word *feed-forward* was categorised as an adjective. I first double-checked the context to find whether it was used as an adjective.

According_PRPto_PRPthe_ATKprogramming_NNWcode_NNWabove_AVK,
*the_ATKcommand_NNWis_VBZused_VVNto_TO0create_VVIa_**ATKfeed forward_AJK**back_NNWpropagation_NNWnetwork_NNW_.*

It was clear that it is an adjective. Thus, the next step was to consult a dictionary and Google to check whether *feed-forward* is written with or without a hyphen. From the dictionary check, *feed-forward* is "The modification or control of a process using its anticipated results or effects" (OED online at http://www.oxforddictionaries.com/): it was a noun modifier, written without hyphen. Thus, the hyphen was deleted.

D) In some cases, students made their own hyphenated adjectives because they can express the meaning s/he intends in a fast way (e.g. *easy-to-use*). These hyphenated adjectives should not be divided into separate words because they then lose their meaning. If it is not actually an adjective, this means that CLAWS made a mistake because of the hyphen, so hyphen/s should be deleted.

These error corrections were rechecked by one of my supervisors to increase the reliability of the corrections. The following stage was to re-insert all corrected files in CLAWS for tagging. The new-tagged wordlists were used to locate the most frequent academic words and their collocations in students' corpora and in the RC (see section 4.5.2 for detailed steps of locating collocations).

3.7 Conclusion

This Chapter first presented the overview methods used in this thesis since a number of methodologies were applied to answer the main research questions addressed in this thesis. It also presented the compilation of the students' and the reference corpora in detail. Steps of designing and processing both students' and reference corpora were fully presented. As students' and reference corpora were designed and processed, the next step is to locate academic collocations in students' corpora and to compare their uses with the RC, which will be covered in the following Chapter.

Chapter 4 The use of Academic Collocations by Non-expert CS Postgraduate Students

4.1 Introduction

This Chapter introduces the first study related to the use of academic collocations in the postgraduate CS students' writing. For this purpose, the frequency-based approach is used in locating the most frequent academic collocations in the students' corpora and the reference corpus. The Chapter will first introduce the most widely used statistical methods developed for identifying collocations, that is, raw frequency, t-score, and mutual information (MI) in section 4.2. Then, sections 4.3 and 4.4 will review previous research that adopted frequency-based methodology. Section 4.5 will present the methodology applied in identifying collocations in this study. Section 4.6 will present the results of the first three research questions related to this study (mentioned below). Finally, section 4.7 will discuss the main findings.

RQ1: What are the most common academic collocations used by Computer Science students in their MSc dissertations?

RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus?

RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?

4.2 Frequency-based methods of identifying collocations

4.2.1 Raw Frequency

"The simplest frequency-based method of identifying collocations is to count the number of times combinations of words occur" (Durrant and Doherty, 2010: 6). Thus, finding that *strong tea* occurs in the BNC 28 times, while *powerful tea* appears only three times, "we may conclude that the former is the more conventional collocation" (Manning and Schütze, 1999: 162-163). Durrant and Doherty (2010) considered this approach problematic, as it does not locate the most frequent collocations only but also the most frequent regular combinations, where co- occurence of words comes about by chance. Thus, it cannot be applied in locating frequent collocations individually.

Stubbs (1995) observed that frequency of co-occurrence is not enough to identify collocations; hence the need for measures of association strength. These measures are based on the assumption that observed frequency (*O*) of a pair of words can be compared to its expected frequency (*E*) in a random hypothetical corpus (Stubbs, 1995). The *(O)* refers to the real number of co-occurrence of a pair of words in a corpus, while the *(E)* refers to the expected frequency of occurrence on the null hypothesis that is no relationship between the words (Durrant, 2009). Expected frequency serves as a reference point for the interpretation of O: The *O* of a pair should not be higher than its *E*. It is rejected if *O* is significantly higher than *E*.

Evert (2008, p. 17) gave an example of the word pair *is to* which is very frequent in the Brown Corpus (Kucera & Francis, 1967) but is not considered a collocation since its observed frequency

in the corpus (*O* = 260) is equal to its expected frequency. The expected frequency of a word pair is calculated using the formula:

$E = f1\ f2\ /N$ (Evert, 2008:18)

Where *f1* stands for the frequency of the first word component in the corpus, *f2* for the frequency of the second word, and *N* for the corpus size. Thus, the expected frequency of the pair *is to* in the Brown Corpus is*:*

E (is to) = 10,000 *26,000 / 1,000,000 = 260

Many formulae have been developed to calculate strength of association based on the expected and observed frequency of a pair in a given corpus (see Evert, 2008 for an overview). These methods can be generally grouped into two main types: hypothesis testing techniques and measures of strength, primarily mutual information (MI). The two types of technique are conceptually different and typically produce rather different types of results (the rationales of these methods will be discussed in the following sections). While hypothesis-testing techniques locate collocations that are unlikely to arise by chance, the MI score measures the strength of association between the components of the collocation. They also differ in terms of the words included in the collocations. Collocations located via MI tend to include infrequent words whereas collocations located via hypothesis testing measures tend to contain frequent words. I shall deal with each in turn.

4.2.2 Hypothesis Testing Techniques

The main hypothesis testing methods of identifying collocations are the *z-score, t-score, chi-squared* and *log-likelihood* tests (Sinclair et al., 2004; Seretan, 2011; McEnery and Hardie, 2012; Barnbrook et al., 2013). These tests check the null hypothesis that the observed frequency (*O*) of a pair is not higher than its expected frequency (*E*). It is rejected if *O* is significantly higher than *E*. Durrant and Doherty(2010) suggested that these hypothesis tests can be seen as formalisations of Hoey's definition of collocations as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991:7). The aim of the hypothesis testing methods is to determine the statistical significance of this apparently greater than chance frequency (Manning and Schütze, 1999: 162-163). These techniques are presented in more detail in a number of publications (Manning and Schütze, 1999; Evert, 2004). For the purpose of the literature survey only t-score, the prominent hypothesis testing method, will be presented and compared with MI below.

4.2.3 Mutual Information and t-score

The MI score quantifies the strength of association between the components of the collocation. MI can be conceptualised as a "measure of how much one word tells us about the other" (Manning and Schütze, 1999:178). In other words, when I encounter one member of a word pair that has a high MI score, I can predict that the other member of the pair is likely to be nearby. The t-score, on the other hand, is a measure of certainty of a collocation. The former is more likely to give high scores to infrequent collocations whereas t-score will yield high scores for relatively frequent collocations, provided they occur even more frequently than expected.

As an illustration of the difference between MI and one hypothesis-testing measure (i.e., t-score), let us consider the pair *heavy rain* in the British National Corpus (BNC) (Davies, 2004). The pair occurs 225 times in the BNC (*O*= 225). The frequency of the word form *heavy* is 9,125 (*f1*) and that of the word form *rain* is 6,253 (*f2*), so the expected frequency of the pair in the BNC (with a total size of 100 million) is:

E (*heavy rain*) = 9,125 * 6,253 / 100,000,000 = 0.57

Given the observed frequency and the expected frequency of the pair, I can now calculate the MI score and t-score according to the following formulae:

MI=log O/E

t-score = (O – E) / √O (Evert, 2008:18)

Thus, for the pair *heavy rain*, the values are:

MI (*heavy rain*) = $\log\frac{225}{0.57} = 8.62$

T-score (*heavy rain*) = (225 - 0.57) / √225 = 14.96

Although the two scores are different, they are both far higher than the required threshold level for 'strong collocations': 3 for MI and 2 for *t*-score (Hunston, 2002a: 71-72). Other clear examples are taken from the RC to demonstrate the MI and t-score clear-cut off. The combinations *'access counts'* was not considered a collocation since t-score =12.5 while MI =2.5. On the other hand, *'data access'* (t-score= 5.7, MI = 14.7) was considered a collocation since it met Hunston's (2002) required threshold level of strong collocations.

Thus, they are quite different, as noted by Clear (1993:279-282), in that "MI is a measure of *the strength of association between two words*", whereas hypothesis-testing methods are measures of

"*the confidence with which we can claim there is some association*". Similarly, Evert (2008: 22) summarised the difference between the two measures well, referring to MI as a measure of 'effect size' and to t-score as measure of 'significance':

"The former [effect size measures] ask the question "how strongly are the words attracted to each other?" (Operationalised as "how much does observed co-occurrence frequency exceed expected frequency?"), while the latter [significance measures] ask, "How much evidence is there for a positive association between the words, no matter how small effect size is?" (Operationalised as "how unlikely is the null hypothesis that the words are independent?"). The two approaches to measuring association are not entirely unrelated: a word pair with large "true" effect size is also more likely to show significant evidence against the null hypothesis in a sample. However, there is an important difference between the two groups. Effect-size measures… are prone to a low-frequency bias (small E easily leads to spuriously high effect size estimates, even for O = 1 or O = 2), while significance measures are often prone to a high-frequency bias (if O is sufficiently large, even a small relative difference between O and E, i.e. a small effect size, can be highly significant)".

Four final points are worth noting about all the frequency-based approaches to defining collocations. First, the approach treats collocations as symmetric units (assuming that the two words comprising the collocation are equally predicted by each other), which is often not the case. Second, although it was claimed above that, there is a specific threshold for each type of association measure; Stubbs (1995) noted that this is an arbitrary decision. Similarly, Evert (2008) claimed that the significance threshold is important when I need to distinguish 'true collocations' from 'non-collocations' but not when the notion of collocation is viewed as a cline

from ‘weak’ to ‘strong’ pairs. However, Durrant and Schmitt (2009) and Siyanova and Schmitt (2008) adopt the significance threshold in their identification of true collocations.

Third, various decisions are important in extracting collocations from a corpus. These include span size (i.e., the number of words to be considered to the right and left of the node, often set between three and five words), word type (whether individual lexical units are defined as word forms, lemmas, or word families, often defined as lemmas), and raw frequency thresholds (the minimum number of occurrences in the corpus for a pair to be considered a potential collocation, often set between three and ten occurrences). Finally, it should be noted that the frequency-based approach is criticised for resulting in linguistically uninteresting combinations such as ‘*children-toy'*, which frequently co-occur based on real world connections rather than any linguistic attraction (Hunston, 2002a:68). Thus, it is important to proceed cautiously when using collocational statistics (Stubbs, 1995; Coxhead and Byrd, 2012).

Consequently, Evert (2008), among various scholars (Bartsch, 2004; Clear, 1993; Stubbs, 1995), stressed the need to combine various measures to compensate for their limitations and the necessity of including a raw frequency threshold for MI to cancel out its low-frequency bias. Hunston (2002a) recommended the use of both MI and t-score to locate strong collocations. Thus, these two measures were selected for locating academic collocations in the NNS and NS corpora in this thesis for two reasons. First, MI and t-score are considered the prominent statistical methods for identifying strong collocations. Second, *ConcGram*, the software I used for locating collocations, employs MI and t-score as the main measures for identifying collocations.

4.3 Previous frequency-based collocation studies

A number of researchers have located NNS and NS students' collocations using the frequency-based approach described above in the EAP context (Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008; Laufer and Waldman, 2011; Granger, 1998) and in the ESP context (Ward, 2007; Gledhill, 2000a). They mainly used raw frequency, MI, and t-score to locate strong collocations. However, since the frequency-based approach has some limitations, comparative approaches were applied such as comparing the frequency of the located collocation in another large corpus (Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008) or they applied non-statistical methods such as checking their existence in collocation dictionaries (e.g., Laufer and Waldman, 2011).Thus, identifying strong collocations is only the first step in locating collocations. The verification of the located collocation is the second necessary step. Each step will be described in turn in the next section.

4.3.1 Identifying strong collocations

Collocations can be extracted either manually or automatically. A number of researchers located potential collocations manually from their learner corpora and then applied the MI and t-score measures to identify strong collocations. For example, Durrant and Schmitt (2009) extracted pre-modified noun collocations (adjective-noun, noun-noun collocations) manually and then inserted them into *Wordsmith Tools* to locate strong collocations using MI and t-score. Similarly, Siyanova and Schmitt (2008) located the most frequent adjective noun collocations used by NS students using both raw frequency and MI. By contrast, Laufer and Waldman (2011) identified verb-noun collocations in a NS student corpus and then located their equivalents in a NNS corpus.

With software like *ConcGram*, collocations can be extracted automatically. Using the text retrieval software TACT, Granger (1998) automatically located all amplifier-adjective collocations from the NS and NNS learner corpora and then manually sorted amplifiers according to certain semantic and syntactic criteria into maximisers (e.g. *absolutely)* and boosters (*highly*).

4.3.2 Verification of collocations

Two approaches are applied to verify the existence of the collocations located in a learner corpus: use of a reference corpus and dictionary checks. Durrant and Schmitt (2009) calculated the strength of their extracted collocations by comparing their frequency with that in the British National Corpus (BNC), which contains 100 million words. They assumed that since the BNC is one of the largest and most representative corpora of general English currently available, collocations that occur frequently in it have common usage in English. Similarly, Siyanova and Schmitt (2008) consulted the BNC to determine the frequency and MI of each NNS and NS collocations.

The second approach involves the use of two general collocation dictionaries to check the existence of the located collocations. For example, Laufer and Waldman (2011) checked their NS and NNS students' verb-noun collocations in two dictionaries: *The BBI Dictionary of English Word Combinations* (Benson, Benson, and Ilson, 1997) and *The LTP Dictionary of Selected Collocations* (Hill and Morgan, 1997). If the verb-noun collocation was listed as a collocation in

either one of the dictionaries, it was accepted as a collocation. A similar procedure of verification of collocations was used by Nesselhauf (2005) and Wang and Shaw (2008).

After verifying the located collocations, some researchers further analysed the collocations by categorising them into bands using association measures. Durrant and Schmitt (2009) used both MI and t-score to classify their collocations into bands. The extracted collocations were divided into seven bands of *t*-score, as follows: (t=2−3.99; *t*=4−5.99; *t*=6−7.99; *t*=8−9.99; *t*=10−14.99; *t*=15−19.99; *t*≥ 20). Similarly, the MI scores were divided into the following bands: (MI=3−3.99; MI=4−4.99; MI=5−5.99; MI=6−6.99; MI=7−7.99; MI=8−8.99; MI=9−9.99; MI≥10). The two kinds of categorisation were not aligned to each other. Unlike Durrant and Schmitt (2009), Siyanova and Schmitt (2008) classified their collocations into five bands using MI only as follows: (0 (failed to appear in the BNC), 1–5, 6–20, 21–100, and >100 occurrences).

4.4 Frequency-based approach of locating collocations in a single genre

A number of researchers have located collocations from research articles either in different disciplines (Ackermann and Chen, 2013; Peacock, 2012) or in a single discipline (Ward, 2007 in Engineering and Gledhill, 2000a in Medical research articles). A similar procedure to collocation identification in learner corpora has been carried out in locating collocations in a specific discipline, with the exception of verifying collocations from a reference corpus. Contextual analysis was vital for understanding the function of the collocations in their located register (Ward, 2007; Gledhill, 2000a; Peacock, 2012).

In his investigation of the most frequent collocations of the five most frequent grammatical words(*has, have, is, of, at*) in his corpus of 120 research article introductions, Gledhill (2000a) applied contextual analysis to gain better understanding of the function and register of these grammatical collocations. For example, the contextual analysis of *is* reveals a limited set of items that can introduce noun-predicate clauses. The following clause is always a biochemical fact. The subject noun varies from empirical to research oriented terms and usually involves explicit evaluation (here underlined). Here are some examples taken from Gledhill (2000a:117):

The most direct evidence is that coagulation factors
A simple explanation is that none of these is currently in use
The expectation is that PTC apparently does not show mutagenesis
An intriguing observation is that these compounds are t-promoters

Although this study looked at grammatical collocations, which are outside the scope of this thesis, I believe manual checking of contexts of collocations for further insights into how they may be used differently in different corpora is valuable.

Peacock (2012) first located the 16 most frequent nouns in 320 research articles across eight disciplines: Chemistry, Computer Science, Materials Science, Neuroscience, Economics, Language and Linguistics, Management, and Psychology using *Wordsmith Tools* and then located their most frequent collocations using MI. To investigate disciplinary variation, the corpus was split into disciplines and context was checked manually.

By contrast, Ackermann and Chen (2013) applied various steps to locate academic collocations from different sub-disciplines. Using both quantitative and qualitative approaches, they first located academic collocations using both MI ≥3 and t-score ≥2 and then filtered the located academic collocations using POS tagging to select noun collocations only. After that, the academic collocations located underwent a qualitative review in which each collocation was assessed independently by the two researchers to determine whether a specific collocation should be included, discussed, or excluded from further analysis. Then the remaining 4,558 collocations were subjected to the expert review to evaluate their developed lists of noun collocations.

In the present study, following Ward (2007) and Peacock (2012), only the 100 most frequent academic words (Coxhead's AWL) used by NNS and NS students in their corpora were selected to be searched for their collocations. Using both MI and t-score, *ConcGram* extracted academic collocation lists from both NNS and NS students first and then verified their existence in the RC. Since the present study focuses on lexical collocations, the next step was to manually extract lexical collocations and then to categorise them according to their specific uses in the three selected sub- disciplines of CS (AI, SE, IS) using two dictionaries. These steps will be presented in detail in the following section.

4.5 Collocation Identification in My Study

In our study, academic collocations were located in five stages. First, the 100 most frequent academic words occurring in the students' corpora were extracted based on Coxhead's (2000) AWL. Second, collocations in students' and reference corpora were located and then compared between each student corpus and the reference corpus. Third, lexical collocations were extracted

from the generated lists of collocations manually. Fourth, the 100 most frequent N and V collocations from each of the students' corpora were tested for their significance and, fifth, two dictionaries were used for checking and categorising significant N collocations in terms of their specificity to CS sub-disciplines. Each stage will be described in turn in the following sections.

4.5.1 Extracting the 100 most frequent academic words from students' corpora

After tagging all students' files for POS using CLAWS, each of the NNS and NS files were merged. The 100 most frequent academic words used by NNS and NS students were located in two stages.

The first stage focused on developing lists of the most frequent academic words used by NNS and NS students. Coxhead's (2000) lemmatised list of AWL families was selected for this purpose. It has been noted by Stubbs (2002) and Evert (2004) that the grouping of all inflected forms (types) under the same lemma is more likely to lead to significant statistical results and helps in detecting strong collocational associations more easily. Stubbs (2002: 82-83) discussed the example of the word *resemblance*, whose collocates in a corpus, and in particular with the verb *bear*, are scattered through different forms. Put together, these forms make up a high proportion of the total number of collocates. This is clearly shown in the following example taken from Stubbs (2002: 82-83):

Resemblance 1.08 %< bears18%, bear 11%, bore 11%, bearing 4% > 44%.

Thus, the lemmatised list of AWL families was inserted as the specific list in relation to which the students' academic wordlists were extracted. For example, the noun *network*, *networks* and *networking* were searched together as one type of AWL families (see appendix C for more examples). By running the 'wordlist function' in *Antconc*, academic wordlists for each student corpus were generated using the lemmatised list of AWL families. 503 out of the 570 AWL families were found in the NNS corpus and 507 were found in the NS corpus. Only the 100 most frequent academic word families were selected for the next phase.

Rather than studying all members of the 100 word families selected, this current study focuses on the most frequent member of each word family, following Coxhead and Byrd's (2012) procedure for identifying the most frequent member of the family. Where two words in a family have similar frequency, the study presents information about both members of the family. However, in most cases one member of such sets is much more commonly used and is thus the focus of my study.

The second phase was carried out by checking and identifying which POS was prominent for the 100 academic word families located. For example, *focus*, which can be categorised as N or V, was checked in the NS corpus to determine whether it was used more frequently as N or V. It was used as N (62) times, while it was used as V (18) times; thus, it was counted as one of the most frequent nouns in the NS corpus.

Nouns and verbs were the most frequent POS in the 100 most frequent academic words located. Therefore, two lists were developed for each student corpus; the first list was for the most frequent nouns and the second list was for the most frequent verbs. For more details about these

N and V lists, see Appendix C. Table 4-1summarises the number of AWL word families for which nouns and verbs were the most frequent forms from each student corpus.

Table 4-1: Summary of the AWL word families and their specific members (noun and verb word forms) that formed the nodes for the collocation search in the NNS and NS corpora

	AWL word families for which noun forms were the most frequent	**AWL word families for which verb forms were the most frequent**	**Total number of frequent word families**
NNS corpus	68	20	88
NS corpus	62	26	88

As can be observed from Table 4-1, the total number of AWL word families selected from each student corpus for the study was similar. These 88 word families were used first in locating academic collocations in students' corpora and then to compare the located collocations with the reference corpus.

4.5.2 Locating collocations

At this stage, 'collocation' was defined as a node word and the word that co-occurs within the span of three words, co-occurring at least five times in total with MI score of at least 3 and a t-score of at least 2. Since Hunston (2002a: 75) noted that a collocate with an MI score of at least 3 and a t-score of at least 2 is considered "a strong collocate, and a certain one", the present study will analyse collocations using both MI and t-score applying the values recommended by Hunston (2002a) as conditions for locating strong collocations.

In order to locate collocations for the 88 AWL word types focused on in this study, *ConcGram* (Greaves, 2005), a special program for locating collocations, was used. This software is useful for calculating the significance of collocations in context. It locates all of the contiguous and non-contiguous words, including both constituent (AB, ACB) and positional (AB, BA) variations. *ConcGram*" generates t-score and MI to help to decide the significance cut-offs for ConcGram lists, and to provide the user with indications as to which word co-occurrences are more likely to prove to be meaningful, and which ones the user can afford to ignore" (Cheng et al., 2006:8-9).

The automatic nature of the search is time-efficient and reliable in retrieving all possible permutations that may otherwise be difficult and cumbersome to find manually. Concgramming is efficient, but its automatic nature needs to be complemented with manual analysis in order to focus on strong collocations and to avoid focusing on grammatical words, which frequently co-occur with contiguous and discontiguous collocations (Yuldashev, Fernandez, and Thorne, 2013).

Two main steps were followed to locate academic collocations used by students: first, academic collocations for the most frequent academic nouns and verbs were located in students' corpora as well as in the reference corpora applying the same criteria.

Three criteria were set for locating noun and verb collocations in the three corpora:

1- Both MI≥3 and t-score≥2 were applied.
2- Collocations were located using a span of three words from both sides of the node word.

3- N and V lists developed in stage 1 (section 4.4.1) were used as the wordlists in *ConcGram* in order to locate collocations of these nouns and verbs.

Second, noun and verb collocations located in the students' corpora were searched in the reference corpus for verification of collocations. Each step will be presented in detail below.

4.5.2.1 Locating academic collocations in the students' and reference corpora

Applying the aforementioned three criteria, N and V collocations were located first in students' corpora. Four lists of collocations were developed, two for each corpus. Table 4-2 shows the number of tokens of collocations for both nouns and verbs in each of the students' corpora.

Table 4-2: Number of tokens of N and V collocations in NNS and NS corpora

	NNS N collocations	**NS N collocations**	**NNS V collocations**	**NS V collocations**
Number of tokens of collocations	872	1608	258	591

Regarding the reference corpus, the same procedure was applied with the exception that four lists of collocations were extracted: two lists were created using the academic nouns and verbs that I chose as nodes in NNS corpus and two lists were created in the same way for the NS corpus. In this way, parallel sets of N and V collocations were located in RC for each student corpus, including many collocations that were not used in students' corpora. Table 4-3 presents the number of tokens of noun and verb collocations located in RC.

Table 4-3: Number of tokens of N and V collocations located in RC

	NNS-N-RC collocations	**NS-N-RC collocations**	**NNS-V-RC collocations**	**NS-V-RC collocations**
Number of tokens of collocations	5843	6282	1454	1648

After locating N and V collocations in the three corpora, the next step was to compare between the collocations located from the students' corpora and those located for the same nodes in the RC to determine the shared set of collocations.

4.5.2.2 Comparing collocations located in the NNS and NS corpora against those in the RC

Following common practice (e.g., Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008) verification of located collocations by using the BNC to check whether the located collocations exist in a large corpus, students' located N and V academic collocations were compared with the RC for verification. To avoid the need for manual checking of academic collocation lists, *ConcGram* provides a useful technique to compare lists of collocations between two corpora. Thus, the verified lists of N and V collocations (tokens) located in the RC were searched in each of the NNS and NS corpora to locate the shared set of collocations among the corpora.

As a result, 3559 N[11] collocations (tokens) from the NNS corpus were shared with the RC, while 3652 N collocations (tokens) from the NS corpus were shared with the RC. Verb collocations were also compared following the same procedure. 1126 V collocations (tokens) from the NNS corpus were similar to the RC collocations, whereas 1294 V collocations (tokens) from the NS corpus were similar to the RC collocations. Table 4-4 shows the number of tokens of students' N and V collocations compared with RC verified academic collocations.

Table 4-4: Number of tokens of students' N and V collocations after they had been compared with RC verified academic collocations.

	NNS N collocations located in RC	**NS N collocations located in RC**	**NNS V collocations located in RC**	**NS V collocations located in RC**
Number of tokens of academic collocations located in the RC	5843	6282	1454	1648
Number of tokens of collocations in students' corpora after verification	3559	3652	1126	1294

The verified lists of collocations were all sorted by their raw frequency to facilitate the search for the 100 most frequent lexical collocations. The resulting lists of N and V collocations were all saved in Excel sheets for further analysis.

[11]Note that the3559 N collocations (tokens) in the RC are tokens of the same collocation types that there are 872 tokens of in the NNS corpus. This is applicable to NS N collocations as well as to V collocations.

4.5.3 Manual vetting to limit the collocations to patterns of interest

Lexical collocations that fall into the following four types of POS combinations were the major targets of our subsequent investigation: verb + noun (e.g. *gather data*), verb + adjective (e.g. *stay safe*), adjective + noun (e.g. *systematic approach*), and noun + noun (e.g. *data user*). This conforms to the literature of conventional corpus-based collocation research. For example, verb + noun combinations were investigated by Altenberg and Granger (2001), Laufer and Waldman (2011), Nesselhauf (2005), and Howarth (1996, 1998a); adjective + noun combinations by Siyanova and Schmitt (2008) and Durrant and Schmitt (2009); and noun + noun combinations by Peacock (2012).

The verified lists of N and V collocations (tokens) from each of the students' corpora underwent a qualitative review to exclude grammatical collocations (e.g. *access for, access may*). After excluding all grammatical collocations from the lists, the 100 most frequent N and V lexical collocations (types) from each student corpus were selected for further analysis. Thus, 400 lexical collocations (types) in total were selected for testing their significance in the RC.

4.5.4 Significant collocations

Sinclair et al. (2004: 10) describe a significant collocation as the "regular collocation between items, such that they co-occur more often than the respective frequencies and the length of the text in which they occur would predict". To answer the second research question related to students' tendency of over or underuse of N collocations and V collocations, the 100 most frequent nouns and verbs collocations (types) from each students' corpora were tested for their significance. Thus, 400 collocations (types) were tested in total. A chi-squared test, with 5 per

cent as the critical level of statistical significance, was used to test the significance of most collocations (Sinclair et al., 2004; Gries, 2010). A few collocations (20 N collocations and 30 V collocations) were tested using Fisher's exact test because the expected count cells were below 5.

The chi-squared tests compared the times each collocation occurred in a student corpus in relation to the times it did not occur there with the times each collocation occurred in the reference corpus in relation to the times it did not occur there. At the time of data analysis, I was not aware of a way I could calculate the times a collocation did not occur in a corpus and consulted Mr. Phil Scholfield, one of the statistics experts in the Department of Language and Linguistics at University of Essex. He developed the formula[12]below, which was inputted into SPSS.

Trunc ((total words in corpus – (2 x number of collocations)) / 2)

This formula assumes that, in theory, a two-word collocation can occur in a corpus as many times as half the words in a corpus. The command 'trunc' was used because in some cases the expected output number could include decimal places, whereas I cannot have one collocation and a half, for example. The command 'trunc' deleted decimal places from the number of expected collocations in a corpus. Detailed chi-square values for the 400 tested collocations (types) are given in Appendix D.

[12] Even though Barnbrook, Mason, and Krishnamurthy (2013) mentioned another formula for calculating the times a collocate did not occur in a corpus, it was not used for two reasons.First, the statistical part of this study had already been completed and, second, that formula was developed for calculating frequencies for single words not collocations.

4.5.5 Dictionary Checks

From the initial analysis of some of the significant N collocations, some collocations appeared to be flexible in their uses while others appeared to be fixed terms of CS. To determine whether this impression was correct, two dictionaries were used to check the specificity of the 100 most frequent N collocations (types) from each of the students' corpora.

Two dictionaries were selected to check the meaning and use of these collocations. The first dictionary was a general CS dictionary, which was available free online;[13] the other dictionary was *The BBI Combinatory Dictionary of English* (Benson et al., 1997). I followed the following procedure in categorising the N collocations. Collocations were categorised as general academic collocations (GAC) if they were found in both dictionaries. They were categorised as general Computer Science collocations (GCSC) if they appeared only in the CS dictionary. For example, *available resources* and *code number* were found in both dictionaries, thus they were categorised as GAC, while *data layer* and *data user* were only found in CS dictionary. Therefore, they were categorised as GCSC. In some cases, collocations were not found in either dictionary (e.g. *data amount*, *method class*) so they were marked as 'not found' and were left to be categorised by CS experts (for detailed information about the dictionaries check see Appendix E).

4.6 Results and Discussion

In this section, I will present the results on both significant noun collocations and verb collocations to answer the first three research questions:

[13] http://www.specialist-online-dictionary.com/computer-dictionary.html

RQ1. What are the most common academic collocations used by Computer Science students in their MSc dissertations?

RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus?

RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?

To address the first question, I will present the most frequent academic collocations used by CS students in their writing. Then to answer the second question, I will compare the 100 most frequent noun and verb collocations from each of the students' corpora with the RC according to their frequency. To address the third question, I will compare the use of the shared noun collocations between the NNS and NS students' corpora.

RQ1. What are the most common academic collocations used by Computer Science students in their MSc dissertations?

After locating the most frequent members of the 100 most frequent AWL families in the students' corpora, a short list of the most frequent members of the 88 word families (see Table 4-1for details) for each of the students' corpora was inserted into *ConGram* to locate their collocations. Collocations were located applying MI of 3, t.score of 2, and span of three words from the left and the right of the node words.

The results reveal that both NNS and NS students tend to use noun collocations more than verb collocations, as displayed in Table 4-5 below. This finding seems to be in agreement with Halliday(1966) and Coxhead and Byrd (2007) who claim that Science discourse is characterised by the use of nominalisations and thus can be described as more noun centric than verb centric. Surprisingly, both NNS and NS use only few verb collocations significantly.

Table 4-5: The frequency of noun and verb collocations in the NNS and NS corpora

Type of collocations	**Corpus**	**Frequency**
N collocations	NNS corpus	3559
N collocations	NS Corpus	3652
V collocations	NNS corpus	1126
V collocations	NS corpus	1294

RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus?

To check whether students had a tendency to over or underuse N collocations and V collocations, the 100 most frequent collocations of each type from students' corpora were tested for their significance. Thus, 400 collocations were tested in total. The chi-squared test, with 5 per cent as the critical level of statistical significance ($p<.05$), was used for most of the collocations. A few collocations (20 N collocations in total, 30 V collocations) were tested using the Fisher exact test because the expected cells count was less than 5.

Most of the noun lexical collocations were significant. 57 of the 100 NNS N collocations were significant while 81 of the 100 NS N collocations were significant. On the other hand, the 100 selected verb collocations from both NNS and NS corpora were not all significant. Only three of the 100 NNS V collocations were significant whereas 13 out of the 100 NS V collocations were significant. Table 4-6 below presents the percentage of over/underused collocations in each of the students' corpora as compared to the reference corpus.

Table 4-6: Percentages of significantly over/underused noun and verb collocations in the NNS and NS corpora as compared to the reference corpus

	NNS N collocations	**NS N collocations**	**NNS V collocations**	**NS V collocations**
Significant overuse	52%	78%	3%	13%
Significant underuse	5%	3%	0%	0%
Total	57%	81%	3%	13%

The overuse of noun collocations was slightly higher in the NS students' corpus than in the NNS students' corpus. However, there was no significant difference between the percentages of the overused N collocations by both NNS and NS students as the z-score was 0.29 (the z-score between the NNS and NS corpora was compared to the z-score that would be expected under the null hypothesis with a=0.05 was supported). In addition, both NNS and NS students underused only a few noun collocations, while they did not significantly underuse any verb collocation. This suggests that CS students were perhaps exposed more to noun collocations in their studies and thus noun collocations were more frequently used by students regardless of whether they

were NNS or NS. This claim is in accordance with Coxhead and Byrd (2007) who found that the style of academic writing in Science is more noun centric than verb centric.

Another tentative explanation for the overuse of N collocations by students can be related to their frequent exposure to these collocations in their years of study. Jones and Durrant (2010) suggest that "the collocations students are most frequently exposed to are stored in users' mental inventories and therefore frequently retrieved from memory in preference over less conventional expressions" (390).

Unlike the findings from previous EAP studies on collocations' use that indicate that NNS are limited in their use of collocations since they overused a small set of collocations compared to NS (e.g., Durrant and Schmitt, 2009; Laufer and Waldman, 2011), in this study NNS students seem to be similar in their collocations' use to NS students. Thus, it can be concluded that NNS students tend not to face a great difficulty in using N collocations in ESP context since they were using N collocations like NS students**.** The contrast between my findings with EAP findings can be related to the different use of collocations in an ESP context. Collocations tend to be more scientific and discipline-specific in ESP registers.

On the other hand, V collocations were significantly overused for a small number of collocations. NS students overused only thirteen V collocations, while the NNS students overused only three V collocations. These collocations are presented in Table 4-7 below. Numbers in brackets are the normalised frequencies per 100,000 words in all of the tables.

Table 4-7: Raw frequency and normalised frequency (in brackets) of the significantly overused verb collocations in the NNS and NS corpora as compared to the RC.

NNS V collocations	RC frequency	NNS frequency	NS V collocations	RC frequency	NS frequency
extracted features	7(1.1)	15(4.9)	defined section	9(1.4)	32(10.8)
obtained result	5(0.8)	9(2.9)	ensure system	5(0.8)	26(8.8)
achieve goal	4(0.6)	8(2.6)	created new	8(1.33)	24(8.1)
			created object	7(1.1)	16(5.4)
			affect performance	6(0.9)	14(4.7)
			extracted data	4(0.6)	14(4.7)
			required information	8(1.33)	10(3.3)
			creates new	5(0.8)	10(3.3)
			required work	5(0.8)	10(3.3)
			implemented method	4(0.6)	10(3.3)
			demonstrates section	4(0.6)	10(3.3)
			found algorithm	4(0.6)	8(2.7)
			created data	4(0.6)	8(2.7)

Since N collocations were the significant collocations that were used more frequently than V collocations by CS postgraduate students, further analysis will be carried out focusing on N collocations. Two comparisons were carried out. The first comparison was made between the frequencies of the 100 most frequent N collocations in each of the students' corpora and in the RC. The second comparison was carried out between the shared set of N collocations that occurred in both NNS and NS student corpora. 30N collocations from the 100 most frequent N collocations were shared between the NNS and NS student corpora (this comparison will be answered in the third research question). Table 4-8 shows that the two groups of students

overused different sets of noun collocations. If I compare the top 10 overused N collocations from each of the students' corpora with their use in the RC, a clear difference can be observed.

Table 4-8: Raw frequency and normalised frequency (in brackets) of the top 10 overused noun collocations in the NNS and NS corpora as compared to the RC

NNS collocations	**RC frequency**	**NNS frequency**	**NS collocations**	**RC frequency**	**NS frequency**
network traffic	8 (1.33)	70 (23)	code source	70 (11.6)	128(43.5)
simulation results	5(0.8)	56 (18.5)	data test	34 (5.6)	50 (17)
sites web	29(4.8)	39 (6.5)	design system	13 (2.1)	50(17)
error rate	9(1.4)	32(10.6)	environment development	6(0.9)	50(17)
extraction information	3(0.4)	28(9.2)	computer vision	13 (2.1)	48(16)
allocation dynamic	4(0.6)	27(8.9)	process development	8(1.3)	46(15.6)
data layer	21(3.4)	26 (8.6)	source open	33(5.4)	44(14.9)
data different	15(2.4)	26 (8.6)	data database	7(1.1)	42(14.2)
data amount	10(1.6)	26(8.6)	data raw	6(0.9)	42(14.2)
data access	4(0.6)	19(6.3)	layer application	10(1.6)	42(14.2)

Since this comparison involves two different genres – dissertations and research articles – collocation overuse might be explained if I compare the demands of writing dissertations with the demands of writing research articles. Students work with the word limits of MSc dissertations, but these word limits are much higher than the word limits of research articles. According to the CS Department website, MSc students are required to write approximately 50-60 pages/10,000-15,000 words to fulfil the departmental requirements for writing a complete dissertation. By contrast, after checking the journals' article submission requirements, it was

clear that most CS journals tend to provide their writers with word limits that should not be exceeded. The length of an article, for example, in the *ACM Journal of Information Systems* is 25-30 pages, as stated clearly in the 'Guidance to Authors'.

Thus, the overuse of noun collocations in the student corpora as compared to the RC may be due to the fact that the former consisted of MSc dissertations whereas the latter of journal articles. Since MSc, dissertations are much longer texts than journal articles, the former are bound to include more lexical repetition than the latter. This claim is supported by research on lexical variation, that is, the variety of vocabulary deployed by a speaker or writer (Malvern and Richards, 2002) in written or spoken texts: various studies (e.g., Arnaud, 1984; Richards, 1987) have indicated that lexical variation decreases as the number of words in a text increases. Generally, lexical variation is lower in dissertations rather than in research articles. Thus, the chance of repeating same collocations in dissertations would be greater than in the research articles. Therefore, N collocations were overused in students' dissertations rather than in research articles. On the other hand, unlike NS students, NNS students tend to underuse few N collocations compared with experts as shown in Table 4-9.

Table 4-9: Raw frequency and normalised frequency (in brackets) of the significantly underused noun collocations in the NNS and NS corpora as compared to the RC

Underused collocations	**RC frequency**	**NNS frequency**	**Underused collocations**	**RC frequency**	**NS frequency**
data training	70 (11.6)	18 (5.9)	design architectural	47(7.8)	12(4)
code source	70 (11.6)	11 (3.6)			

document query	33(5.4)	7(2.3)			
method class	32(5.2)	7(2.3)			

Since few N collocations were underused by NNS and NS students that may indicate that postgraduate CS students are exposed frequently to N collocations in their study. Thus, they are rarely underused. The mentioned underused collocations might be used in specific topics rather than others.

Missing Noun Collocations

Some noun collocations that appeared in the reference corpus did not appear in the NNS corpus, the NS corpus, or in both of them. They can be considered extreme cases of underused collocations. Table 4-10 displays the top ten missing noun collocations in each of the students' corpora as compared to the RC.

Table 4-10: Raw frequency of the top 10 missing noun collocations from NNS and NS corpus as compared to the reference corpus

Missing collocations from NNS corpus	**RC frequency**	**Missing collocations from NS corpus**	**RC frequency**
files vulnerable	113	files vulnerable	113
function ranking	100	function ranking	100
document ranking	88	document ranking	88
document scope	65	network effects	71
files neutral	60	document scope	65
document cohesion	58	files neutral	60

code base	57	document cohesion	58
analysis program	53	attributes methods	53
document function	53	instance database	53
code lines	52	functions ranking	52

The absence from the student corpora of this set of collocations might be related to their use in specific topics that might not be included in the students' corpora. These collocations occurred only in two files of the RC. These were the articles with the titles below:

1- ***Progressive Alignment Method Using Genetic Algorithm for Multiple Sequence Alignment (AI 2)***
2- ***Approximating the Genetic Diversity of Populations in the Quasi-Equilibrium State (AI3*)**

Both titles of the research articles were from the same journal, *IEEE Transactions on Evolutionary Computation*, which was selected for the AI sub-discipline of CS. Thus, perhaps the missing collocations could be specific collocations for certain sub-disciplines of CS and, therefore, might not be frequently encountered by NNS and NS students.

RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?

The second comparison was carried out between the shared set of noun collocations in both the NNS and NS corpus. 30 noun collocations were shared between the students' corpora. Table 4-11 presents the 30 shared noun collocations and their over/underuses in both NNS and NS corpora. The missing ticks in some cases indicate the non-significance of the collocation.

Table 4-11: The 30 significant over/underused shared N collocations in each of the students' corpora

Collocations	**Non-native speaker corpus**		**Native-speaker corpus**	
	Significantly overused	**Significantly underused**	**Significantly overused**	**Significantly underused**
code following	√		√	
code number	√		√	
data layer	√		√	
data amount	√		√	
data access	√		√	
data user	√		√	
data information	√		√	
data Web	√		√	
data time	√		√	
data other	√		√	
data type	√		√	
design system	√		√	
environment development	√		√	
features other	√		√	
layer application	√		√	
network traffic	√		√	
resources available	√		√	
resources system	√		√	
method class		√		√
code source		√	√	
data input			√	
data structure			√	
data available			√	
design implementation			√	

section previous			√	
section following			√	
site web	√			
source open			√	
components different			√	
data different	√			

It can be clearly seen from Table 4-11 that there are some similarities and some differences in terms of the overuse and underuse of the shared noun collocations between the NNS and NS students' corpora. 18 of these collocations were similarly overused by both NNS and NS students. The only underused collocation was *method class*. Another clear difference between the NNS and NS use of these collocations was in their different frequency of use of *code source*. It was underused by NNS students while it was overused by NS students. The remaining 10 noun collocations differed in terms of their overuse and underuse as compared to the expert writers' corpus. Eight were significantly overused by NS students only and the other two were significantly overused by NNS students only.

4.7 Summary and Discussion

Results related to the first research question showed that both NNS and NS postgraduate CS students used noun collocations more frequently than verb collocations in their MSc dissertations. This finding is consistent with Coxhead and Byrd (2007) and Halliday (1966) who described scientific register as noun centric and that nominalisation is a distinctive feature of scientific register. Thus, it might be concluded that CS postgraduates are exposed more to N collocations in their years of study and thus N collocations become 'entrenched' in their mental lexicon (Jones and Durrant, 2010: 390). Since both NNS and NS students overused noun

collocations, this finding contrasts with previous collocation studies in an EAP context (e.g., Durrant and Schmitt, 2009; Nesselhauf, 2005) that show that NNS overuse a limited set of collocations and that their uses are not native-like.

Another interesting finding related to the second research question is the overuse of most noun collocations by students when compared with the RC. Genre requirements could be an important factor in the use of these collocations. Writing dissertations is different from writing in published articles. Expert writers have to follow the writing demands of journals when writing academic journal articles. On the other hand, MSc dissertations have larger word limits than journal articles so students can write in detail about their MSc projects. In addition, differences in the topics between either of the students' corpora and the RC might be another factor that could explain the overuse and underuse of some of the collocations.

Regarding the third research question, a number of possible factors could explain the similarities and differences between NS and NNS students' use of the 30 shared collocations. First, after comparing the concordance lines of these collocations in the NNS and NS corpora, different patterns were observed. In this thesis, a pattern is defined as "if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it" (Hunston and Francis, 1996:37).Thus, it was supposed that the overuse of these collocations could be related to the patterns of use that may occur more in one or both of the students' corpora rather than in the expert writers' corpus and/or of patterns that occur in one or both of the students' corpora but not at all in the expert writers' corpus. The underuse of some collocations could be related to the more frequent use of some collocation patterns in the expert writers' corpus than in the students' corpora and/or to the use of patterns

that occur in the expert writers' corpus but not at all in one or both of the students' corpora. Patterns of the shared N collocations will be identified and discussed in detail in Chapter 5.

Second, the degree of the specificity of some collocations might affect their occurrences. In the preliminary analysis of these collocations, two dictionaries were used to check whether the collocations were GAC or GCSC. I judged the collocation to be GAC if it was found in both dictionaries and to be GCSC if it was only found in the CS dictionary (see section 4.5.5 for more details).

Some of the collocations could be classified as fixed expressions, as Handl (2009) noticed that some collocations can be classified as fixed expression "if a word occurs very rarely and in almost every case with the same partner, then it has a tendency towards being used as fixed expressions"(2009:74). Some of the N collocations tend to be fixed expressions as they occur with the same partner across all corpora as well as in the dictionaries' entries. For example, *network traffic* and *layer application* were used as fixed expressions in CS as they tend not to have other collocates, whereas other collocations tend to be more flexible as they include words that also form part of many other collocations, such as *data*, which can collocate with more than three words: *access*, *information*, and *input*. This distinction between fixed and flexible collocations was confirmed by Handl (2009) in her classification of collocations. Shoppen (1985) refers to fixed collocations as compounds since they collocate with the same partner and always have the same word order, unlike flexible collocations, which allow for word order variation (see differences between compounds and collocations in section 2.4.1).

4.8 Conclusion

Since the quantitative data analysis summarised so far in this Chapter could not indicate conclusively which of the factors mentioned above has caused the overuse and underuse of some collocations in the students' corpora as compared to the expert writers' corpus, further qualitative analyses will be carried out to investigate the aforementioned factors. The 30 shared N collocations were selected for further analysis because they occurred in both student corpora and thus, differences and similarities can be identified. The next chapter will investigate the patterns of the 30 shared N collocations in detail and explore the factors behind various uses of these collocations according to CS experts' views.

Chapter 5 Factors Underlying the non-experts' Over/underuse of Noun Collocations

5.1 Introduction

This Chapter presents the second study, which will examine the factors that were thought to explain the over/underuse of the 30 shared N collocations. These factors are various collocations patterns, effect of genre and topic on the use of collocations, and discipline-specific collocations. The Chapter has four main sections. Section 5.2 will first review the literature related to pattern identification and the effect of genre and topic on the use of collocations and then will present the research questions that this study seeks to explore. Section 5.3 presents the procedure followed in identifying patterns and in verifying the results using two other approaches: categorisation judgement task and in-depth interviews with CS experts. Section 5.4 reports the results of pattern identification, the categorisation judgement task, and experts' interviews and discusses them in detail. Finally, section 5.5 will provide a summary of the chapter.

5.2 Literature Review

5.2.1 Research on the grammar and lexis of collocations

Researchers differ in their views about the relationship between grammar and lexis. Three main approaches have claimed a relationship between lexis and grammar. These are Sinclair's idiom principle approach (1991, 1996), Hoey's lexical priming approach (2003, 2004, 2005), and

Hunston and Francis' pattern grammar approach (1996, 2000). This section will describe briefly these approaches.

Sinclair (1991) proposes two different principles of interpretation to explain the way in which meaning arises from a linguistic text. These principles are the open-choice principle and the idiom principle. The open-choice principle, which is often called 'slot-and-filler' model, represents the traditional assumption that grammar is the main restraint in seeing and describing language. That is, "language text is a series of slots which have to be filled from a lexicon which satisfies local restraints grammar" (Sinclair, 1991: 109). On the other hand, the idiom principle emphasises that a large number of multi-words units are constructed as single choices in the language user's mind (Sinclair, 1991). Thus, collocations are considered as single units even though they might be analysable into segments.

The link between these two principles has been clearly established by Sinclair (1991) who proposes that, ideally, when reading, the idiom principle is the normal mode applied since "the majority of text is made of the occurrence of common words in common patterns or in slight variants of those common patterns" (Sinclair, 1991:108). Nevertheless, whenever lexical choices appear, which are unexpected, a switch to the open-choice principle will occur. Thus, it can be concluded that both principles are connected and that the switch between them is based on the reader's existing store of collocations. Grammar is the output of repeated collocational groupings as words are mentally 'primed' for use through our experience of their infrequent association with others (Hyland, 2008).

Sinclair (1991: 111) summarised seven features of the idiom principle. First, many phrases have an open-slot. For example, *set eyes on* attracts a pronoun subject. Second, many phrases allow for internal lexical variation e.g. *set x on fire* or *set fire to x.* Third, many phrases allow internal lexical syntactic variation e.g., *it is not in his nature to* can be replaced by changing *is* to *was*, *not* to *hardly*, and *his* to another possessive. Whereas *it*, *in*, and *nature* cannot be changed. Fourth, many phrases allow some variation in word order, e.g., *it is not in the nature of an academic to…, to recriminate is not in his nature*. Fifth, many uses of words and phrases attract other words in strong collocations e.g., *hard work, hard luck.* Sixth, many uses of words and phrases show a tendency to co-occur with certain grammatical choices, e.g. *set about* always occurs with the *–ing* verb form. Seventh, many uses of words and phrases show a tendency to occur in a certain semantic environment, e.g. *happen* is associated with unpleasant things such as *accidents*. According to Sinclair (1991), collocations function as single lexical items. Therefore, all the features of the idiom principle apply to them.

Sinclair (1996, 2004), in his model of extended lexical units, proposes different sets of lexical meaning to the words: starting with collocation (focus on the meaning of words), moving to colligation (in which focus is related to grammatical patterns of the words), then to semantic preferences (focus on the context of the words), and, finally, to semantic prosody (focus on the discourse function of the unit). His model claims the importance of the lexis as it is described in the centre: lexis related to other lexis, to the world, and then to the speaker.

Hoey (2003, 2004, 2005) has also viewed lexis as an important unit of language to be observed. His lexical priming approach claims that language learners can subconsciously notice the collocations, colligations, and semantic and pragmatic associations of the lexis whenever they encounter them. Language learners can also notice the contextual features of a word or cluster of

words. That is, they can identify the genre, style, and social situation in which a word or cluster of words will be used. Furthermore, linguistic-textual features are the third dimension of his approach, whereby learners can observe the textual features of the words. Their textual collocations, textual colligations, and textual semantic associations can be noticed subconsciously (Hoey, 2003, 2004, 2005). It should be noted that Hoey's approach is concerned with a psychological view of native speakers' mind. Thus, collocation is, in his view, a psychological association between words that is merely "evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (2005, pp. 3-5).

In their pattern grammar approach, Hunston and Francis (2000) have built on Sinclair's work to propose a description of language in terms of patterns. Hunston and Francis (2000: 3) claim that "pattern is a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups, and clauses that follow the word. Patterns and lexis are mutually dependent, in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns". Thus, according to pattern grammar, grammar and lexis cannot be treated as distinct phenomena in the description of English (Hudson, 1984; Hunston and Francis, 2000).

It can be seen that the three approaches present and confirm the connection between lexis and grammar, as they are interrelated. Sinclair's idiom model was explained by Hoey's lexical priming approach. Like Sinclair, Hoey points to associations between words and other words (collocation) and between words and groups of semantically related words (semantic association, equivalent to Sinclair's semantic preference). Hunston and Francis (1996) describe lexico-grammatical combinations, which are generalisations from the multitude of collocations and colligations observable in corpus data.

Pattern grammar will be applied in identifying collocations' patterns grammatically since the aim of the current study is to locate patterns of collocations. It will mainly involve dealing with grammar rather than focusing on meaning. No attempt will be made to identify patterns according to their meaning since an understanding of CS discourse is needed. This is a task for which the researcher must be well-versed in CS. In the following section, a more detailed description of patterns and approaches of pattern identification will be presented and previous studies related to pattern identification will be summarised.

5.2.2 What is a pattern?

In general, *pattern* means repetition. If single symbols '*' are repeated twice '**' or more it becomes a minimal pattern. The minimal pattern "may form a sequence that when repeated comprises a more noticeable pattern" (Hunston, 2010:152).

******** ******** ********

******** ******** ********

In language, pattern is observed when words, sounds, rhythms, or structures are repeated (Hunston, 2010:152). Patterns have been described as an approach of describing language that focuses on grammar and meaning of words (Francis et al. 1996; Hunston and Francis, 1996, 2000). This approach is called pattern grammar (Hunston and Francis, 1996), which describes the syntactic and semantic behaviour of words, in a certain environment (Mason and Hunston, 2004). Thus, the pattern of a word consists of the words that follow it and precede it.

Even though patterns are related to the grammatical and lexical features of the word, they cannot be described under the headings of either ‘lexis' or ‘grammar’ (Hunston and Francis, 1996). Sinclair (1991) and Hunston and Francis (1996:251) argue that “patterns are so central to the description of language, this cross-classification cannot be dismissed as a marginal peculiarity, but it must count as a challenge to the distinction between lexis and grammar itself, so that the word grammar, if it is used at all, must comprise information about lexis as well as information about syntax”.

5.2.2.1 Importance of Pattern Identification

Patterns have been considered useful in describing linguistic variation. First, presenting the link between lexis and grammar is one way of describing linguistic variation. Second, expressing a single meaning in different lexis-pattern combinations is another way of highlighting linguistic variation. Finally, identifying significant patterns in a specific register of a language helps to indicate the meanings that are prevalent in that register. Mason and Hunston (2004) comment that patterns in particular disciplines can reveal the phraseology of that discipline.

Recognising patterns is also important in language teaching because it facilitates the development of both accuracy and fluency (Hunston and Francis, 1996, 2000). Observing NS and NNS use of patterns will reveal how much control the NS have over their second language. Even advanced NS learners tend to have imperfect control over patterns. If NNS learners use a word in a correct grammatical pattern, their usage may be unidiomatic rather than wrong. Moreover, when a learner learnt a word with its pattern a series of words phrased together can be produced. This can be interpreted by ‘pattern flow’ in which a word of one pattern is the starting word for another pattern (Hunston and Francis, 1996, 2000). Thus, it would be useful to locate

patterns of collocations located in NNS and NS students writing to investigate their control of language.

Coxhead and Byrd (2012) also point to the importance of identifying patterns in academic settings that will inform researchers, teachers, and material designers. Durrant (2009) argues for the pedagogical importance of teaching patterns of collocations to students. He claims that drawing learners' attention to patterns that are needed will be of great advantage and will make vocabulary teaching more beneficial. Hunston and Francis (1996) suggest that the awareness-raising approach is the most suitable for teaching patterns. A number of researchers have applied this approach to raise their readers' awareness about the use of patterns (Lewis, 1997; Jones and Durrant, 2010; Jiang, 2009).Thus, a sample of awareness-raising activities were designed for NNS in Chapter 6.

5.2.2.2 Types of Patterns

Patterns were first identified by Francis et al. (1996, 1998) for the main four open classes (noun, adjective, verb, and adverb). The first aim of identifying patterns for these classes was to develop coding for their inventory in the *Collins COBUILD* English Dictionary (CCED) (Sinclair et al., 1995). Patterns were categorised by observing what follows a word and what precedes it. Verbs are mostly identified by subsequent words since most verbs have complementation patterns that follow them. Though complementation patterns are usually the most interesting facts about verbs, there may be other reasons for identifying their following patterns, as this would show how often a verb occurs in the passive or infinitive, which modals it is often used with, which nouns are its typical subjects, whether it is frequently negated, and so on.

Similarly, adjective patterns can be identified by the following words that represent types of nouns the adjective modifies (e.g., ADJ N, ADJ –ing) and their complementation patterns (e.g., ADJ that, ADJ prep, ADJ to-inf)). Moreover, in some cases, identifying the preceding words reveal interesting facts about the kinds of modifiers that commonly collocate with the adjective. For example, the 'predictive adjective' that always occurs after a link verb has the pattern 'v-link ADJ'.

In the case of a noun, identifying the complementation of nouns is the most revealing since it shows the various ways in which the noun is modified (Francis et al., 1998). Even though verb and adjective patterns are important to be identified and recognised by both linguists and teachers, no attempts are given to present them in detail in the current study since the focus of this study is on lexical collocations and mainly noun collocations. Thus, a detailed description of noun patterns will be presented.

The following are the main noun patterns used in the CCED and in Francis et al. (1998). Noun patterns were identified into groups according to POS of preceding words (Group A) and to POS of following words (Group B) (Hunston and Francis, 1996: 56-58).

Group (A): Patterns with POS preceding the noun

1- a N, the N: the noun is preceded by an indefinite or definite article

2- Poss N: the noun is typically preceded by a possessive determiner like 'my' or 'your' or a possessive-formed noun group

3- ADJ N: the noun is preceded by an adjective

4- NN: the noun is preceded by another noun

5- from N, to N, on N, etc.: the noun is preceded by a specific preposition. The prepositions most frequently used in patterns like this are as follows: *at, by, from, in, into, on, out of, under, with.*

6- Supp N: the noun is preceded by a range of the elements given above: determiner, possessive determiner or possessive noun group, adjective or noun.

Group (B): Patterns with POS following N:

1- N to inf
2- N that
3- N N
4- N prep
5- N of N, N for N, N from N, etc. The noun is followed by a prepositional phrase introduced by a specific preposition (*e.g. about, against, among, as, at, behind, between, for, from, in favour of, in, into, of, on, over, to, towards*).
6- N with supp, which means that the noun is both preceded by a range of the elements mentioned above and followed by them.

The focus of this study is on identifying patterns for N collocations. Thus, both ADJ-N and N-N will be searched for.

5.2.2.3 Previous Studies on Identifying Collocation Patterns

A number of corpus-based studies have been conducted on identifying patterns of words for different purposes (e.g., semantic sequences (Hunston, 2008), categorisation of collocations (Coxhead and Byrd, 2012), and lexical bundles (Hyland, 2008)). However, few studies on pattern identification have been conducted in the ESP context. Gledhill (2000a) was the only

study conducted to investigate the discourse function of medical research articles. Searching for certain grammatical collocation patterns, he has identified the collocation framework in Medical discourse.

Different methods have been applied in identifying word patterns. These are mainly automatic recognition of patterns versus manual identification. A number of studies applying a manual identification of patterns have been conducted (Cacchiani, 1984; Hunston and Francis, 1996; Hunston, 2008, 2010). In this method, the researcher, after sorting out the concordance lines either by the node word itself or by the left or the right context of the node word (Tribble, 2010; Hunston and Francis, 1996), locates similar and different patterns of words in randomly selected concordance lines (Hunston and Francis, 1996; Mason and Hunston, 2004).

Hunston (2008) presents three alternative approaches for searching for semantic sequences. The first approach involves starting with specific words or phrases and searching for their patterns – by looking at their grammatical similarities and differences – and then grouping them semantically. It can be considered a method useful for locating patterns that could be generalised. The second approach focuses on a certain pattern to be searched and located in the selected concordance lines. A search for the pattern *N that*, for example, will yield a number of nouns that follow this pattern (e.g., *suggestion that, observation that*). Even though this approach is a targeted search on a grammar pattern, it could be useful in a specific piece of discourse. Hunston and Francis (1996) have also considered these two approaches in their identification of patterns. They suggest that patterns can be identified in two perspectives; the researcher can begin by a single word and look for their different patterns or begin with a certain pattern and search for different words that are associated with that pattern.

The third approach focuses on identifying certain grammatical words in a specific discourse. The search is based on 'small words': that is, grammar words such as prepositions. Gledhill (2000a), in his identification of the collocation framework, applied this approach. Examining these 'small words' in a specific discipline reveals a surprising amount about the "epistemology and ideology of the discipline because they reveal phraseologies that are linked to recurrent meanings" (Hunston, 2008:293).

The three aforementioned methods of pattern identification do not seem to be different in nature. Starting with either specific words or small words will yield a number of patterns. These patterns can be further classified according to their syntactic or semantic meaning. Thus, a pattern will be specified (Hunston, 2008).

Coxhead and Byrd (2012) have adopted Hunston's (2008) first approach to locate collocations for their AWL words. Using the same 3.5 million-word corpus compiled for locating the 570 AWL families (Coxhead, 2000), the most frequent word members of each family were selected. Moreover, using *log likelihood* for collocation indication, a list of the most frequent collocates of the selected words was compiled. *Wordsmith Tools* 4.0 presents the most frequent collocates for the selected academic word to the left and to the right of the word. Thus, collocations have been identified in both directions. Focusing on the five most frequent collocations for the selected academic words, collocations were categorised according to their *log likelihood* results into

strong (*e.g., create, analysis*), weak (*ongoing*) and lonely [14](*e.g., nonetheless*).This study is useful in highlighting steps of identifying collocations to the left and to the right of the words under investigation since it would be useful in locating as many collocations as it can.

Similarly, Cacchiani (1984) has investigated the complex collocations of intensifiers using the BNC (100 million words) as the main corpus. Using the software *Sketch Engine*, 250 random concordance lines were checked for their intensifiers from both left and right side of the nodes. Intensifiers' patterns and degrees of complexity were identified manually. Three main categories of intensifiers were identified: intensifiers' ability to occur in complex collocations clearly originates in the lexico-semantic features of intensifiers. "The less grammaticalised and more subjective the type of evaluation, the more likely is the intensifier to modify other intensifiers. The more undistinguished the emotion, the more likely the intensifier is to occur in complex collocations" (Cacchiani, 1984:244).

However, identifying patterns of collocations in ESP disciplines has not received much attention. Gledhill (2000a) follows Hunston's third approach (beginning with small grammar words) to investigate the collocation framework in Biomedical discourse. A corpus of 120 cancer research articles was compiled accounting for half a million words. The top ten silent grammatical words were selected to be searched for their patterns. His focus was on the verb forms *has, have, been, is* and the prepositions *to* and *of*. He looked for these sets of word patterns individually and in

[14]Coxhead and Byrd (2012) categorised collocations according to their strength of collocational relationship with the words preceding and following the node. This collocational relationship was measured by Log Likelihood (LL). The high significant frequent nodes are the ones that have more collocates and are called the strong; the low frequent nodes are the ones that have fewer collocates and are termed the weak; while the nodes that have weak collocational relationships based on the log likelihood statistic and equally weak patterning in the set of three-word patterns are labelled the lonely.

combinations. For example, the verb forms *has/have* were searched for their patterns and then a search of the combination of each of them and *been* (*has been/have been*) was carried out. To identify the patterns of these silent words, an understanding of the context was required. As a result, patterns were identified by focusing on their meaning and functions. An example of specific pattern in biomedical research articles was:

[Biochemical process] (Possessive) ability to [biochemical process]

As presented in the following example from Gledhill (2000a: 127):

Calibrating their [leukocytes'] ability to modify factor specific DNA

Exemplified by its [Xpa3] ability to undergo epoxidation.

Another method of identifying collocations patterns called 'accumulative collocations' was described by Hunston (2008, 2010). This method is "used to perform a recursive search to refine what is observed" (Hunston, 2010:163). The search starts with a single word and looks for its most frequent adjacent-word collocate. For example, *distinguishing* has *between* as its most frequent adjacent word. Then the words *distinguishing between* will be the starting point for another search. The most frequent adjacent collocate of *distinguishing between* is *of* (Hunston, 2010:163). The search can be held in both directions – that is, by looking at preceding words and following words. Even though this method is productive in terms of understanding collocations, it will not be followed, as I am only interested in finding patterns of collocations in the limit of the two words under investigation.

So far, all studies identified patterns of individual words or their collocations manually. An automatic recognition system for identifying verb patterns has been developed by Mason and Hunston (2004). They claim that developing an automatic recognition system will be an essential

step towards large-scale textual analysis. Two pre-processing steps have been applied: parsing the verb lists and tagging the POS of these verbs. Moreover, this was done by focusing on limited linguistic information; that is, only syntactic categorisations were taken into consideration. To evaluate their software, 100 cases were tested for their patterns for the verb *decide*; about 85% were correctly identified while the remaining 15% were wrongly identified. These wrong identifications was related to a number of problems encountered during pattern identification: ambiguous patterns and intervening words, tagging errors, multiple patterns, and non-canonical patterns[15].Therefore, identifying patterns automatically is not an easy task.

The observer should be aware that there were a number of problems encountered during the identification of patterns. Pre-processing of word lists and their patterns needs to be manually checked. Thus, human judgement is necessary in identifying patterns even in automatic systems. Coxhead and Byrd (2012) claim that generating word lists based on statistical analysis is not enough; they continue enforcing the importance of human checking of concordance lines to seek additional information and to take a long, careful look at how words and their typical phrases are being used in context. For this reason, the patterns of selected collocations will be identified manually in this thesis.

[15]These patterns refer to "where the word order does not follow the prototypical sequence" (Mason and Hunston, 2004: 264)

5.2.3 Genre Effects on the use of collocations in corpus-based studies

5.2.3.1What is genre?

Different spoken and written genres have different communicative purposes. Thompson (2001) notes that different genres have different conventions. He defines written genre as "a socially constructed concept to describe a set of texts that are perceived to perform similar functions. Texts belonging to a genre are conventionalized, to differing degrees, in terms of sequencing, of layout, of phraseology, and there are expectations of, and constraints on, the structure and linguistic expression of such texts. These expectations can vary from one disciplinary community to another. The forms that the texts take can also vary, depending on the range and diversity of purposes that exponents of the genre are asked to serve" (Thompson, 2001: 33-34). Johns et al. (2006:247) summarised the purpose of genre studying as to cover "the complexities of texts, contexts, writers and their purposes, and all that is beyond a text that influences writers and audiences".

Thus, genre knowledge does not relate to the understanding of textual features only, but also to the "understanding of the social and cultural context in which genres occur as well as how these factors impact the language choices made within them" (Paltridge, 2001: 7). For Hyland (2004: 55-56), genre knowledge is "not simply grammatical competence but involves the ability to understand how to participate in real-world communicative events" and thus genre knowledge is "knowledge of the culture in which writers, readers and text are found".

In the area of ESP/EAP, "genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognised by the expert

members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and constrains the choices of content and style. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience" (Swales, 1990: 58). A seminar presentation, a university lecture, or an academic essay are different genres in academic discourse (Paltridge, 2001).

Hyon (1996) identifies two types of ESP genre approach. The first type of genre research takes a global approach, looking at the overall structure of the texts rather than at a specific type of language. An example of this approach is the Swalesian Move Analysis, which "describes global organisational patterns in genres" (e.g. Bhatia, 1993; Swales, 1981, 1990). The second type of ESP genre analysis concentrates on specific grammatical features, such as verb tense, hedges, and passive voice. Flowerdew's (2002) paper is a good example of this micro approach to genre analysis, which concentrates on specific features of language rather than on general patterns of text. His approach begins with textual analysis and then looks in detail for grammatical and lexical features covered in the text under analysis.

5.2.3.2 Genre-based studies of academic writing

A number of genre-based studies have been conducted to reveal inter- and intra-disciplinary variations in academic writing. Research into disciplinary variation either across various disciplines (inter-) or within a specific discipline (intra-) has largely focused on expert writing, i.e., research articles (RAs) (e.g., Harwood, 2005, 2006; Hyland, 2004; Samraj, 2002). A research article is defined as "a formal article reporting original research that could be submitted to an academic journal. Rather than a format dictated by the professor, the writer must use the

conventional form of academic journals in the relevant discipline" (Cooper and Bikowski, 2007: 213), and is considered the prestigious type of experts' writing. Thus, investigating linguistic variation in experts' writing will yield the most conventional academic features either across multi-disciplines or within a specific discipline (Hyland, 2008).

Recently, researchers have turned their attention to the disciplinary variation in students' writing, mainly the Doctoral thesis and the Master's dissertations to compare between NNS and NS students in their academic writing (e.g. Altenberg and Granger, 2001; Bunton, 2002; Hyland, 2004; Samraj, 2008). However, few studies have been conducted to compare experts' writing with novice students' writing.

In corpus-based phraseological studies, the main comparison was carried out between NNS and NS students' use (Durrant and Schmitt, 2009; Siyanova and Schmitt, 2008) to investigate their over/underuse of certain types of collocations. Even though the findings from the studies reveal differences between NNS and NS students in their use of lexical collocations, they have not compared their uses to experts' uses in the field. One of the aims covered in this thesis was to compare the uses of academic lexical collocations between NNS and NS students and experts in the field of CS.

Hyland (2008) has identified the most frequent four-word clusters' (which are called 'extended collocations') functions and structures in three different genres – research articles, PhD theses, and MA dissertations – in a corpus of multi-disciplines to investigate their over/underuse among experts' and students' writing. Clusters in research articles were less frequently used compared

to MA dissertations and PhD theses. Hyland confirms that genres' variations influence uses and structures of these clusters. Students' genres are more 'phrasal' than the research articles and they tend to depend on using these four-word clusters in developing their arguments.

Hyland (2008) notes the importance of using the four-word clusters in a particular genre to signal the users' involvement in a given community. These clusters are more frequently used by writers and readers in a specific genre, thus, the "absence of them might reveal lack of fluency of novice or newcomer". As a writer matures, they use more collocations and extended collocations in their writing (Haswell, 1991). "Gaining control of a new register therefore requires a sensitivity to expert users' preferences for certain sequences of words over others that might seem equally possible" (Hyland, 2008:42).

Moreover, the three genres used the identified four-word clusters differently. The three identified functions of clusters (participant-oriented, text-oriented, and research-oriented) were employed differently in each genre. Research article clusters were more participant-oriented and text-oriented rather than research-oriented. On the other hand, MA dissertations and PhD theses were more research-oriented. Thus, these findings confirm that variation between experts' writing and students' writing depends on their purpose and audience as well as the written context.

Hyland (2008) goes on to compare the variation of the purpose of experts' writing and students' writing. While both research articles and students' dissertations present arguments, the purpose is completely different. Writing in research articles is concerned with "persuasive reporting through the review process and engagement with the professional world" (Hyland, 2008: 56);

thus, it is related to norm developing rather than the norm developed, as described by Swales (1990). The main aim of writing a research article is to "disseminate academics' research and establish their reputations, exhibiting to colleagues both the relevance of their work and the novelty of their interpretations" (Hyland, 2008: 57). On the other hand, when writing dissertations for either MA or PhD, students are concerned with only the reader of their works.

5.2.3.3 Experts' and students' writing

Other researchers have been investigating variation between expert writers and novice writers in their writing for academic purposes and in a specific domain (Bereiter and Scardamalia, 1987; Geisler, 1994; Tardy, 2009). Bereiter and Scardamalia (1987) describe the differences between expert and novice writing by distinguishing between two types of the composing process: knowledge telling and knowledge transforming.

Knowledge telling is the process in which inexperienced writers simply employ the knowledge readily available to them. Knowledge transforming, on the other hand, is a more complex process, of which knowledge telling is one part. The knowledge transforming writing processes consist of two problem spaces: subject matter content and rhetorical. In this knowledge transforming model, writers go beyond knowledge telling to rework and transform their knowledge. Bereiter and Scardamalia (1987) refer to the two-way interaction between context and rhetoric as "dual problem space". Thus, it seems that novice writers may not have reached the level to work on this "dual problem space" as experts do.

Experts' writing can be described as knowledge transforming since the process of "peer review works as a control mechanism for transforming beliefs into knowledge". The writing in the research articles is expected to be prestigious and the model of good academic writing. Therefore, experts' writing transform the knowledge differently. Unlike experts' writing, students' writing can be seen as an example of knowledge telling since they "demonstrate a suitable degree of intellectual autonomy while recognising readers' greater experience and knowledge of the field" (Hyland, 2008: 47).

A number of studies have been conducted to investigate students' difficulties in their writing of dissertations. From his in-depth interviews with 22 NNS students, Shaw (1991) found that ESL students had difficulty in their writing of dissertations and were influenced by genre and discipline specific vocabulary rather than by their first language or cultural factors. Moreover, Dong (1998) carried out his survey of 169 NNS students and their advisors' views in two US institutions, finding that NNS graduate students show more writing difficulties with discipline-specific, genre-specific, and audience-specific knowledge. When asked what areas of English were most important in writing research articles 100% of NNS graduate students indicated vocabulary, as compared with 40% of NS graduate students.

5.2.4 Topic Effects on the use of collocations in corpus-based studies

Another factor that could explain the variation of academic collocations' use in the students and experts' writing is topic. The topic selected for inclusion in the sub-disciplines may play a role in highlighting the use of some collocations rather than others. These collocations could be classified as standard terminology within the discipline.

From his investigation of the most frequent noun collocations in eight disciplines, Peacock (2012) found that most of the collocations presented were standard terminology within the discipline. For example, *crystal data* was a specific term used in Chemistry only, while *software process* and *user model* were specific terms that occurred only in CS. He concludes that there is a set of discipline-specific collocations that occur only in specific disciplines and which play an essential role in conveying meaning in that discipline. Furthermore, the sharp discipline differences presented indicate that the high frequency collocations of common nouns are part of the favoured terminology by which disciplines can be differentiated (Groom, 2005).

Ward (2007), from his investigation of common nouns and their collocations in Chemical Engineering textbooks, found that the three most common nouns *gas, heat,* and *liquid* were collocated with certain words to express certain meanings in the discipline. His explanatory study has been valuable in highlighting the discipline-specific collocations in Chemical Engineering.

Findings from the aforementioned studies confirmed the importance of discipline-specific collocations' research since evidences reveal that there are sharp discipline differences in their uses of collocations. Since a set of high-frequent collocations occur in a certain genre they would represent disciplinary norms and if they are presented in different patterns from other disciplines they may be accepted as writers' recognised ways of writing in that discipline(Hyland,2000:78). Schmitt and Carter (2004) confirm that frequent collocations in a corpus indicate that they are conventional within a discourse community.

5.2.5 Conclusion

Having reviewed the literature related to the suggested factors(patterns, topic, genre and experts' and non-experts' writing) underlying the over/underuse of the 30 shared N collocations, the next sections will investigate in more detail the patterns of the 30 shared N collocations in the student's corpora and the RC (see Table 4-11 in the previous chapter). By analysing each collocation concordance line qualitatively, I will try to explain why some collocations were overused and others were underused in one or both of the student corpora, compared to the RC. The rationale behind this collocation pattern analysis is that a collocation may be overused in one or both of the students' corpora because it appeared in patterns that occurred rarely, if at all, in the RC. On the other hand, a collocation may be underused in one or both of the students' corpora if it appeared in fewer patterns in them than in the RC. In addition, other factors will be investigated by CS experts' interviews and categorisation judgement task (CJT).

Research Questions investigated in this study:

RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora?

RQ5. To what extent do the shared collocations differ in their patterns?

RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views?

RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?

5.3 Methodology

To address the aforementioned research questions and to determine the reasons behind over/underuse of the located collocations in the students' corpora, a series of quantitative and qualitative methods were employed: patterns identification, CS experts' in-depth interviews, and categorisation judgement task.

To answer the first and second research questions, patterns were identified for the 24 shared N collocations among students and reference corpora (the 30 shared N collocations fell to 24 collocations after cleaning the concordance lines; this will be explained in detail in section 5.3.1.3). To answer the third and fourth research questions, CJT was given to CS experts to verify our findings about dictionaries' information about the specificity of the collocations, and in-depth semi-structured interviews were conducted with three CS experts to find out the factors behind over/underuse of some of the collocations as well as their located patterns. Each method will be presented in detail in the following section.

5.3.1 Pattern Identification

5.3.1.1 Pattern identification in previous studies

Reviewing the literature, there has been a number of different methods for collocation pattern identification. Hunston (2008) first suggested three methods for identifying semantic preferences: starting with certain words to locate their patterns, starting with a certain pattern to locate words that can be categorised under the pattern, and starting with small words (e.g. prepositions) to be searched in a specific discourse. The first two methods were seen as two sides of the same coin (Hunston and Francis, 1996).

These methods can be applied either manually or automatically. In the first approach, the researcher identifies patterns of words by observing the words following and preceding the word under investigation. On the other hand, identifying patterns automatically requires a lot of time and effort. Mason and Hunston (2004) developed their automatic system recognition for certain types of verbs by first tagging their texts by POS and then made their list of verbs and their patterns that were extracted from Sinclair et al. (1995) and Francis et al. (1996). Even though their evaluation of the system yields good results, complete dependence on an automatic system for pattern identification cannot be reliable. Ambiguous patterns and error POS tagging were all found and the best solution was to re-check automatically identified patterns manually.

Most of the previous studies in EAP contexts applied Hunston's (2008:277) first method "starting with certain words" to locate collocations of academic words (Coxhead and Byrd, 2012) or to categorise intensifiers of collocations manually (Cacchiani, 1984). Coxhead and Byrd (2012) categorised their academic-word collocations into three groups – weak, strong, and

lonely –according to their *log likelihood* results. Following Shin and Nation's (2008) criteria of selecting collocations by their word types rather than by their word families, as different word types have different collocates, Coxhead and Byrd (2012) located patterns of the most frequent word type of each word family. Then the top five patterns were identified for the most frequent academic words by observing the preceding word and the following word. That is, patterns were located by recognising the context of the word from both sides (the right and the left side). Cacchiani (1984) used the same procedures in identifying the complexity of intensifiers in collocations.

In his identification of grammatical collocations' patterns in his corpus of medical research articles, Gledhill (2000a) applies Hunston's (2008) third method, "starting with small words". The focus was to reveal the epistemology and ideology of medical discourse. By recognising patterns for a small set of verbs (*has, have, been, is*) and prepositions (*of, to*) of Medical articles' introductions both individually and in conjunction with other forms of verbs, he identified collocations' framework. He categorises the identified patterns, as he understands the meaning of the context (for detailed information about Gledhill's (2000a) procedure see section 2.4.5.3. However, it would be difficult for the researcher to understand the context of a specialised register if he is not a member of that community.

Another important issue raised when identifying patterns of collocation is whether to identify them in all possible levels of the sentence or in a limited level. One of Shin and Nation's (2008) criterion for collocation identification can be also related to collocation pattern identification (for more details about Shin and Nation's (2008) other criteria see section 2.4.5.3). Criterion 5 that was related to the grammatical well-formedness was explained as "collocation should not cross

an immediate constituent boundary… Immediate constituents are components that immediately make up larger parts of a sentence" (2008:342). They consider phrases, clauses, and sentences as immediate constituents. For example, consider the following sentence adopted from their study (2008:342):

{In [(saw v you n) vp (at prep (that det place n) np) pp] pred} s

It consists of five immediate collocational constituents:

1 'I saw you at that place',

2 'saw you at that place',

3 'saw you',

4 'at that place', and

5 'that place'

'You at the place' however does not meet this criterion because it crosses an immediate constituent boundary. Thus, criterion 5 considers that collocations can occurred in the phrase level, clause level, and in the sentence level. In the current study, collocations were identified at the phrase level only; this will be explained in more detail in section 5.3.1.3.

5.3.1.2 Steps for identifying patterns and skills needed

It is obvious that concordance programs only find and organise data; interpretation is a human activity. To identify patterns from a set of concordance lines, a number of skills need to be applied. First, it is important to formulate the search to produce a manageable set of concordance lines (Evison, 2010; Tribble, 2010; Scott, 2010). For example, searching for the words *it is surprising that* in the Bank of English will result in 176 concordance lines. A number of patterns

will be observed in these lines. However, limiting the search by adding the word *not* to be searched, as in *it is not surprising that*, will yield fewer and, therefore, manageable number of concordance lines (Hunston, 2010:158).

Second, when concordance lines are obtained, the next step is interpretation. To identify patterns from the selected concordance lines, observing similarities and differences among them is required. It involves identifying the words preceding and following the word under investigation. Ignoring distracters is also an important skill. The researcher should be able to separate what is a pattern from what is unlikely to be so (Hunston, 2010). Even though computer software such as *Wordsmith Tool* is useful in finding and organising concordance lines, it cannot group and identify similarities and differences among the lines.

After identifying similar and different patterns, the next step is to group them linguistically. For example, observing ten random concordance lines of *react* (these lines were adopted from Hunston, 2010: 159) yielded four linguistic patterns if they are grouped by the words following the node word *react*. These linguistic patterns are a subordinating conjunction (lines 1 and 2), a preposition (lines 8 and 9), an adverb (lines 3 and7), and to-infinitive clause (line 10).

1 could not believe the way Vieira ***reacted*** after he was dismissed. The…
2 at all. When asked today how they'd ***react*** if the White House sent them a ne…
3 step, which will enable viewers to ***react*** immediately to what they have see…
4 two-thirds of the radical pairs ***reacting*** (in a field of typically only…
5 anymore, I don't know how he would ***react***. Is there any point in making…
6 growth because stock markets could ***react;*** Mr Visco said stock markets in…
7 police officer at Selhurst Park ***reacted*** similarly to the Cantona incident…

8 mail, in New York, Adrian Clark ***reacted*** to Simon Hoggart's discussion of…

9 market has come, and how people will ***react*** to it, .The best seats and places…

10 strength of a substance and the body ***reacts*** to fight off any diseases which…

Further observation of these lines may yield different grouping. For example, *react* in lines (1, 2, 5, 6, and10) occurs at the end of a clause, while in lines (3, 7, 8, and 9) it is followed by the preposition *to* as a necessary part of the clause.

Another set of patterns can be observed by looking at what preceded the node word *react*. Patterns can also be further identified according to their meaning, that is, the function and use of the pattern in the selected lines (Hunston, 2010). Thus, patterns will be grouped semantically. For example, the subject of *react* can be grouped into two groups: intentional (as in lines 1,3,5,8 and 9) or non-intentional (as in lines 4, 6, and 10).

To identify patterns appropriately, a focus on the purpose of identifying patterns should be taken into consideration. Whether the researcher is keen on investigating similar semantic or linguistics aspects of the word needs to be pointed out. Thus, an accurate set of patterns will be identified. Another important issue related to the presentation of the pattern is whether it should be presented in a linear or hierarchical way (Hunston and Francis, 1996; Mason and Hunston, 2004). In the linear presentation, each pattern stands alone and will not be considered a part of the next pattern; for example, in the sentence *if you decide you want to get pregnant*, the verb pattern 'V to-inf' stands for only *want to* and the following 'V ADJ' pattern is not related to the previous pattern.

Figure 5-1: Linear presentation adopted from Mason and Hunston (2004: 259)

V			*that*				
V					to-inf		
						V	ADJ
If	*you*	*decide*	*you*	*want*	*to*	*get*	*pregnant*

On the other hand, the hierarchal presentation shows the relation between the lexical words by introducing the identified pattern as the starting point of another pattern (Mason and Hunston, 2004). Thus, it clearly displayed the 'pattern flow': "Pattern flow occurs when an item that is a component of one pattern is also the starting-point of another pattern" (Mason and Hunston, 2004:259). As shown in Figure 5-2, the pattern 'V to–inf' is the starting point for the following pattern 'VADJ'.The hierarchical presentation is advantageous as it presents the relations between clauses.

Figure 5-2**:** Hierarchical presentation adopted from Mason and Hunston (2004:259)

V			*that*				
V					to-inf		
						V	ADJ
If	*you*	*decide*	*you*	*want*	*to*	*get*	*pregnant*

Even though some researchers prefer to use the linear way in their presentation of patterns (Tognini-Bonelli, 2001) as it is considered a new way of looking at language, the traditional hierarchical presentation will be used in this study since relations between phrases are needed for the identification of collocations at the phrase level.

5.3.1.3 Steps of Pattern identification in the current study

Following Hunston's (2010) and Coxhead and Byrd's (2012) procedure in identifying patterns, manual identification of the 30 shared noun collocations' patterns were carried out. No attempt was made to identify patterns according to their meaning for two main reasons. First, I am not a Computer Scientist, so I would not be able to understand the context of the written text. Second, understanding the context of a written CS text will be difficult, as it comprises many scientific terms and expressions. Thus, the decision was taken to identify and group patterns only linguistically (syntactically) (Hunston, 2010); meaning was not taken into consideration.

5.3.1.3.1 Cleaning concordance lines from erroneously located collocations

Before locating patterns, cleaning collocations that were wrongly located by *ConcGram* was necessary. To check each collocation's concordance lines for any instances of mistaken collocation identification, Shin and Nation's criterion 5 (grammatical well-formedness) was adopted and applied at the phrase level only. That is, two words were defined as a collocation if they occurred in the same phrase. They can be in the same noun phrase (NP), prepositional phrase (PP), or adverbial phrase (ADVP). Collocations that occur in the same phrase were counted and if not were excluded.

Some of the collocating words did not form a syntactic structure according to Shin and Nation's (2008) criterion 5. For example, in extract (1), *data* and *tim*e do not form the collocation *data time* as they occur in two different noun phrases; while in extract (2), *data* and *time* occur in the same prepositional phrase *time of data*. Thus, only extract (2) features the collocation *data time*.

1-…Using the unbalanced ***data*** and reduced the ***time*** for evaluation. Figs… provide the pseudo code… (4RC).[16]

2-Mozilla Firefox had 34 releases at the ***time*** of ***data*** collection developed over four years (6NNS).

In addition, criterion 5 were used with two exceptions. These exceptions were identified using a semantic criterion and a syntactic criterion. These two criteria did not form part of Shin and Nation's (2008) criteria but were developed in consultation with my supervisor committee. The semantic criterion refers to a group of words that are in different phrases but can be paraphrased to form the required collocations. For example, the collocation *following code* in the below extract occurs in the longer NP *the following section of code*, which can be paraphrased to express the same collocation. The following extract was taken from this study's corpora, as is the case from this point for all extracts presented here:

The_ATK ***following_AJK*** section_NNW of_PRF ***code_NNW*** checks_VVZ that_CJT a_ATK graph_NNW is_VBZ(5NNS)

In *the following section of code*, *code* belonged to the noun phrase *section of code* and *following* modified *section of code*, so *following* and *code* were related syntactically. Moreover, *the following section of code* can be paraphrased as *the following code*. Thus, it was classified as a collocation.

The concordance line with the collocation *data access* below also illustrates this semantic criterion:

[16] 4RC means that this line was taken from the reference corpus (RC) and it was line number 4 from the concordance lines of the located collocation.

and_CJC track_VVB a_ATK write_VVB ***access_NNW*** **to_PRP the_ATK protected_AJK** ***data***_NNK using_VVG (3RC)

Because *access* and *data* were separated by 'PRP+ADJ', as can be seen from the extract, *access* and *data* occurred in two different NPs. *Access* occurred in NP as object to the verb '*write*' while *data* occurred in NP related to the prepositional phrase. Thus, it should be excluded if I apply Shin and Nation's criterion 5 only. However, applying the semantic criteria we developed, the phrase '*access to the protected data*' can be rephrased as *protected data access*; that is, it can be rephrased into a single noun phrase. Thus, it was accepted as a collocation.

On the other hand, the syntactic criterion refers to the occurrence of ellipsis, where one word of the collocation is implied. In the extract below, for example, *data* and *information* can form the collocation *data information* even though the words are separated by *swab*. The noun phrase *data or swab information* can be divided into two noun phrases, *data information* and *swab information:*

…it_PNP implies_VVZ that_CJT the_***ATK data_NNK*** or_CJC swab_***NNW information_NNW*** communicated_VVN (5 NNS).

To ascertain whether or not the putative noun collocations detected in my corpus are to be classed as collocations for the purposes of this study, Shin and Nation's criterion 5 relating to restricted phrase level, together with the semantic and the syntactic criterion explained above, were applied in checking all concordance lines of the 30 shared noun collocations.

Consequently, several concordance lines were excluded for a number of reasons. First, if a concordance line did not meet my version of criterion 5, it was excluded. If the two words of the collocation occurred in two different syntactic structures, they were judged as a case of violation of criterion 5. For example:

the_ATK ***following_AJK*** hypothesis_NNW on_PRP ***code_NNW*** complexity_NNW:_PUN Vulnerable_AJK files_NNY (4RC)

The two words of the collocation *following code* in the extract above were not parts of the same phrase. *Following* belongs to the NP *the following hypothesis* and *code* belongs to the prepositional phrase *on code complexity*. This means that they violate criterion 5 because each of them belongs to a different syntactic structure. Consequently, this line was excluded..

Second, some collocations were wrongly tagged by CLAWS. Tagging-errors were one of the main problems encountered by Mason and Hunston (2004) in developing their automatic system recognition for verb patterns and Coxhead and Byrd (2012) therefore highlight the importance of human manual checking of computer software analysis. Ackerman and Chen (2013) checked all their lists of academic collocations for any POS tagging errors as one of the main steps towards refining their collocation lists. Even though I checked and corrected wrongly tagged wordlists developed in the first stage of the study (see section 3.6), some POS tagging-errors were found. For example, *type* that was the collocate word of *data* was wrongly tagged as a noun in the following extract, while it should be a verb:

…their_DPS ability_NNW **to_PRP** ***type_NNW data_NNK*** in_PRP quickly_AVK and_CJC flawlessly_AVK (23 NS)

Thus, this concordance line was excluded since the collocation *type data* did not fall under the categorisation of noun collocations 'N+N' (Hunston and Francis, 1996, 2000).

Third, if a concordance line includes Computer Science names of programs or systems that form part of the collocations, it was excluded. For example, the following concordance line that included the name *ByDesign System* was included in the computer-generated *design system* collocation list, but was manually excluded:

For_AVK example_AVK, _PUN the_ATK SAP_NNW ***ByDesign_NNW system_NNW*** has_VHZ thousands_CRD of_PRF (1RC)

In some cases where the concordance cut-off makes it difficult to appreciate the meaning in context, checking the full context is required to judge whether the collocation met my version of criterion 5. For example, it cannot be decided from the extract below whether the collocation *system design* was in the same phrase or not. Thus, checking the full context was necessary.

… of_PRF the_ATK ***system_NNW*** architecture_NNW ***design_NNW*** should_VMK be_VBI taken_VVN with_PRP care_NNW (18NNS)

CONTEXT: The_ATK system_NNW architecture_NNW of_PRF the_ATK web_NNW application_NNW being_VBG developed_VVN will_VMK be_VBI illustrated_VVN in_PRP the_ATK following_AJK sections_NNY:_PUN Architectural_AJK Considerations_NNY.

The_ATK decision_NNWof_PRF the_ATK system_NNW architecture_NNW design_NNW should_VMK be_VBI taken_VVN with_PRP care_NNW.

After checking the full context, I can see that *system* and *design* were in the same noun phrase *system architecture design*. Thus, this line of concordance was included in the analysis. However, in the following extract *system and design* occurred in two different noun phrases but checking the context was important to clarify whether this example should be excluded from the analysis:

…definition_NNW ***System_NNW*** and_CJC software_NNW ***design_NNW*** Implementation_NNW and_CJC unit_NNW testing_NNW (19NNS).

Context: Normally_AVK, _PUN there_EXK are_VBB five_CRD stages_NNY in_PRP a_ATK systems_NNY development_NNW Requirements_NNY analysis_NNW and_CJC definition_NNW ***System_NNW and_CJC software_NNW design_NNW Implementation_NNW*** and_CJC unit_NNW testing_NNW Integration_NNW and_CJC system_NNW testing_NNW Operation_NNW and_CJC maintenance_NNW…

It can be seen that *system* belongs to the noun phrase *definition system* whereas *design* belongs to the noun phrase *software design*. This means that they violate criterion 5 and the line was not counted in the analysis.

5.3.1.3.2 Re-checking the significance of the 30 shared N collocations

Having checked manually all concordance lines of the 30 shared N collocations, the next step was to re-check the significance of these collocations with reference to the RC. A chi-square test

was carried out. As a result, six collocations (*data other, data web, code number, design implementation, data available*, and *data different* (see Appendix F for more details of the results)) were non-significant since the Fisher exact test resulted in a *p* value that was (>0.05) (for detailed information about the chi-square test and Fisher exact test, see section 4.5.4). Thus, they were excluded. The remaining 24 collocations were examined and their patterns were analysed in detail.

5.3.1.3.3 Identifying Patterns

To identify patterns for each collocation from the three corpora, their concordance lines were first extracted from each corpus and then grouped into a single text file. Following Hunston's (2010) steps of pattern identification, I first looked at the collocation in all concordance lines and identified similarities between and among the corpora.

For example, looking at concordance lines of *data access* from the RC, I can observe that there are only three patterns:

1 worth_NNW of_PRF application_NNW server_NNW ***access_NNW*** log_NNW ***data_NNK*** to_TO0 simulate_VVI user_NNW

2 _PUN and_CJC track_VVB a_ATK write_VVB ***access_NNW*** to_PRP the_ATK protected_AJK ***data_NNK*** using_VVG

3in_PRP terms_PRP of_PRP ***data_NNK*** object_NNW ***access_NNW***._We_PNP note_VVB here_AVK

1- *Access* log *data*

2- *Access* to the protected *data*

3- *Data* object *access*

Then, patterns were also identified in each of the students' corpora following the same steps.

Observing concordance lines from the NNS corpus, three patterns were identified:

NNS

1 being_VBG developed_VVN, _PUN the_ATK ***Data_NNK Access_NNW*** layer_NNW should_VMK contain_VVI the_ATK
2 Logic_NNW Layer_NNW, _PUN and_CJC ***Data_NNK Access_NNW*** Layer_NNW ._Any_DTK changes_NNY
3 Business_NNW Logic_NNW Layer_NNW *_UNC ***Data_NNK Access_NNW*** Layer_NNW The_ATK User_NNW Interface_NNW
4 Interface_NNW layer_NNW and_CJC ***Data_NNK Access_NNW*** layer_NNW. The_ATK Data_NNK
5 The_ATK ***Data_NNK Access_NNW*** layer_NNW should_VMK contain_VVI all_DTK the_ATK
6Data_NNK Access_NNW Layer_NNW The_ATK ***Data_NNK Access_NNW*** layer_NNW has_VHZ a_ATK class_NNW named_VVN
7Business_NNW Logic_NNW layer_NNW or_CJC ***Data_NNK Access_NNW*** layer_NNW. The_ATK website_NNW
8directly_AVK with_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW. Instead_AVK, _PUN
9is_VBZ contained_VVN in_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW to_TO0 communicate_VVI with_PRP
10 the_ATK classes_NNY in_PRP the_ATK ***Data_NNK Access_NNW*** Layer_NNW. Group_NNW class_NNW
11 this_DTK class_NNW are_VBB The_ATK ***Data_NNK Access_NNW*** Layer_NNW. The_ATK Data_NNK
12 The_ATK ***Data_NNK Access_NNW*** layer_NNW should_VMK have_VHI all_DTK the_ATK
13class_NNW in_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW
14see_VVB Appendix_NNW B._NPK The_ATK ***Data_NNK Access_NNW*** Layer_NNW The_ATK Data_NNK Access_NNW layer_NNW
15(_PUL business_NNW logic_NNW and_CJC ***data_NNK access_NNW***)_PUR can_VMK make_VVI the_ATK application_NNW
16 the_ATK parameters_NNY to_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW.The_ATK Business_NNW
17 is_VBZ located_VVN in_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW
18logic_NNW from_PRP the_ATK ***Data_NNK Access_NNW*** layer_NNW and_CJC User_NNW Interface_NNW
19 and_CJC external_AJK ***data_NNK source_NNW Access_NNW*** errors_NNY

These patterns are:

1- *Data access*

2- *Data access* followed by layer, which expresses a specific name of layer of data access used in Computer Science

3- *Data* source *access*

Turning to the last set of concordance lines extracted from the NS corpus, two patterns were observed.

NS

1…intermediate_AJK layers_NNY to_TO0 gain_VVI ***access_NNW*** to_PRP the_ATK desired_AJK ***data_NNK*** ._SENT
2…intermediate_AJK layers_NNY to_TO0 gain_VVI ***access_NNW*** to_PRP the_ATK desired_AJK ***data_NNK*** ._SENT
3…The_ATK ***data_NNK access_NNW*** components_NNY present_VVB in_PRP this_DTK
4…presentation_NNW aspects_NNY with_PRP ***data_NNK access_NNW***aspects_NNY if_CJS poorly_AVK written_VVN ._SENT
5…processes_NNY logic_NNW and_CJC the_ATK ***data_NNK access_NNW***
6 …presentation_NNW aspects_NNY with_PRP ***data_NNK access_NNW*** aspects_NNY if_CJS poorly_AVK written_VVN ._SENT
7…processes_NNY logic_NNW and_CJC the_ATK ***data_NNK access_NNW***.The_ATK data_NNK server_NNW
8…The_ATK ***data_NNK access_NNW*** components_NNY present_VVB in_PRP this_DTK

These patterns are:

1-*Access* to the desired *data*, which can be paraphrased as *desired data access*

2-Data access

Then, patterns were tabulated so that similarities and differences would be clear:

Table **5-1**: RF, NF, and number of users for the patterns of *data access* in each corpus

Corpus	data access		data+noun+access		Access+(prp+adj)+data		Access+noun+data	
	NF	No.of users	NF	No. of users	NF	No. of users	NF	No. of users
RC			0.16	1/63	0.16	1/63	0.16	1/63
NNS	5.9	1/29	0.33	1/29				
NS	2.7	3/26			1.30	2/26		

As a result, *data access* is associated with four patterns, three of which were shared between two corpora. I can group the first pattern located in the RC, *data object access*, in which *data* and *access* were separated by a noun with the second pattern of the NNS corpus, *data source access*. Another similar pattern was also located between the RC and the NS corpus in which *access* and *data* are separated by a prepositional phrase.

Taking into consideration Hunston's (2010) procedure of rechecking identified patterns for any possible merging, the third pattern, '*access+PRP+ADJ+data*,' can be merged with the first pattern, *data access*. Moreover, applying the semantic criteria specified above, the phrase '*access to the desired/protected data*' can be paraphrased as '*desired/protected data access*'.

Interestingly, similar overuse of the first pattern *data access* was detected by both NNS and NS students. Nevertheless, surprisingly, this collocation was not used by the expert writers. Another notable pattern '*access+N+data*' was used only by expert writers and did not occur in either the NNS or NS students' writing. This variation could be explained if I consider the topics that were in focus in each corpus. It could be that the students' corpora consists of topics that would be associated with the *data access* collocation more than the expert writers' topics in the RC. The single occurrence of the pattern used by expert writers could be related to the writer's personal style.

To determine which of these patterns were significantly over/underused as compared to the RC, a chi-square test was computed. If the test result is less than 0<.05, it will be significant. Following this procedure, the remaining 23 collocation patterns were identified.

5.3.2 Categorisation Judgement Task (CJT)

5.3.2.1 Aim of the categorisation judgement task

To verify my findings from dictionaries' check of whether the 49 collocations displayed in section 4.6 could be categorised as GAC or GCSC, CS experts from the School of Computer Science and Electronic Engineering at the University of Essex were asked to complete the categorisation judgement sheets.

5.3.2.2 Design

49 N collocations were included in the CJT. These were the 31 N collocations shared between the NNS and NS corpus, the top 10 overused N collocations from each of the students' corpora excluding the 10 overlapping collocations, four underused N collocations from both students' corpora, and four N collocations missing from both the NNS and NS students' corpora. For detailed information about the CJT, see Appendix G.

Since some collocates were adjacent and others were non-adjacent, the decision was made to have two separate sections in the CJT of these two kinds of collocations so that respondents can clearly recognise the difference in their uses. The adjacent noun collocates were all presented in

one table whereas the non-adjacent collocates were presented individually followed by two concordance lines. An example of adjacent and non- adjacent collocates are given below.

Figure 5-3: An example of adjacent collocate presentation in the CJT.

Phrases with adjacent words	**General academic phrases**	**General CS academic phrases**	**Specific CS academic phrases**			
			Artificial Intelligence	**Software Engineering**	**Information System**	**Comments**
1-code following/ following code						

Figure 5-4: An example of non-adjacent collocate presentation in the CJT.

method …class

As shown in the following extracts from Computer Science students' writing

1-…to the document using the ***method*** of the Dataset ***class***.

2-…It uses the ***method*** from the Membership ***class…***

General academic phrases	**General CS academic phrases**	**Specific CS academic phrases**			
		Artificial Intelligence	**Software Engineering**	**Information System**	**Comments**

Another issue was related to the degree of specificity of the GCSC, that is, whether these collocations were discipline-specific in their uses. Looking for GCSC in a CS dictionary was not enough to reveal whether a collocation was specifically used in one of the selected CS sub-disciplines, thus it was decided to add the third category Specific Computer Science Collocation (SCSC) to the CJT so that CS experts will be able to classify sub-disciplinary differences. Thus,

three categories were defined and exemplified to the respondents. To avoid misunderstanding of what collocation means to CS experts, the term *phrases* were used instead. Respondents were also provided with detailed instructions and were given some examples in order to complete the task successfully as shown below.

Figure 5-5: Definitions and examples of the three types of collocations provided in the CJT.

a- **General academic phrases** (these phrases can be found in Computer Science as well as in other academic disciplines, e.g. *available data, different components*)
b- **General Computer Science (CS) academic phrases** (these phrases can be found in Computer Science only, but in ANY discipline of Computer Science, e.g. data input)
c- **Specific Computer Science (CS) academic phrase** (these phrases can be found in Computer Science only, but can only be found in certain disciplines of Computer Science, e.g. *network traffic: in the sub disciplines of software engineering and information systems only*).

phrases	**General academic phrase**	**General CS academic phrase**	**Specific CS academic phrase**			**Comments**
			Artificial Intelligence	**Software Engineering**	**Information Systems**	
1-Available data	√					
2-Data input		√				
3-Netwrok traffic				√	√	A very common phrase in some types of CS.

Figure 5-6: Detailed instructions and examples given in the CJT.

In the first example, the Computer Science specialist felt that the phrase *available data* can be found in ANY or ALL disciplines, not only Computer Science, and so s/he ticked the **'General Academic phrase'** box. In the second example, the Computer Science specialist felt that the phrase *data input* is a phrase used in Computer Science only and can be used in ANY discipline of Computer Science and so s/he ticked the **'General CS academic phrase'** box. In the third example, the Computer Science specialist felt that the phrase *network traffic* is a phrase used only in SPECIFIC types of Computer Science, Software Engineering, and Information Systems, but not in Artificial Intelligence. S/he has also added a comment, saying *network traffic* is very common in certain fields of Computer Science.

5.3.3 Expert Interviews

5.3.3.1 Aim

In addition to the CJT, semi-structured in-depth interviews were conducted with three CS experts in order to gain a deeper understanding of the over/underuse of some of the collocations and their located patterns. They were also used to ask experts their opinions of which of the preliminary factors found from my analysis and from my supervisors' analysis were more important. These factors were genre, topic, NNS vs NS in their writing style, Expert vs Novice writers in their use of language, and personal style.

Genre could be one of the main factors behind the over/underuse of some of the collocations since dissertations and journal articles are different genres and therefore may exhibit different collocation patterns due to differing genre requirements and norms (e.g., Harwood, 2005, 2006; Hyland, 2004, 2008; Samraj, 2002). Another factor that could explain the various use of the collocations among corpora could be the topics of the texts in which collocations appear. It has

been noted by Peacock (2012) in his investigation of noun collocation from eight different disciplines that some collocations were restricted in their use due to their topics.

From my preliminary checks of MSCs dissertations' topics, most of the collocations occurred in various topics. Due to the researcher's limited knowledge of CS topics, it could not be verified whether these collocations were topic-specific or not, thus, CS experts were consulted (see Appendix H for detailed information about the topic classification of some of the 49 N collocations).

Another factor that could explain the variation of the use of N collocations between NNS and NS students could be related to their different writing style (e.g., Altenberg and Granger, 2001; Bunton, 2002; Hyland, 2004; Samraj, 2008). Perhaps NNS might use long extended collocations, unlike NS. Moreover, expert writers write in different ways if compared to novice writers (students writing in this study) (Bereiter and Scardmalia, 1987; Geisler, 1994; Tardy, 2009) could perhaps explain the various uses of the N collocations. Hyland (2008) found variation between students and experts in their use of lexical bundles in their academic writing. The last factor was related to writers' personal style; each writer has his own style of writing.

5.3.3.2 Respondents

I was advised by my supervisory committee to conduct interviews with CS experts who specialise in one of the three CS sub-disciplines considered in this thesis. Interviewing two experts from each sub-discipline will yield a fair amount of subjective opinions about collocation use in that discipline.

After I checked CS experts' profiles online from the School of Computer Science and Electronic Engineering at the University of Essex, twelve CS experts were selected according to their specialisations, four specialists in each sub-discipline. No concern was given to respondents' nationalities as the focus was on respondents' specialisations. I contacted all of them via e-mail asking for their participations (see Appendix I for more details about the e-mail sent to the CS experts asking for their participations).

From their initial replies, six of the twelve selected CS experts replied positively, two from each sub-discipline. They were all asked to complete the CJT and to return it to me before giving their interview. Thus, enough time was given (2-3 days) to compare between experts' categorisation with my categorisation, which was based on the process of dictionary consultation described in section 4.5.5. Only four respondents completed the CJT. Information about their specialisations, positions in the CS Department, and their working experiences is presented in the following Table.

Table 5-2: Detailed information about respondents' position in the CS Department, specialisations, and working experience.

Respondents	**Position in the CS Department**	**Specialisation**	**Working experience**
P1	Senior Lecturer	AI	20 years
P2	Reader	SE and AI	25 years
P3	Reader	IS	20 years
P4	Professor	IS	20 years

The interviews were conducted with the first three CS experts, one from each sub-discipline. The fourth respondent who completed the CJT withdrew from the study due to departmental commitments. I attempted to recruit additional respondents to address these withdrawals but was unsuccessful.

5.3.3.3 Interview design

The interview consisted of three main parts: general questions about the MSc dissertation requirements in the School of Computer Science and Electronic Engineering at the University of Essex, questions related to the CJT, and detailed questions about some collocations' use and patterns(for more information about each section of the expert interview, see Appendix J). After respondents replied to the general questions related to the requirements of writing MSCs in their department, they were asked to comment on their categorisation of some collocations that did not match the dictionary findings. For example, *design system* was classified as general academic phrase. I looked up its meaning in the previous mentioned dictionaries but it was categorised as specific academic phrase for CS only. Thus, CS experts were asked for explanation.

P2 explained in detail how *design system* is related to CS as follows:

> *"the two words come together only in CS. I know it might be used in other ways but it is very colloquial. But in CS, it is a very specific term. System design or design system is a topic we teach our students about. I have actually got a book here…in Software Engineering; there is a lot about design system, by system we mean information system. It is a big area in CS and everybody who works in CS should learn it. It has very specific meaning because of the word 'design'. Design here means how the parts fit together. And system is related to the whole picture of the design program".*

The third section was related to some of the 24 shared collocations' use and patterns. Respondents were given the table of results of some of the collocations to comment on and to

explain the results according to their subject knowledge. An example of the collocation *environment development* table of results (see below)was given and then followed by general questions first to encourage respondents provide their own explanations.

Figure 5-7: An example of the collocation *development environment* table of results**.**

Corpus	Development environment	
	Normalised frequency(NF)	No.of users
Expert writers(journal articles)	.83	1/63
Non-Native writers	3.9	8/29
Native writers	16.9	12/26

1-*You can see from the table that both native and non-native students use environment development more than the expert Computer Scientists writing journal articles. Please comment on why you think this may happen.*

Questions about other factors that may explain the different uses of collocation among corpora were asked. These factors were genre, topic, NS vs NNS in their use of language, experienced and inexperienced writers' use of language, and personal style. A sample of these questions are presented below.

Figure 5-8: Sample questions from the interview about topic and other factors suggested.

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?

3-To what extent do you think that the dissertation or journal article **topic** might affect writers' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

Topics:

NS 1: Implementation of Game Agents in Unreal Tournaments

NS13: Mobile Phone Training for the Elderly People

NNS14: Advanced Web Application Programming

NNS23: Intelligent Web Search Using Named Entity Recognition

RC 16SE: A Logical Verification Methodology for Service-oriented Computing.

4. Can you think of any other reasons which may explain why only the native and non-native students use this phrase?

5. Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

5.3.3.4 Procedure

The interviews were conducted over a period of 10 weeks. They were held in respondents' offices and were tape-recorded. Three tapes were used and each one was labelled with the interviewee's name for data to be well-organised and ready for transcription and analysis.

Before starting the interviews, all respondents were thanked for their agreement to take part in this study and were reminded about the anonymity of the interviews and confidentiality of the recorded tapes. At the beginning of each interview, a brief description of the structure of the interview was given. Each interview lasted between 70- 90 minutes. Interviews were then transcribed to organise them into a manageable and analysable base of information (Mackey and Gass, 2005). One of the CS expert interview's transcription is given in Appendix K.

5.3.3.5 Data coding

After full transcriptions of all the interviews were made, including pauses and repetitions, a list of codes were generated using techniques adapted from Miles and Huberman (1994), Coffey and Atkinson (1996), Dornyei (2007), and Saldana (2009). Codes are defined as "tags or labels for assigning unities of meaning to the descriptive or inferential information compiled during study. Codes usually are attached to chunks of varying size-words, phrases, sentences or whole paragraphs, connected or unconnected to a specific setting" (Miles and Huberman, 1994:56). Codes are used by researchers "to retrieve and organise the chunks [of text]… so the researcher can quickly find, pull out and cluster the segment relating to a particular research question, hypothesis, construct, or theme" (Miles and Huberman, 1994:57).

According to Dornyei (2007:253), researchers can define a list of codes "as a result of preliminary scanning of the data" or when they have "sufficient background information" on the topic under study. Miles and Huberman (1994) suggested developing a list of codes with their examples and definitions to be used. Thus, the idea of developing my initial list of codes were drawn from two sources: (1) the notes I kept during the preliminary scanning of the data; and (2) the research questions of the present study.

With regard to the first source, I made some initial codes and placed them besides the quotes throughout my first reading. For instance, genre effect (writing in dissertations is different from writing in research articles) was one of the codes for the following extracts:

(1)P3 [source code]: Yeah, I think students use it more as they are talking about programming in their dissertations. Expert writers might use it in a moderate way, not to mention programming in detail

(2)P3: [following code]: I think because dissertations are focused more on industrial professional style rather than academic writing, both NNS and NS are not writing in an academic way. Their formal reports may include some academic writing, but it cannot be compared to research articles.

(3)P3 [development environment]: I think the low frequency of expert writers; it ideally talks about things relatively practical. If you had a theory or a problem and you want to develop a solution, you normally talk about those issues rather than talking about development environment. Development environment is a kind of computer software so everybody knows about it, so you do not need to talk about it.

(4) P3 [method class]: No effect.

For my revised list of codes, I grouped codes thematically by comparing their similarities and differences. Thus, numbers of sub-codes were grouped under one code. For example, extracts (1), (2), and (3) were given the code 'genre effect' but when I compared to other extracts (4) that were also coded as genre effect they were different in their focus. Thus, the code genre effect was revised to have number of sub-codes: writing in dissertations vs writing in research articles (1), dissertations' writing demands (2), research articles' writing demands (3), and no effect of genre (4). A sample of thematic coding is given below:

(1)P3 [source code]: Yeah, I think students use it more as they are talking about programming in their dissertations. Expert writers might use it in a moderate way, not to mention programming in detail (writing in dissertations vs. writing in research articles).

(2) P3 [following code]: I think because dissertations are focused more on industrial professional style rather than academic writing, both NNS and NS are not writing in academic way. Their formal reports may include some academic writing, but it cannot be compared to research articles (dissertations' writing demands).

(3) P3 [development environment]: I think the low frequency of expert writers; it ideally talks about things relatively practical. If you have a theory or a problem and you want to develop a solution, you normally talk about those issues rather than talking about development environment. Development environment is a kind of computer software so everybody knows about it, so you do not need to talk about it. (research articles' demands).

(4) P3 [method class]: No effect (genre-no effect).

The second source informing the shaping of the revised list of codes was "to start from foreshadowed research question[s] that inspired the research project" (Coffey and Atkinson, 1996:32). Since the third research question sought to be answered by experts' explanations, therefore, questions asked in the third section of the interview were all related to the main factors that sought to affect the use of the collocations by students and experts. I added new codes about experts' comments on the use of the collocations as well as other factors that emerged from my analysis; genre, topic, experts vs. novice writing and NS vs. NNS in their writing style, and personal style. Table 5-3 shows the list of codes for the factors which were thought to affect the use of the collocations.

Table 5-3: List of codes generated for the factors affecting collocations' use and patterns

1- Genre effect
a- Writing in dissertations vs. writing in articles
b- Dissertations' writing demands
c- Articles' writing demands
d- Genre-no effect
2- Topic specific collocations
a- Agreement
b- Disagreement/Uncertainty
c- General collocations – not topic-specific
d- Specific collocations – topic-specific
3- Other factors mentioned(a, b, c)
a- NNS vs. NS in their use of language (effect/no effect)
b- Experience and non-experienced writers (effect/no effect)
c- Personal style (effect/no effect)
d- Other factors(a,b,c) – No effect
4- **Interviewees' additional reasons**

a- Cultural factors
b- UK-oriented or US-oriented collocations
c- Subject-related collocations
d- Use of equivalent L1 terms

Before I began my data analysis, two-second judges, who are my supervisors, were asked to check the reliability of the codes. The total percentage of agreements between the second raters was 90%, which is quite satisfactory, being above the minimum acceptable agreement percentage as indicted by qualitative scholars (Mackey and Gass, 2005:244; Miles and Huberman, 1996:64). Slight changes were made for some codes. For example, extract (5) was coded under students' writing style, but the second inter-rater suggested coding it under lack of writing competence. Thus, it double-coded as students' writing style+ lack of writing competence.

(5) Students may not thinking of writing in a professional way; they just write in a linear style. They are not writing to publish their work. This is the reason. Another issue is that some students are repetitive in their writing; they just keep mentioning the same word distributed everywhere in their dissertation. They do not have that sense of narrative flow [that occurs] in academic writing.

For detailed information about the codes and their definitions, see Appendix L. The analysis of themes and categories created from the coding procedure are presented in detail with the results.

5.4 Results and Discussion

5.4.1 Categorisation Judgement Task Results and Discussion

The four CS experts who completed the CJT agreed in most of their categorisation; they disagreed only for a few collocations (6%); similarly, they could not categorise only a few collocations (8%).

Table 5-4 shows that there was a great match between dictionaries' categorisation and CS experts' categorisation for the General Academic Collocations (GAC) but a great mismatch between dictionaries' categorisation and CS experts' categorisation for General Computer Science Collocations (GCSC). This could be explained by the various categories given in the judgement task. When definitions of the 49 collocations were checked in the two dictionaries mentioned previously in section 4.5.5, the specificity of collocations in the selected CS sub-disciplines (AI, SE, IS) could not be identified due to the limited information given in the dictionaries; no specific CS sub-disciplines were mentioned. For detailed results of the CJT, see Appendix M.

Table 5-4: Percentages of CS experts' dis/agreement with dictionaries' information.

Categorisation	**GAC**	**GCSC**	**SCSC**	**GAC/ GCSC**	**GCSC/ SCSC**	**Various Marks**	**Not Marked**
No. of collocations (49)	15	13	6	7	1	3	2
Agreement between CS experts	30.6%	26.5%	12%	14%	2%	6%	8%

Dictionaries' agreement	25%	65%	-				10%

However, few collocations were categorised as SCSC to certain sub-disciplines. *Data layer* and *layer application* were categorised as collocations specific to IS while *query document*, *document cohesion*, *document ranking*, and *data training* were marked as collocations specific to AI. This finding may be seen as incongruent with Hyland and Tse's (2007) objection to the AWL (Coxhead, 2000) as they claimed that academic words have different collocations in different academic disciplines.

Even though collocations of some AWL words seem to be discipline specific, as claimed by Hyland and Tse (2007), there are a number of collocations that were categorised as GAC by CS experts and were found in the new Academic Collocation List (ACL) developed by Ackermann and Chen (2013). These are *available data, available resources, previous section, following section*, and *process development*. This finding could confirm the usefulness of AWL in locating collocations in various academic disciplines.

The data in Table 5-4 also indicates that GAC and GCSC were the most frequent collocations used by CS students. This evidence is in line with our hypothesis that the most frequent collocations could be the overused ones and, therefore, might not be considered problematic to students as they encounter them frequently. Jones and Durrant (2010) suggest that the most frequently used collocations are being stored in learners' mental inventories and therefore become 'entrenched' in their lexical production.

5.4.2 Categorisation Difficulty

Various explanations were given when CS experts were asked about the collocations they had not categorised. The difficulty of categorising some of the collocations was due to their meaning. For example, *data information* consists of words that seem to carry the same meaning. P3 commented on this collocation by describing it as "[a] *strange phrase. As a pair of words, I have not recalled seeing it before. I do not know what it means, 'data information' or 'information data' ... actually these two words have the same meaning, so I wonder how they occur together as compounds. This is why I have trouble classifying it. This is not a kind of English I would write. You could say that it is a GA phrase, but I would rather say that even general academics will not use it in this way*".

Moreover, the difficulty of categorising some collocations as GCS or SCS was mentioned by one of the CS experts (P3): "*some of them were very tricky ... Some terms could be categorised in both of them; some were very hard to say. Actually, there were no clear cut connections between these two categorisations.*" The difficulty of categorising some collocations as general or specific to CS seems to be in agreement with Spack's (1988) claims that the specificity should be limited to some extent in some cases in order not to be too specific.

Other CS experts found some collocations difficult to categorise as they commented that these collocations' use have changed over time. For example, *computer vision, network traffic,* and *layer application* were specific collocations to certain sub-disciplines but nowadays they are used in all CS and are thus considered GCSCs, as P3 commented" :*It is kind of becoming more*

general CS. It is kind of debatable. Layer application occurs a lot in CS. You could argue it is related to SE or IS, but it is mainly related to networking."

P2 provided similar explanation to the use of *network traffic* by saying, *"This term used to be very specific to IS, but nowadays it can be used in any discipline of CS. It is more commonly used than before. It was used in network specialisation, but since almost everything we do is related to network traffic it becomes a common term.*" Similarly, P3 categorised *computer vision* as GCSC for the same reason. He said, "*Computer vision was more specific to AI in the 1990s; it has its roots in artificial intelligence, but now it becomes more mainstream. Overtime, this term becomes more general in CS.*"

On the other hand, two collocations were problematic for all CS experts as none of them categorised them. It seems that *neutral files* and *vulnerable files* are not common collocations in CS, as P1 commented:"*It is not a phrase I have heard before … my immediate thought will be there are some files which were hostile depraved … other files are normally depraved … But it is like a neutral opinion. I could not think of an ordinary use of that phrase*." Even when I showed respondents some examples of these collocations' use, they could not categorise them; as P3 said, "*I was reluctant to categorise neutral files. I cannot imagine context with this one. I struggle to classify this one. (Showing examples) … it is really hard to say. I am not sure*."

Since these two collocations were selected from the top ten missing N collocations from both NNS and NS students' corpora, the difficulty that the CS experts faced when they tried to categorise them may suggest their infrequent uses in CS.

5.4.3 Pattern Identification Results and Experts' Views

A different number of patterns were identified for each of the 24 shared collocations, ranging from one to six patterns. These patterns will be displayed in detail in this section to answer the three research questions related to the collocation patterns:

RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora?

RQ5. To what extent do the shared collocations differ in their patterns?

RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views?

RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?

To address the first question, number of patterns for each collocation will be first identified and then counted in the three corpora to be checked with my pattern hypothesis according to which number of patterns could explain the over/underuse of collocations. The rationale behind this collocation pattern analysis is that a collocation may be overused in one or both of the student corpora because it appeared in patterns that occurred rarely, if at all, in the RC. On the other hand, a collocation may be underused in one or both of the student corpora if it appeared in fewer patterns in them than in the RC. Then the last three questions will be answered in parallel: each collocation pattern will be displayed individually so that similarities and differences across

corpora will be obvious. Explanations given by CS experts about the over/underuse of some of the collocation as well as about the reason for using different patterns will be discussed.

RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora?

Following Hunston and Francis' (1996) steps for pattern identification, a different number of patterns were identified for each collocation in the three corpora, as can be seen in Table 5-5.

Overall, the over/underused patterns hypothesis regarding the number of patterns used among corpora was only supported in the comparison between NS and RC as t (23) =-1.683, p= 0.05. While there was no significant difference between NNS and RC as the paired test was t (16) = 0.169, p0>05. However, some individual collocations have been overused by NNS or NS students because of their use of more patterns than the patterns used in the expert writers' corpus.

Table 5-5: Over/under used patterns' hypothesis were checked for the 24 shared N collocations

Collocations	NNS No. of patterns	NS No. of patterns	RC No. of patterns	Overused hypothesis	Underused hypothesis
code following	1	3	1	√ NS only	
data layer	2	2	1	√ both	
data amount	3	2	5		√ both
data access	2	2	3		√ both
data user	5	2	3	√ NNS only	

data information	3	2	1	√ both	
data time	3	1	2	√ NNS only	
data type	3	2	3		√ NS only
design system	4	3	3	√ NNS only	
development environment	1	1	1		
features other	2	1	2		√ NS only
layer application	1	1	2		√ both
network traffic	4	4	2	√ both	
resources available	2	2	2		
resources system	1	2	2		
method class	1	1	4		√ both
code source	1	1	2		√ both
data input	0	3	2	√ NS only	
data structure	0	1	2		
section previous	0	1	1		
section following	0	1	1		
site web	1	0	1		
source open	0	1	1		
components different	0	3	2	√ NS only	

16 of the 24 shared N collocations met my pattern hypothesis (that is, a collocation may be overused in one or both of the student corpora because it appeared in patterns that occurred rarely, if at all, in the RC. An underused collocation can be explained if the RC includes patterns of a collocation that appear not at all or rarely in one or both of the student corpora). Three N collocations met the overused pattern hypothesis since both NNS and NS students used more patterns than the expert writers did. Only three N collocations were overused only by NS

students (*code following, different components,* and *data input*) and three other N collocations were overused by NNS students only (*data user, data time,* and *design system*) because they used more patterns compared to expert writers. In addition, five other N collocations (*data amount, data access, layer application, method class,* and *code source*) met the underused-pattern hypothesis by both NNS and NS students. That is, students underused these collocations due to their limited use of patterns, unlike expert writers.

Interestingly, other N collocations were used similarly by students and expert writers. A tentative explanation might be that some of these collocations could be fixed terms in CS and there is therefore only one way of expressing them. For example, *development environment* and *web site* each had only one pattern. It was evident from their definitions in the online *Oxford English Dictionary* that they are fixed terms in CS.

***Development environment**: "a computer system, including hardware and software, that is specifically designed to aid in the development of software and interfaces" (online Oxford English Dictionary).*

***Web site**: "a document or a set of linked documents, usually associated with a particular person, organization, or topic, that is held on…a computer system and can be accessed as part of the World Wide Web" (online Oxford English Dictionary).*

To summarise, the use of different patterns for expressing the same collocation by students and experts could be considered as one of the reasons behind the over/underuse of 16 of the 24 shared N collocations. However, the findings indicate that differences in pattern variation between the student corpora and the RC were not significant for many collocations, so it seems that other factors behind collocation over/underuse for these collocations and also, possibly, even for the collocations whose pattern number differ significantly between the RC and one/both of the student corpora. Both my analysis and CS experts' views will be discussed in the following section so that other suggested factors will be investigated.

Each of the 24 collocation patterns will be compared between NNS and NS students' corpora, and between both students' corpora with the RC. Moreover, each pattern's frequency will be presented as a normalised frequency (NF) (per 100,000 word tokens) in each corpus. CS experts' explanations for some overused or underused collocations will also be discussed in the following section.

RQ5. To what extent do the shared collocations differ in their patterns?

RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views?

RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?

The data in Table 5-5 indicates that different numbers of patterns were identified for the 24 shared collocations. Thus, the 24 shared collocations were grouped into three groups according to their number of patterns: single pattern collocations, two pattern collocations, and three-plus pattern collocations. Each group will be presented respectively in the following section.

5.4.3.1 Single Pattern Collocations

Five of the 24 collocations had only one pattern. These are *web site*, *development environment*, *open source*, *previous section,* and *following section*. Their patterns will be presented in turn.

Development Environment

As can be seen from Table 5-6, only one pattern was associated with the collocation *development environment*, which was *development environment*, as shown in the concordance lines below:

1…the_ATK client_NNW project_NNW in_PRP the_ATK ***development_NNW environment_NNW***, _PUN which_DTQ automatically_AVK (1RC)

2…the_ATK user_NNW in_PRP the_ATK Eclipse_NNW ***development_NNW environment_NNW*** and_CJC take_VVB the_ATK form_NNW (2RC)

Table 5-6: RF, NF, and number of users for the patterns of *environment development* in each corpus

Corpus	**Development environment**		
	Raw frequency	NF	No. of users
Reference	5	.83	1/63
NNS	12	3.9	8/29
NS	50	16.9	12/26

Even though the pattern was overused by both NNS and NS students, it was significantly overused (Fisher's chi-square was 36.818, $p<0.0001$) by NS students but not by the NNS students, as the chi-square test was non-significant (Fisher's chi-square test was 2.881, $p>0.05$). The fact that this collocation occurs in all corpora with the same pattern would probably be related to its specificity in CS. This was confirmed by CS experts' categorisation of *development environment* as GCSC. Moreover, the dictionary definition indicates that this collocation is used more as GCSC rather than SCSC. *Development environment* refers to "a computer system,

including hardware and software that is specifically designed to aid in the development of software and interfaces" (*online Oxford English Dictionary).*

The overuse of this collocation by students rather than by experts could be explained if I compare the demands of writing in two different genres. Writing in dissertations is different from writing in research articles. First, students have more room to describe in detail the construction of their software programs when writing their dissertations, unlike experts who are restricted to a limited number of words in their writing of research articles. Thus, there is less of a requirement for journal article authors to describe the step-by-step development of their program. As P1 commented, "*I think that people writing dissertations are more interested in describing the construction of their programs, while people writing in journals are just assuming their programs are written*". The extract below confirms P1's explanation that experts only mention that their programs are developed:

We found three client programs that met our study criteria: a task-focused environment, a ***development environment*** *for the JBoss web application server, and the Java debugging environment in Eclipse. Once we analysed the framework's source history, we tried to compile the first version of the client programs with Eclipse 3.3. For each call to a framework method that could not be resolved by the compiler, we ran the SemDiff recommender and noted its recommendations. (16SE)*

More explanation about the effect of genre on the use of *development environment* was given by P3: "*if you have a theory or a problem and you want to develop a solution, you normally talk about those issues rather than talking about development environment. Development environment is a kind of computer software so everybody knows about it, so you do not need to talk about it. You do not need to waste your words talking about it, nobody cares*."

P2 added: *"I would assume that people who write in journals will not talk about software used in developing another program. But students in all levels, when they write, have to say which software they used. So this would be related to the writing required in the journal. Development Environment will not be a piece of important information, people are not interested in which program you used to develop the software. Whereas for students it would*." Since students are required to write their programs in detail, *development environment* will occur frequently. The extract from NS4 below confirms CS experts' explanations that students need to write their programs in detail:

The goal of the project was to build a software visualisation tool for task execution times. It is intended to help designers compare features of task execution times such as the average case execution time and standard deviation for tasks running under different platforms. As a design tool, it will work with the eclipse integrated ***development environment*** *so that it integrates with the designer's normal* ***development environment****. Within the 3 months available for the project, the following was the list of objectives for the project… (NS4)*

Thus, it can be concluded that since *development environment* is a GCSC it may not be important to be mentioned by experts as expert genre requirements do not require writers to focus on it, as explained by P2. On the other hand, the overuse of this collocation by students tends to be related to their detailed writing about the software developed for their MSc projects.

Moreover, it is not only the demands of writing in two genres that matter but also the style of writing in these two genres. MSc dissertations are affected by the style of writing common in the CS industry while the published research articles are more academic in style, as commented by P3, who had published numerous book chapters and research articles in IS: "*I think because dissertations are focused more on industrial professional style rather than academic writing, both NNS and NS are not writing in an academic way. Their formal reports may include some academic writing, but it cannot be compared to research articles*". Students tend to prefer

writing in industrial style as most of them are supposed to work in companies and industries rather than in academic sections. The apparent overuse of this collocation by the NS students can be related to the topics selected to compile this corpus. It could probably be summarised from Table 5-6 that the overuse of this collocation by NS compared to NNS is due to their selected number of topics that may include the construction of programs and, thus, more cases of *development environment* will occur.

When CS experts were asked about the possibility of this collocation's occurrence in selected topics from each corpus, they were able to confirm that topic affects the use of *development environment*. If topics selected were about implementation of programs, a program will be written and *development environment* will be used. For example, some of the given topics to the CS experts were about programming, thus, *development environment* occurred. For example, one of the given topics was *Intelligent Web Search Using Named Entity Recognition (NNS23);* P1 confirmed that it involves using *development environment*: "*yes, this is definitely talking about construction of a program, so it will have occurrences*".

P2 also confirmed "*It involves programming as it specified the name of program used for their intelligent web search which is Named Entity Recognition*". All CS experts agreed that these topics involve the use of *development environment*, as they might include a type of programming.

Web site

The collocation *web site* has also one pattern. This pattern was found in both the RC and NNS corpus only. This pattern was non-significantly overused by NNS (Fisher's chi-square was 0.073, *p>0.05*).

Table 5-7: RF, NF, and number of users for the patterns of *web site* in each corpus.

Corpus	**Web site**		
	Raw frequency	NF	No. of users
Reference	37	6.16	6/63
NNS	30	9.9	10/29
NS			

Web site is a fixed term that is defined as "a document or a set of linked documents, usually associated with a particular person, organization, or topic, that is held on…a computer system and can be accessed as part of the World Wide Web"(*online Oxford English Dictionary*).This could explain the single pattern of this collocation.

Another possible explanation of the overuse of *web site* by NNS students could be the topics selected for writing their dissertations. A great number of NNS dissertations included topics related to *web site,* such as ***Web Application Security (NNS28), Advanced Web Application Programming (NNS14)***. Both topics include the development of *web site* in their titles and thus will use *web site* frequently. On the other hand, NS dissertations did not include topics about *web site.* NS seem to prefer other topics that are not related to *web site.*

This finding has been confirmed by CS experts when they provided their explanations about the effect of the topics in the use of web site. P1 said, "*The variation is entirely explained by the topics*". NNS students tend to prefer to select topics related to *web site*, while NS seem to prefer topics unrelated to it, as P2 commented:

"*Web site is a really general term. My guess will be the choice of topics. It might be worth looking at the curriculum we have for the last five years; we have massive projects based on websites. But the level of skills required were not high enough for the degree of the MSc. Most of the NS students from UK had already done that in their undergraduate level or before they come to the university. We discourage NNS students to do their projects on websites.They tend to use web site earlier. I would think the NNS might like to choose projects related to web site; it might be more useful to know where they come from, while here they assumed they already have it*".

Therefore, the variation could be likely related to the NNS vs. NS preferences in their choice of topics.

Since the writing of dissertations is longer than the writing in research articles, this would involve a repetition of the collocation. It seems that if the topic was about *web site*, thus, more repetition will occur throughout the dissertation. As P3 explained, "*students are most likely talking about web site in their projects as they are developing maybe some websites*". Another tentative explanation could be related perhaps to the idiosyncratic style of the writer. That is, each writer has his personal style in writing. P2 suggested that personal style could play a role in using *web site* in the students' writing: *"In personal style; I think if we exclude the choice of projects, so let's leave that aside, then I could think that personal style can play a secondary role in that people use web site in more generic projects when they are trying to think of general terms.*"

Open Source

As can be seen from Table 5-8, only one pattern was located for the collocation *open source* in the reference corpus and NS corpus only. Even though the located pattern was overused by NS students compared to the RC, it was non-significantly so (p>0.05).That is, it did not support the overuse pattern hypothesis as it only has one pattern and thus the lack of a significant overuse in the NS corpus could be explained.

The overuse of *open source* by NS students compared to the experts' writers tends to be related to the genre requirements. Students write in more detail in their writing of dissertations, while experts' writing is constrained by the rules of the journal in which they aim to publish. Unlike NS students, NNS students did not overuse *open source*, eventhough *open source* had been categorised as GCSC, which means that it can occur in any discipline. Perhaps NNS students' selected topics do not require the use of *open source*.

Table5-8: RF, NF, and number of users for the patterns of *open source* in each corpus.

Corpus	**Open source**		
	Raw frequency	NF	No. of users
Reference	33	5.4	7/63
NNS	0		
NS	44	14.9	11/26

Previous Section and Following Section

Unlike previous single pattern collocations that were all GCSC, these two collocations were marked as GAC by all CS experts in the categorisation judgement task. Their categorisation was congruent with dictionary information. It seems that these two collocations can be used in any academic discipline, as they were not found in the CS dictionary. This finding is in line with Ackermann and Chen's (2013) ACL, in which *previous/following section* were included as general academic collocations.

Table5-9: RF, NF, and number of users for the patterns of *following/previous section* in each corpus.

Corpus	**Following section**			**Previous section**		
	Raw frequency	NF	No. of users	Raw frequency	NF	No.of users
Reference	14	2.33	7/63	13	5.4	6/63
NNS						
NS	20	6.8	8/26	27	9.1	10/26

The striking result was that these two collocations were used only in the NS students' corpus and the experts' corpus, but their use was non-significant in the NNS corpus, as is shown in Table 5-9. This could be explained by NNS students' lack of signposting in their academic writing, as they may not be taught about the function and importance of such expressions in guiding the reader through the text. One of the CS experts commented, "*It is striking. These phrases are learnt, they are vocabulary items, which might be learnt by NS more than NNS as they are advanced in their use of language. Thus, they are used freely by NS. NNS writers are very different…It would be interesting to find what NNS writers used instead of these phrases.*"

Another explanation was that NNS are different from NS in their use of signposts in their academic writing. As P2 commented, "*they might be related to the style of the writing. Most people like to give signs to the readers but foreign students may not use these signs in their writing*."

The apparent overuse of these two collocations by NS students can be explained by different genre writing demands between dissertations and articles. Writing in research articles is condensed, unlike the long detailed writing of dissertations. Students are required to write more sections in each chapter, thus, signposts will occur more frequently. As P1and P2 explained:

P1: "*When you write a dissertation, it will be mainly read by your supervisor or assessor. So it is as if you are making a conversation with them and you want to guide them. While writing in articles is like writing to a crowd. So there is only big headings and you expect the reader to find their way.*"

P2:"*People writing in articles are more likely to mention the name of the chapter or section instead of using the term. They do not have room for that. I agree with you in that the number of chapters and sections will be fewer in journals. But in dissertations, students write more sections and chapters, it is more preferable.*"

The style of expert writers is more professional than NS students' writing. Expert writers are more experienced than students are and thus likely to have a more accomplished style of academic writing. Thus, their style can be characterised as "*elegant and compressed*" to avoid repeating words, as commented by P3: "*Even though NS overused these two collocations, there was still variation among them*". It was clear that only few NS use this signpost language in their writing more than others, as indicated by the number of users in Table 5-9. Thus, the overuse of these collocations might be explained by the fact that different writers write in different ways. As commented by P1: "*Yes, the last factor [personal style] can explain the different uses of these*

phrases between NNS and NS. I can imagine that there are also differences among NS use of these phrases."

To summarise, single pattern collocations were either CS terms, like *web site* and *development environment*, GAC, like *previous section* and *following section*, or GCSC, like *open source*. These findings were in line with the CS experts' choices in the CJT. The single pattern collocations did not support the overuse pattern hypothesis since single pattern was located in their use across all corpora. Thus, patterns identification did not explain the over/underuse of these collocations. However, they were explained by other factors: genre, topic, and experts vs. novice writers. The remaining 19 collocations will be discussed applying the same analysis in the following sections.

5.4.3.2 Two Pattern Collocations

A number of collocations have two patterns. These are *data structure, code source, layer application, available resources,* and *other features.* The patterns of each collocation will be presented individually.

Data Structure

Looking at *data structure* patterns, only two patterns were identified. The first pattern, *data structure*, was used by both expert writers and NS students. This pattern was non-significantly overused by NS students (p>0.05) since differences between frequencies across the corpora were non-significant. The overuse of the pattern by NS students could be related to the rules of writing MSc dissertations, which differ from those related to research articles' writing.

Table 5-10: RF, NF, and number of users for the patterns of *data structure* in each corpus

Corpus	**Data structure**			**Structure of data**		
	RF	NF	No. of users	RF	NF	No. of users
Reference	25	4.1	8/63	1	0.16	1/63
NNS						
NS	30	10	5/26			

The second pattern, *structure of data*, occurs only once in the reference corpus. Thus, it is an infrequently used pattern. The chi-square test could not be performed since the pattern appeared only in one corpus.

Code Source

The collocation *code source* has two patterns identifiable from the corpus, as displayed in Table5-11:

1-Adjacent collocation: *source code*;

2-'Source + NP that includes 'code' in prepositional phrase that is dependent on the noun'.

These two patterns are presented in the extracts below:

1…knowledge_NNW of_PRF the_ATK ***source_NNW code_NNW*** and_CJC _UNC or_CJC better_AJC user_NNW (RC1)

2…with_PRP ***source_NNW lines_NNY of_PRF code_NNW***, _PUN alert_AJK density_NNW from_PRP a_ATK (RC69)

Table 5-11: RF, NF, and number of users for the patterns of *source code* in each corpus.

Corpus	**Source code**			**Source lines of code**		
	RF	NF	No. of users	RF	NF	No.of users
RC	66	11	7/63	2	0.33	1/63
NNS	11	3.6	4/29			
NS	126	42.8	9/26			

The first pattern, *source code*, was located in all corpora and thus used by students and expert writers. However, it was overused by NS students and underused by NNS students. The overuse of this pattern by NS students was significant (Fisher's chi-square=18.750, p 0<0.05) and its underuse by NNS was significant (Fisher's chi-square= 39.286, p 0<0.05).

From my preliminary analysis of the CJT, *source code* was categorised as GCSC by three CS experts, thus it means that *source code* can be used in any discipline and thus may be used in any topics. To confirm that, I checked topics of dissertations in which *source code* was used. I could not find any clue from the titles whether *source code* was used in them. These titles were:

NNS17: Intelligent Control of an Unmanned Aerial Vehicle

RC (15 se): A Framework for the Checking and Refactoring of Crosscutting Concepts

(16 se): A Logical Verification Methodology for Service-oriented Computing

My finding was confirmed by two CS experts when they looked at the topics, P1 said, "*It cannot be explained by topics. Like the phrase front door, source code can occur in any branch of Computer Science.*" Thus, it is not topic specific. As P2 said, "*Actually, in all sub-areas of CS,*

there is programming somewhere, so source code will be there. I would not think it is related to specific topics".

However, one of the CS experts who specialised in SE suggested that *source code* could be used more in SE than in the other two CS sub-disciplines: "*People in SE definitely talk about source code more*". His suggestion was confirmed when he checked the topics given as they were all from SE. "*From research articles all from SE, I am not surprised to see this. As I said before, it is mostly used in SE in programming*".

As mentioned by all CS experts, *source code* is regularly used in programming, and since in most the MSc dissertations students talk about programming and produce lines of programming code, the overuse of this collocation is expected in NS students' dissertations. In contrast, experts do not talk about programming in detail when writing research articles, as commented by P2:"*But you can see in journal articles, they will say that they will not mention the codes; they only mentioned that they use source codes for their programming*". Thus, variation of this collocation's use can be explained.

Another factor that may explain the overuse of *source code* by NS rather than by NNS was the cultural factor. P2 confirmed that *source code* is more British-oriented than US-oriented and, thus, might be used more by NS students: "*Something you may not be aware of is that this term is more British-oriented. In the US, they tend to use different terms, they say this is the program, lines of program, but in the UK, we used source code and lines of code instead. So there is a difference between the countries as well. When I came to this country, I realised the difference of*

how many times they use source code and lines of coding compared to the US". Thus, NNS students may use the US style in their writing of their programs.

However, P3 suggested that differences in the frequency of this expression could be related to students' preferences: "*NS tend to pick up projects that involve programming and NNS may like to choose different topics which might not use programming*". The second pattern *source lines of* code occurred only in the reference corpus. Since it occurred only twice, it is considered an infrequent pattern.

Layer Application

Two patterns were identified for *layer application*: *application layer* and *application +N+ layer*, as shown in Table5-12.

Table 5-12: RF, NF, and number of users for the patterns of *layer application* in each corpus.

Corpus	**Application layer**			**Application +N+ layer**		
	RF	NF	No. of users	RF	NF	No. of users
Reference	8	1.33	5/63	1	0.16	1/63
NNS	13	4.3	7/29			
NS	38	12.9	6/26			

The first pattern was identified in all corpora. Even though *application layer* was overused by both NNS and NS students, it was only significantly so for the NS students (Fisher's chi-square =19.565, $p < 0.05$).The overuse of *layer application* in the NS corpus was due to the significantly more frequent occurrence of the pattern *application layer* in the NS corpus than in the RC.

It is most likely that the apparent overuse of this pattern could be related to the fact that writing in dissertations is quite different from writing in research articles. Students are required to write in more detail and to include more chapters in their dissertations, whereas expert writers are restricted to a certain number of words according to the rules set for the journal. The extract below shows the use of application *layer* in one of the student's dissertations:

A simplified model of the architecture deployed by both platform specific application and web application has been shown in Figure. This model…both the ***application layer*** *and web* ***application layer****. […]The high-level extensions implemented within this project are targeted at the web* ***application layer*** *and thus utilise the Browser Interface which is provided by the web browser. However, the Operating System…general* ***application layer*** *will also be briefly discussed since both the iOS (NS27)*

The second pattern occurred only once in the RC. This pattern could perhaps be merged with the first pattern if the added noun can be removed. When the concordance line from the RC was checked, the added N was a name of a specific application layer.

cluster_NNW mode_NNW serve_VVB as_PRP the_ATK ***application_NNW server_NNW layer_NNW*** (8RC)

To confirm whether it was a specific name of layer application, I examined the context of this line:

Eight_CRD instances_NNY of_PRF WebSphere_NPK v7_UNC with_PRP JDK_NPK 1.6_CRD in_PRP cluster_NNW mode_NNW serve_VVB as_PRP the_ATK ***application_NNW server_NNW layer_NNW***

The context suggests that *application server layer* is the name of a certain application layer.

Resources Available

As can be observed from Table 5-13, the collocation *resources available* has two identified patterns: *resources available* and *available resources*. The two patterns were located in all corpora, but their frequency of use was different.

Table 5-13: RF, NF, and number of users for the patterns of *available resources* in each corpus.

Corpus	**Resources available**			**Available resources**		
	RF	NF	No.of users	RF	NF	No.of users
Reference	3	0.4	2/63	1	0.16	1/63
NNS	3	0.99	3/29	10	3.3	6/29
NS	8	2.7	4/26	4	1.3	1/26

Both expert writers and NS students use the pattern *resources available* slightly more than the *available resources* pattern. The NNS students prefer *available resources* to the first pattern *resources available*. The frequencies of these two patterns did not differ significantly between each student corpus and the RC with the exception of the overuse of the pattern *available resources* by NNS students.

Other Features

The collocation *other features* has two identifiable patterns: *other features* and *other+ADJ+features*, as shown in Table5-14.

Table 5-14: RF, NF, and number of users for the patterns of *other features* in each corpus.

Corpus	**Other features**			**Other +ADJ+ features**		
	RF frequency	NF	No.of users	RF	NF	No. of users
Reference	3	0.5	2/63	3	0.5	1/63
NNS	5	1.6	3/29	2	0.66	1/29
NS	10	3.3	5/26			

An example is given for each pattern in the extracts below:

1…there_EXK are_VBB still_AVK ***other_AJK features_NNY*** that_CJT arise_VVB in_PRP course_NNW (RC4)

2…in_PRP ***other_AJK key_AJK product_NNW features_NNY*** such_PRP as_PRP capacity_NNW ,_PUN speed_NNW (RC5)

The first pattern, *other features*, was used by NNS, NS students, and expert writers. This pattern was significantly overused by NS students only (Fisher's chi-square=3.769, $p<0.05$).The second pattern was located in the NNS corpus and in the expert writers' corpus only. The overuse of this pattern by NNS was non-significant (Fisher's chi-square = 0.200, $p > 0.05$).

5.4.3.3 Three-plus Pattern Collocations

The remaining 14 collocations have more than two identified patterns. The patterns of these collocations will be discussed in detail below.

Different Components

Three patterns were identified for the overused collocation *components different*: *different components, 'different+N+components'*, and *'different+ADJ+components',* as is shown in Table 5-15.

Table 5-15: RF, NF, and number of users for the patterns of the *different components* in each corpus.

Corpus	**Different components**			**Different N components**			**Different ADJ components**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
RC	5	0.8	3/63	4	0.66	2/63			
NNS									
NS	12	4.07	5/26	6	2.03	2/26	2	0.67	1/26

The first two patterns, which were identified in the NS student corpus and expert writers' corpus, were both non-significantly overused by NS students. However, the pattern *'different +ADJ+ components'* was only found in the NS corpus. It seems that this pattern was infrequently used by NS students as it occurred in just one dissertation. Thus, it could be related to the user's idiosyncratic style. The overuse of *different components* has supported the overuse pattern hypothesis since NS students used more patterns than experts did. Since *different components*

was categorised as GAC by all CS experts, this collocation would be expected to occur in any academic discipline; therefore, its occurrence would not be explained by the selected topics of MSc dissertations and research articles.

Code Following

Three patterns were identified for the collocation *code following*:

1. Adjacent: *Following code*.
2. Non-adjacent: *following + ADJ + code*.
3. *Following* + NP, which includes *code* in the prepositional phrase that is dependent on the noun.

Table 5-16: RF, NF, and number of users for the patterns of the *code following* in each corpus.

Corpus	**The following code**			**The following +ADJ+code**			**The following section of code**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
RC	3	0.5	1/63						
NNS	8	2.7	2/29						
NS	26	8.8	5/26	6	2.03	2/26	4	1.3	1/26

The first pattern following *code* was shared among all corpora, but was significantly overused by NS students only (Fisher's chi-square=18.241, $p<0.05$).However, the last two patterns occurred only in the NS corpus. The pattern overuse hypothesis was only supported by NS students as they use more patterns for the collocation *following code* than expert writers do.

The apparent overuse could be explained by the fact that different writing demands are required in writing in two different genres. Two CS experts were able to confirm that writing a program required much repetition in the students' writing of dissertations. P1 commented: "*It would be most likely that students are more interested in demonstrating their ability in and their understanding of using the code. They show examples of codes. Thus, they use demonstrative phrases like the following code. Whereas in writing an article, they would not be likely to include examples of code, so they might not be used.*"

P2 gave a more detailed explanation of the overuse of this collocation by students: "*The following code means that students are going to show real lines of programming, lines of code or lines of programming. This phrase is really used when students write about programming … but in journal articles they will not list codes. Most journals will say explicitly, "do not put these codes"; they do not want to see the codes. If you want to see the codes, you can make it available online so people can download it.*" This has been confirmed when context was checked for the following code, number of codes were presented as shown in the extract below:

This address takes two parameters, which holds the start address and the destination address. The parameter is set to be the current latitude / longitude location of the user; this is obtained by calling the method of the Location Data class. ***The following code*** *demonstrates how the request is constructed and the parameter is set to equal the location of the user [codes appeared]…the call to the Directions API is a valid one.* ***The following code*** *demonstrates how the parameter is set to the Encoded destination String: [codes appeared].The service call, when completed looks something like: [codes appeared].This is then used within an object to obtain from the Google Directions API.* ***The following code*** *demonstrates this: [codes appeared]. (NS12).*

Since reporting on the development of a program required demonstrating the codes used in detail, dissertations are expected to mention many codes. Students are asked to retain them on

CDs, as commented by P2:"*In both projects nowadays, since there are a lot of codes used… for example, a project of three or four months can easily have hundreds of pages of code. Obviously, looking at all these codes will be difficult. Even now, we ask our students not to add them in the project; they are better keeping them in CDs. If we need them, we will look at them. So we expect students to use less and less codes in their writing.*"

Another reason for the underuse of *following code* by experts suggested by P3 is that expert writers tend to use "*pseudo codes in their writing of programming rather than using 'the following code'*". They also seem to prefer to display their codes in figures rather than in writing, as P3 commented: "*In articles, codes are displayed in figures, so you do not have to mention 'the following code'; it is shown in the figure. But in students' writing, they like to display codes in lines, not in figures, even though we encourage them not to do so*". It seems that students' limited academic writing experience may prevent them from using the appropriate style. Both P2 and P3 claim that they asked students not to overuse codes in their writing and to follow expert writers' style:

P3:"*Students may not think of writing in a professional way; they just write in a linear style. They are not writing to publish their work.*"

P2:"*Another issue is that some students are repetitive in their writing; they just keep mentioning the same word distributed everywhere in their dissertation. They do not have that sense of narrative flow in academic writing.*"

When CS experts were asked whether the occurrence of this collocation could be topic-specific, two of them replied, "*the topic will not make a great deal of difference*" (P1 and P3). Only P2 thought that following code could be topic-specific, as he commented on the topics given as follows, "*It will definitely make a difference, but I am surprised how codes will be used in these specific topics. For the second and fourth topics (Optimising for High-Performance Cache*

Utilisation& Advanced Web Application Programming), it will definitely occur because of the level of programming used there. It is mostly used in hardware. We can expect some codes… about Cache Utilisation…but for No.3 and No.1 (Optical Information System &Minimum Spanning Tree with Uncertainty), I think no codes will be used". Therefore, it seems that *following code* may not be connected to specific topics, but it will very likely be used and used much more heavily in projects that involve programming

The last two patterns were identified only in the NS students' corpus. Their infrequent use might be explained if I examine these patterns in their concordance lines; this examination may help us decide if the use of these patterns is user- or topic-specific. The pattern *following +ADJ+code* was found in two files only (dissertations No. 12 and No. 16) and is just a variation of *following code*. In other words, I do not think that the use of the adjective to modify *code* is connected with writing style. It is just that these two writers happened to want to characterise the code they were presenting to the reader and had to use adjectives to do so.

1… the ***following_AJK, _PUN abbreviated_AJK, _PUN code_NNW*** demonstrates_VVZ this_DTK: _PUN Added_VVN as_PRP (NS12)

2… by_PRP the_ATK ***following_AJK pseudo_AJK code_NNW*** demonstrates_VVZ how_AVQ to_TO0 calculate_VVI (NS16)

On the other hand, when I examined the concordance lines for the third pattern *following section of code* all occurrences were extracted from the same file (dissertation No. 28). Thus, this pattern appeared to be user-specific.

…The_ATK ***following_AJK section_NNW of_PRF code_NNW*** checks_VVZ that_CJT a_ATK graph_NNW is_VBZ (NS28).

Resources System

Three patterns were identified for the collocation *resources system*, two of which occurred in more than one corpus, as shown in Table5-17. Both NNS and NS students use the pattern *system resources* slightly more than the '*resources+PRP+system*' pattern. In addition, expert writers tend to prefer *using* '*resources +PRP+ system*' as a pattern. However, in both cases, the numbers are far too small to reach any firm conclusions about over- and underuse. There were no-significant difference between the frequencies across corpora in their use of these patterns.

Table 5-17: RF, NF, and number of users for patterns of the *resources system* in each corpus.

Corpus	**Resources +PRP+ system**			**System +PRP+ resources**			**System resources**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	3	0.5	2/63	1	0.16	1/63			
NNS	1	0.33	1/29				4	1.3	3/29
NS	2	0.66	1/26				8	2.7	3/26

Data Time

Five patterns were identified for the collocation *data time*, as can be seen from Table 5-18. There were no shared patterns across the corpora, except *time+N+data* that was identified in NNS students' corpus and expert writers' corpus. Due to the small number of occurrences, this pattern was non-significant since Fisher's chi-square test was not valid because of small expected frequencies (two of the expected count cells were below five). Since NNS used more patterns than expert writers did, the overuse hypothesis had been supported.

Table 5-18: RF, NF, and number of users for the patterns of *data time* in each corpus.

Corpus	**Time+PRP+data**			**Time+N+data**			**Data per time unit/ data for each time point**			**ADJ time+ADJ data**			**Execution time data**		
	RF	NF	No. of users	RF	NF	No.of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	3	0.5	2/63	1	0.16	1/63									
NNS				1	0.33	1/29	2	0.66	1/29	1	0.33	1/29			
NS													20	5.8	1/26

On the other hand, the overuse of *data time* by NS students tends to be related to the use of the pattern *time data* preceded by *execution*. It seems that *execution time data* is a fixed term in CS, as was confirmed by dictionary consultation. Moreover, this pattern was used in a single dissertation. When the topic of this dissertation (NS4) was checked, a clear indication of execution time was confirmed (*System Timing Visualiser: A Software Tool to Visualise Task Execution time for a System under Timing Constraints*).Thus, it was related to a specific topic.

Data User

It can be noted from Table5-19 that *data user* was overused by both NNS and NS students. Since NNS students used more patterns than experts did in the RC, *data user* overuse in the NNS corpus can be explained by the pattern overuse hypothesis. There were only two patterns shared between the three corpora: *user data* and *data+PRP+user*. Even though the pattern *user data* was overused by both NNS and NS students, it was significantly overused by NS students only

(Fisher's chi-square=5.333, $p < 0.05$). The pattern *data+PRP+user* was non-significantly overused by both NS and NNS students.

Table 5-19: RF, NF, and number of users for the patterns of the *data user* in each corpus.

Corpus	**Data user**			**User data**			**Data +PRP+user**			**User +N+ data**			**Data collection of user information**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
RC				2	0.33	2/63	1	0.16	1/63	1	0.16	1/63			
NNS	2	0.66	1/29	4	1.3	2/29	2	0.66	2/29	2	0.66	1/29	1	0.33	1/29
NS				10	3.3	4/26	4	1.3	2/26						

The two other patterns (*data user* and *data collection of user information*) were only identified in the NNS corpus with a small number of occurrences. When their concordance lines were examined, these patterns occurred in two files only (NNS14 and NNS1).

1…radio_NNW channel_NNW to_PRP a_ATK mobile_AJK ***data***_NNK ***user***_NNW ,_PUN works_VVZ by_PRP dedicating_VVG (NNS14)

2…The_ATK first_ORD stage_NNW is_VBZ the_ATK ***data***_NNK collection_NNW of_PRF ***user***_NNW information_NNW (NNS1)

Therefore, they can be considered infrequent patterns used by single writers (as shown by the number of users in Table 5-19) who prefer to use various extended patterns rather than the usual pattern of the collocation. Since these patterns were used by some NNS students only, they could be related to the NNS style of using long noun phrases instead of collocations established in the language.

A CS expert claimed that NNS students tend to use long phrases instead of using a brief collocation with the same meaning. P2 commented: "*I realised that NNS students tend to use long phrases with a lot of chopped part, whereas experienced NS use short phrases. This is just a general comment, not applicable only to this term.* "It can be clearly observed in the pattern *data collection of user information*, which was only used by one of the NNS students who might not be aware of the fixed use of the collocation and prefers to use long extended phrases. The shared pattern between NNS corpus and RC *user + N + data* was infrequently used as it occurred in single files, as shown by the number of users in Table 5-19.

Therefore, the overuse of the collocation *data user* by NS students was due to their use of two similar patterns (*user data* and *data+PRP+user*) used by expert writers. However, the overuse of *data user* by the NNS students was related to their use of different patterns, which were non-native like.

Data Information

Four patterns were identified for the collocation *data information*, as shown inTable5-20.Only one pattern was found in all corpora: *information +PRP+N +data*. It was non-significantly overused in both the NNS and NS students' corpora. The other three patterns were identified in either NNS or NS corpora. The second pattern *information as metadata* occurred only in the NS corpus; it was accepted as a pattern since it was identified as N pattern (N as N) by Hunston and Francis (1996) (see section 5.2.2.2 for more details).

the_ATK blob_NNW ***information***_NNW as_CJS meta_NNW ***data***_NNK was_VBD found_VVN to_TO0 be_VBI inconvenient_AJK

Table 5-20: RF, NF, and number of users for the patterns of *data information* in each corpus.

Corpus	**Information +PRP+N+data**			**Information as +N+data**			**Data (or/and) information**			**Information content of the data**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	5	8.3	1/63									
NNS	3	9.9	3/29				7	2.3	2/29	1	3.3	1/29
NS	4	1.3	4/26	2	6.7	1/26						

The last two patterns (*data* and/or *information* and *information content of the data*) were located in the NNS corpus only. The small number of occurrences indicates their infrequency.

1…of_PRF guaranteeing_VVG that_CJT ***data***_NNK or_CJC ***information***_NNW may_VMK only_AVK be_VBI (NNS2)

2…in_AVK order_AVK to_TO0 swap_VVI ***data***_NNK and_CJC ***information***_NNW using_VVG electronic_AJK (NNS6)

3…***information***_NNW content_NNW of_PRF the_ATK ***data***_NNK stream_NNW. (NNS2)

As a result, the overuse of *data information* by NNS and NS students could be explained by the overused pattern hypothesis. Students tend to use more patterns in their use of *data information* than expert writers do.

Data Amount

Examining patterns for *data amount*, five patterns were identified, as shown in Table5-21. Two patterns were identified in all corpora: *amount+PPR+data* and *amount+PRP+ADJ+data*. The pattern *amount+PRP+data* was overused by both NNS and NS students, due to the writing

demands in two different genres. Students writing is more detailed and repetitive, thus, more occurrences of the collocation might occur, as shown in the extract below:

*The design should minimise the **amount of data** transfers between the cloud and the premise. Large **amount of data** transferring can affect in slower performance. Consider computation tradeoffs between cloud and premise (NS20).*

Table 5-21: RF, NF, and number of users for the patterns of the *data amount* in each corpus.

Corpus	**Amount+PRP+data**			**Amount+PRP+ADJ+data**			**Amount+PRP+N+data**			**Amount+PRP+ADJ+ADJ+data**			**Amount+PRP+ADJ+N+data**		
	RF	NF	No.of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No.of users
Reference	3	0.5	2/63	2	0.33	2/63	4	0.66	2/63	1	0.16	1/63	2	0.33	1/63
NNS	20	6.6	6/29	2	0.66	1/29	3	1.9	2/29						
NS	20	5.8	8/26	4	1.3	2/26									

This pattern was significantly overused (Fisher's chi-square=12.565 for both NNS and NS corpora, $p < 0.05$). *Data amount* was marked as GAC by all CS experts in the CJT, thus, it is considered not topic-specific. In contrast, the pattern *amount +PRP+ADJ+ data* was non-significantly overused by NS and NNS students. Thus, no explanation is provided. The third pattern *amount+PRP+N+data* was identified only in the RC and NNS students' writing and it was no significantly overused.

The last two patterns were only used by expert writers. Their small number of occurrences could be related to either specific writer style or specific topic. To test these possibilities, concordance lines of these patterns were examined.

1…do_VDB not_XXK have_VHI a_ATK sufficient_AJK ***amount_NNW of_PRF labeled_AJK training_NNW data_NNK (RC18 AI)***

2…a_ATK dramatic_AJK increase_NNW in_PRP the_ATK ***amount_NNW of_PRF volumetric_AJK image_NNW data_NNK (RC13 IS)***

3…click_NNW data_NNK and_CJC the_ATK small_AJK ***amount_NNW of_PRF multi-grade_AJK labeled_AJK data_NNK (RC13 IS)***

When the two occurrences of the pattern *amount +PRP+ADJ+N+data* were checked, a clear indication of their specific use by single writers was confirmed. Only two expert writers used this pattern (the writer of article No. 18 from AI and the writer of article No.13 from IS). Moreover, the pattern *amount+PRP+ADJ+ADJ+data* was used by a single expert writer who also used the previous pattern (writer for article No.13 from IS). Therefore, these two patterns (*amount+PRP+ADJ+N+data* and *amount+PRP+ADJ+ADJ+data*) can be considered user-specific. Consequently, the overuse of this collocation by students could not be explained by the pattern overuse hypothesis. In contrast, experts tend to use more patterns than students do.

Data Type

As can be seen from Table 5-22, three patterns were identified for *data type*. Two patterns were shared between the corpora: *data type* and *type+of+data*. Both experts and NS students prefer to use *data type* instead of *type+of+data*, whereas NNS students tend to prefer *type+of+data* rather than *data type.*

Table 5-22: RF, NF, and number of users for the patterns of *data type* in each corpus

Corpus	**Data type**			**Type of data**			**Type+of+ADJ+data**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	4	0.66	3/63	3	0.5	2/63			
NNS	2	0.66	2/29	5	1.6	4/29	2	0.66	2/29
NS	22	7.4	10/26	4	1.3	2/26			

The overuse of *data type* by NS students, which was significant (Fisher's chi-square=14.440, $p<0.05$), could be explained by genre variation. One of the CS experts (P3) thought that the variation of the overuse of *data type* could be explained by genre constraints: "*students are going to focus on programming; they surely will use this phrase*". Other CS experts felt that genre could not explain this variation as this collocation is used as a general collocation rather than a specific one as reported by P2: "*data type is a very general term in CS*".

Another tentative explanation could be the effect of topics selected in compiling the NS corpus. Even though there was disagreement among CS experts about the effect of topics on the use of *data type*, it was confirmed that *data type* is used more as a general term in programming, as P2 said when asked about topics: "*I do not think that topics may explain. Data type is a very general term in CS*."On the other hand, another CS expert (P3) thought that topic might play a role in this variation: "*I suspect a lot of narrative description of their programming is in these dissertations. Most dissertations in CS develop programs or some practical experiments*."

The pattern *type of data* was overused by NNS more than by NS students. The slight overuse of this pattern by NNS students can be related to their preference for using long phrases rather than fixed collocations. It was confirmed by one of the CS experts (P3): "*NNS like to extend their*

writing by adding prepositional phrases and using relative clauses". He added, "*NS may be more aware of the fixed technical term and more sensitive to this phrase, but NNS seem not to be aware about the use of this phrase*".

The last pattern *type+of+ADJ+data* was used by NNS students only. Its few occurrences can be related to specific-user style (as shown by the number of users). Examining the concordance lines of this pattern confirms its use by a single user as it was located in a single file (NNS 14), as shown in the extracts:

1…another_***DTK type_NNW of_PRF unwanted_AJK data_NNK*** which_DTQ need_VVB to_TO0 be_VBI removed_VVN (NNS14).

2…are_VBB a_***ATK type_NNW of_PRF unwanted_AJK data_NNK*** available_AJK on_PRP web_NNW pages_NNY (NNS14).

To summarise, the pattern overuse hypothesis was only supported in relation to NNS students since they used more patterns than the expert authors did.

Data Layer

Three patterns were identified for *data layer*, but only one pattern *data layer* was shared between the corpora. It was significantly overused by NS students only (Fisher's chi-square=15.696, $p<0.05$). Even though NNS and NS students' overuse of *data layer* supported the pattern overuse hypothesis they used different patterns. NS students prefer to use *data layer* while NNS students prefer to use *data+N+layer*.

Table 5-23: RF, NF, and number of users for the patterns of *data layer* in each corpus.

Corpus	**Data layer**			**Data+N+ layer**			**Data+PRP+layer**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	21	3.5	15/63						
NNS	2	0.66	1/29	24	7.9	12/29			
NS	24	8.1	9/26				2	0.67	1/26

When some of the concordance lines of *data+N+layer* were examined, it was obvious that the added nouns (*link* and *access*) were all proper names as they were all capitalised as shown in the extracts below:

1…are_VBB totally_AVK based_VVN on_PRP ***Data_NNK Link_NNW Layer***_NNW technology_NNW (NNS12).

2…layer_NNW acts_VVZ as_PRP a_ATK system_NNW ***Data_NNK Link_NNW Layer***_NNW that_CJT can_VMK be_VBI (NNS3).

3…TOOLONG_NNW class_NNW in_PRP the_ATK ***Data_NNK Access_NNW Layer_NNW***._ (NNS6).

4…Diagram_NNW for_PRP the_ATK Linq_AJK Meeting_NNW ***Data_NNK Acess_NNW Layer_NNW*** is_VBZ shown_VVN … (NNS15).

Perhaps *data+N+layer* is used in specific topics or specific sub-disciplines of CS. *Data layer* was marked as SCSC by CS experts in CJT. It was categorised as a collocation specific to IS. The NNS corpus consists of more dissertations written in the IS sub-discipline than the NS corpus (see Table 3-4). This fact could explain NNS students' use of different patterns as they may include different topics talking about *data layer*. The pattern *data+PRP+layer* was only

used infrequently by a single NS user, thus it might be related to the idiosyncratic style of the writer.

Network Traffic

As can be seen from Table 5-24, five patterns were identified for *network traffic*, two of which were located in all corpora. These patterns were *network traffic* and *traffic+PRP+the network*.

Table 5-24: RF, NF, and number of users for the patterns of *network traffic* in each corpus.

Corpus	**Network traffic**			**Traffic +PRP+ the network**			**Traffic +N+PRP+ network**			**Traffic network**			**Traffic +PRP+ADJ+ network**		
	RF	NF	No.of users	RF	NF	No.of users	RF	NF	No.of users	RF	NF	No.of users	RF	NF	No.of users
RC	5	0.8	3/63	3	0.5	2/63									
NNS	43	14.3	13/29	8	2.7	5/29	3	0.99	1/29	1	0.33	1/29			
NS	4	1.3	3/26	2	0.66	1/26	2	0.66	1/26				1	0.33	1/26

The pattern *network traffic* was used more across all corpora compared to the use of the second pattern *traffic+PRP+the network*. The first pattern was significantly overused by NNS students only (Fisher's chi-square=30.08, p 0< 0.5).

NNS students' overuse tends to be related to their choices of topics included in their corpus. CS experts marked *network traffic* as a collocation specific to IS in the CJT and the NNS corpus included sixteen dissertations written in the IS sub-discipline compared to six NS dissertations

written in the same sub-discipline. Thus, the overuse of *network traffic* could be related to the topics written in IS. As a result, *network traffic* would be a discipline specific collocation.

Similarly, the pattern *traffic+PRP+the network* was overused more by NNS than NS students, as shown from the extracts:

1…video_NNW **traffic_NNW in_PRP the_ATK network**_NNW by_PRP assigning_VVG a_ATK video_NNW (NNS1)

2…the_ATK incoming_AJK **traffic_NNW at_PRP the_ATK network**_NNW router_NNW interface_NNW by_PRP using_VVG (NNS3)

3…monitor_VVI all_DTK **traffic_NNW on_PRP the_ATK network**_NNW links_NNY._They are_VBB able_AJK (NNS6)

4…the_ATK internal_AJK traffic**_NNW of_PRF the_ATK network**_NNW so_CJS that_CJS they_PNP can_VMK be_VBI (NNS4)

5…encrypted_VVD **traffic_NNW across_PRP the_ATK network_**NNW._Encryption_NNW protocols_NNY (NNS10)

On the other hand, only two occurrences of this pattern were found in the NS corpus and they occurred in the same file (NS23).

1…**traffic_NNW ingress_NNW to_PRP the_ATKnetwork_NNW**._ Whilst_CJS Ethernet_NPK has_VHZ

2…data_NNK **traffic_NNW in_PRP the_ATK mobile_AJK network_**NNW. Voice_NNW requires_VVZ low_AJK

A possible interpretation would be that some NNS students might not be aware of the fixed use of the collocation and prefer to use long extended phrases. The number of users of NNS students proves this possibility, as shown in Table5-24.

However, the third pattern *traffic+N+PRP+the network* was used infrequently by both NNS and NS students. Examining the concordance lines of this pattern, it was found only in two files (NNS14, NS23).Thus, it could be related to user specific style.

1…machine_NNW ***Traffic_NNW load_NNW on_PRP the_ATK network_NNW*** The_ATK detection_NNW engine_NNW can_VMK (NNS14)

2…each_DTK ***traffic_NNW type_NNW in_PRP the_ATK network_NNW*** ._This_DTK node_NNW is_VBZ (NNS14)

3…***traffic_NNW engineering_NNW within_PRP the_ATK network_NNW***. There_EXK are_VBB different_AJK (NS23)

The last two patterns were located either in the NNS corpus or in the NS corpus. Their small number of occurrences could be related to specific user style. To examine this possibility, concordance lines were checked for the last two patterns, *traffic+PRP+ADJ+network* and *traffic network.*

1…data_NNK ***traffic_NNW in_PRP the_ATK mobile_AJK network***_NNW requires_VVZ low_AJK… (NS20)

2…Recent_AJK ***traffic_NNW network***_NNW measurements_NNY, on_PRP the_ATK other_AJK …(NNS22)

It was found that they were all used in a single file. Thus, it seems likely related to the idiosyncratic style of the writer As a result, the overuse of *network traffic* by NNS and NS students could be explained by the pattern overuse hypothesis.

Class Method

Observing patterns for the collocation *class method*, three patterns were identified, as displayed in Table 5-25. Having examined the patterns for *class method*, its underuse by NNS and NS students can be explained by the underused pattern hypothesis, as both NNS and NS students use only one pattern compared to expert writers.

Table 5-25: RF, NF, and number of users for the patterns of *class method* in each corpus.

Corpus	**Class method**			**Method+PRP+class**			**Method+PRP+ADJ+class**			
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	
Reference	3	0.5	1/63	9	1.5	3/63	1	0.16	1/63	
NNS				1	0.33	1/29				
NS	4	1.3	2/26							

Another factor that might explain the underuse of *method class* was the choice of topics selected. CS experts were able to confirm the effect of the topics in the occurrence of *method class.* It was suggested by P2 that topics that use the Java programming language would feature this collocation, unlike other collocations such as *source code,* which could be used with any programming language. CS experts were given some topics that include the use of *class method* to comment on; these topics are:

RC 15SE: A Framework for the Checking and Refactoring of Crosscutting Concepts

(16 SE): A Logical Verification Methodology for Service-oriented Computing

19SE (1) DARWIN: An Approach to Debugging Evolving Programs

18 IS (2&3) Information Technology Implementers' Responses to User Resistance: Nature and Effects.

After checking the topics of the dissertations that included *method class*, P2 added: *"looking at these, some of these topics will use Java, thus the term will occur … the method class will be used in Java in a very specific program. In other Computer Science programming language, it is called a function, but in Java specifically, it is called method class*". Thus, *method class* tends to occur only in specific topics that use Java as its main programming language. This finding had been confirmed when the context of *class method* was checked; the extract below shows that *class method* was used with Java.

*For CFJ and our implementation for **Java**, we prefer to accept this limitation – enforcing constant super classes, return types and field types in all alternative implementations of **a class method** or field –and use the renaming workaround for all other cases, instead of complicating the type system (NS27).*

The first pattern *class method* was non-significantly overused by NS students compared to expert writers' use. The second pattern *method+PRP+class* was significantly underused by NNS students compared to expert writers' use (Fisher's chi-square=4.500, $p<0.05$). The last pattern *method+PRP+ADJ+class* was only used by expert writers. It could be related to specific user style as it occurred only in a single research article (No.5 from IS), as shown below:

1… ***method***_NNW in_PRP the_ATK considered_AJK ***class_NNW***._ (5IS)

Data Input

As can be noted from Table 5-26, *data input* has four patterns identified in the RC and NS corpus. Only one pattern, *input data*, which was overused by NS students compared to expert writers, was shared between the two corpora. Its overuse was significant (Fisher's chi-

square=6.533, p <0.05). The overuse of data input could be explained by pattern overuse hypothesis since NS students used more patterns than experts do.

Table 5-26: RF, NF, and number of users for the patterns of *data input* in each corpus.

Corpus	**Input data**			**Input+N+data**			**Data input**			**Input+PRP+ADJ+ data**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	8	1.3	3/63	1	0.16	1/63						
NNS												
NS	22	3.7	5/26				2	0.66	1/26	2	0.66	1/26

The other three patterns were found only in one corpus. The low number of occurrences of these patterns in either expert writers' corpus or NS corpus tend to be presumably related to the personal style of the writer as number of users indicated. The second pattern *input+N+data* was only found in a single file in the RC, as shown in the following extract:

1…no_***ATK input_NNW test_NNW data_NNK*** can_VMK change_VVI the_ATK programs_NNY (15AI)

Similarly, the last two patterns *data input* and *input+PRP+ADJ+data* were only found in a single NS dissertation (No.1), as shown in the extracts below:

1…so_CJS that_CJS invalid_AJK ***data_NNK input***_NNW is_VBZ less_AVK likely_AJK.

2…***input_NNW*** and_CJC storage_NNW of_PRF raw_AJK ***data_NNK*** and_CJC the_ATK other_AJK package_NNW.

When CS experts were asked whether the use of data input could be explained by the choice of topics selected, (CS experts were given four topics on prompt cards to comment on, as shown below):

NS 1: Implementation of Game Agents in Unreal Tournament

NS 3: The development of a negotiation system using software agents to attempt to resolve the irregularities associated with the transfer of Professional Football Players (E-commerce technology)

RC (15AI): Similarity measure for anomaly detection and comparing human behaviours

(17AI): Text summarisation contribution to semantic question answering: New approaches for finding answers on the web

P1 confidently confirmed that all these topics covered the use of *data input*, while P2 and P3 thought that the first two topics might include the use of *data input*, but for the third and fourth topics, they confirmed its occurrence, as P2 commented, *"in the first two, it is likely, but in the last two, definitely. As you can see these articles are in AI*". Thus, *data input* seems to be discipline-specific to some extent, as P2 suggested, *"data input actually, very common in AI projects, but if in other areas of CS, it might be used less*".

Design System

Five patterns were identified for the collocation *design system*, as shown in Table 5-27.Two patterns, *system design* and *design+PRP+the system*, were located in all corpora and were overused by NNS and NS students compared to the expert writers. The pattern *system design* was significantly overused by both NNS (Fisher's chi-square=4.0, p 0<0.05) and NS (Fisher's chi-square=18, p 0. <0.05), whereas the pattern *design+PRP+the system* was only significantly overused by NS students only (Fisher's chi-square=5.3, p<0.05). The first pattern overuse hypothesis was supported by both NNS and NS students since they both overused the shared

patterns. Moreover, the second pattern overuse hypothesis was supported only by NNS students as they used more patterns than experts writers did.

Table 5-27: RF, NF, and number of users for the patterns of *design system* in each corpus.

Corpus	**Design system**			**System design**			**Design+PRP+the system**			**System to the design**			**System+N+ design**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No.of users	RF	NF	No. of users	RF	NF	No.of users
Reference				4	0.66	3/63	2	0.33	2/63	1	0.16	1/63			
NNS	1	0.33	1/29	12	3.9	9/29	2	0.66	1/29				1	0.33	1/29
NS				28	9.5	8/26	13	3.3	4/26						

Another tentative explanation could be that students significantly overused the shared patterns due to the demands of writing in two different genres, as P2 commented: *"Yes, I think that would be the main reason. The level of information needed is different. In journals, you will not find this term often, like in dissertations*." The other three patterns *design system, system to the design,* and *system+N+design* were not found in all corpora. The low frequencies of these patterns indicate their infrequent uses. They could be related to specific user style, as can be seen from the number of users in Table 5-27.

Data Access

Data access is associated with four located patterns, three of which were shared between two corpora, as displayed in Table 5-28. Interestingly, the pattern *data access* was non-significantly overused by both NNS and NS students. Since experts used more patterns than both NNS and NS do, the pattern overuse hypothesis had not been supported.

Table 5-28: RF, NF, and number of users for the patterns of *data access* in each corpus.

Corpus	**Data access**			**Data+N+access**			**Access+PRP+ADJ+data**			**Access+ADJ+data**		
	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users	RF	NF	No. of users
Reference	8	1.3	3/63	1	0.16	1/63	1	0.16	1/63	1	0.16	1/63
NNS	18	5.9	2/29	1	0.33	1/29						
NS	8	2.7	3/26				4	1.30	2/26			

This variation could be explained if I consider the topics that were the focus of each corpus. It could be that the students' corpora consist of dissertations on topics that would be associated with *data access* more than in the RC.

The low number of occurrences of the second and the third patterns in the students and experts' corpora indicate their rarity; thus, it would be likely related to specific users as number of users indicates in Table 5-28. Turning to the last pattern, *access+ADJ+data*, it occurs only in the reference corpus and was used by only one user. Thus, it could be related to the personal style of the writer.

5.4.4 Discussion

The previous detailed findings reveal that some of the collocations were more explainable than others in terms of their patterns. Few of the overused collocations by both NNS and NS students have supported the pattern overuse hypotheses since both NNS and NS students used more patterns than experts do, as shown in Table 5-5. Both NNS and NS students used different patterns in their use of collocations; the most frequent patterns were N+PRP+N and N+ADJ+N: these patterns were all classified by Hunston and Francis (2000). The use of different patterns to express collocations can be related to Sinclair's idiom principle features (1991) (see section 5.2.1 for a detailed review of these features). He claims that many phrases allow for internal lexical variation; a number of collocations' patterns have been extended in this thesis using different lexical insertion such as ADJ in *following+ADJ+code* or PRP in *System+PRP+resources*. Moreover, the collocations that allow for variation of the word order, such as *available resources* that can also be *resources available*, seem in line with Sinclair's idiom principle feature in which he confirms that some phrases allow for variation of word order. However, in some cases, single patterns were used by single or few writers: they are more likely related to the idiosyncratic style of the writers.

The overuse of the shared patterns in most of the identified collocations could be explained by genre variations. That is, writing in research articles is different from writing in dissertations. Students are required to write in more detail and to include more chapters in their dissertations, whereas expert writers are restricted to a certain number of words according to the rules set by the journal. Different writing demands are required in these two genres.

Writing MSc dissertations that may include some type of programming or experiments required a detailed narrative style, as P3 commented: *"I suspect a lot of narrative description of their programming is in these dissertations. Most dissertations in CS develop programs or some practical experiments*." Thus, students tend to overuse collocations that are related to programming in their writing of dissertations, such as *source code, environment development*, and *open code.* Two CS experts were able to confirm that writing a program required much repetition in the students' writing of dissertations. P1 commented in the overuse of the *following code* by both NNS and NS students: "*It would be most likely that students are more interested in demonstrating their ability in and their understanding of using the code. They show examples of codes. Thus, they use demonstrative phrases like 'the following code'. Whereas in writing an article, they would not be likely to include examples of code, so they might not be used*". On the other hand, writing in research articles does not require writing programs in detail. Instead, expert writers assumed that their programs are developed. I may relate this variation to the purpose of writers.

Hyland (2008) claims that writing in research articles is concerned with "persuasive reporting through review process and engagement with the professional world" (Hyland, 2008: 56) thus it is related to 'norm developing' rather than 'norm developed', as described by Swales (1990). The main aim of writing a research article is to 'disseminate academics' research and establish their reputations, exhibiting to colleagues both the relevance of their work and the novelty of their interpretations" (Hyland, 2008: 57). On the other hand, when writing dissertations, students are concerned with only the reader of their work. One of the CS experts raised this issue by commenting on students' writing style as industrial rather than academic: "*I think because dissertations are focused more on industrial professional style rather than academic writing, both NNS and NS [students] are not writing in an academic way. Their formal reports may*

include some academic writing, but it cannot be compared to research articles". It can be concluded that different writing styles in two genres are related to different purposes (Hyland, 2008).

Another interesting finding is the overuse of the collocations *following/previous section* in NS students' writing rather than in experts'. Since students need to write in more detail, they will include more sections in their writing of dissertations. Thus, the use of these collocations is expected to be more frequent in students' writing. Hyland (2008) has classified these collocations as text reflective markers, which were also used more frequently in his students' corpus of Masters and PhD dissertations rather than in his corpus of research articles. Students' genres are more 'phrasal' than the research articles since students depend on using formulaic language (collocations in this study) in developing their arguments more than experts do (Hyland, 2008).

Moreover, comparing the use of collocations between non-experts (both NNS and NS students) and experts demonstrates their different levels of knowledge. Students are more concerned in their writing of dissertations to show their knowledge of developing certain programs or software, unlike expert writers who write to publish their works. Experts' writing can be described as knowledge transforming since the process of "peer review works as a control mechanism for transforming beliefs into knowledge". Unlike experts' writing, students' writing can be seen as an example of knowledge telling since they "demonstrate a suitable degree of intellectual autonomy while recognising readers' greater experience and knowledge of the field" (Hyland, 2008: 47).

Another factor that could explain the variation of the use of some N collocations is more likely related to the choice of topics selected in compiling the three corpora. CS experts confirm that topic has an effect on the overuse of some collocations such as *development environment, source code,* and *following code*. On the other hand, other collocations were classified as SCSC as they are related to a specific discipline and, thus, they tend to be topic-specific. For example, *data input* was classified as SCSC in the CJT for AI and was confirmed by CS experts as occurring more in AI than the other two disciplines when topics were checked. The underuse of the collocation *method class* was also related to its specific occurrence in a certain topic that required the use of a certain programming language, Java. This finding is in agreement with Peacock's (2012) and Ward's (2007) assertion that collocations are very discipline specific. Analysis of the 24 shared N collocations reveal some disciplinary differences in the collocates of high-frequency nouns.

An interesting variation observed between NNS and NS students was their use of long extended collocations rather than fixed collocation. NNS students prefer to use long extended collocations such as *data collection of user information* instead of *data user* and *information content of the data* rather than *data information.* Even though these extended collocations are accepted, they seem non-native like and unidiomatic. This finding seems in agreement with Hunston and Francis' (2000) claim that observing native and non-native use of patterns will reveal how much control the non-native writers have over their second language. Even advanced learners of language tend to have imperfect control over patterns. If non-native learners use a word in a correct grammatical pattern, their usage may be unidiomatic rather than wrong. In addition, Hill (2000) notes that NNS students use long phrases instead of using fixed phrases when they do not recognise the right expressions. Thus, NNS students need to raise their awareness about the fixed

use of collocations and their patterns. Chapter 6 will be devoted to presenting some awareness-raising activities.

5.5 Conclusion

This Chapter has presented collocations' patterns' identification and experts' interviews as well as the CJT. In particular, I have first referred to the literature on patterns' identification and then presented the methodological steps applied in pattern identification, experts' interviews, and CJT. This Chapter has further presented and discussed the findings of patterns' identification of the 24 shared N collocations among the corpora and experts' views about the factors behind the different uses of some of the N collocations. In addition, CJT findings have also been discussed in detail.

Chapter 6 Academic Collocations' Awareness-raising Activities

6.1 Introduction

While increasingly more EAP units are developing in-house materials for teaching collocations (e.g., McCarthy and O'Dell, 2005; Schmitt and Schmitt, 2005; Barlow and Burdine, 2006), these do not often focus on students' problematic over/underuse of collocations and do not take into account disciplinary variation. Thus, designing specific materials for raising students' awareness of the problematic over/underuse of collocations could be very useful. In this Chapter, I present our third study, designed to answer our seventh research question: What kind of teaching materials are needed to raise NNS students' awareness of the use of academic collocations? The Chapter comprises two sections. First, I review the main issues related to teaching collocations (section 6.2), the corpus-based approach to teaching collocations: Data-Driven Learning (DDL) (section6.3), cognition and L2 vocabulary learning (section6.4),taxonomies of awareness-raising activities (section 6.5), and types of collocation activities in ESL textbooks and corpus-based research (section6.6). Second, I present three types of activities that I designed for raising CS NNS students' awareness of some problematic collocations' use and patterns (section 6.7).

6.2 Teaching Collocations

6.2.1 The Importance of Teaching Collocations

Collocations are considered as one of "the most powerful forces in the creation and comprehension of all naturally occurring text" (Hill, 2000:49) in the mental lexicon of any individual. Any individual needs to have enough collocational competence to be able to recognise and use collocations easily and effectively. Lack of this competence leads to a number of difficulties: mainly the overuse of a limited set of collocations, use of long expressions instead of using precise collocations, and producing odd and foreign combinations of words of English, which might be a translation of words from the students' L1(Sinclair, 2004). Thus, teaching collocations is essential. Hill (2000: 59) insists on teaching collocation from lesson one, since "collocation is not an added bonus which we pay attention to once students have become sufficiently advanced".

Jiang (2009) and Lewis (2000) claim that English language learners do not need to learn new words, but rather learn collocations of the words they already know. Lewis (2000) recommends teaching students collocations of familiar words to extend their collocational competence rather than teaching new words. From their pre- and post-tests of 41 Japanese students, Webb and Kagimoto (2011) found that productive learning of collocations is increased when learners "learn multiple collocates for a small number of node words than to learn a smaller number of collocates for a large number of node words" (270). By applying this approach, the collocation learning burden for NNS learners would be reduced and the productive learning of collocations would be maximised.

It has been assumed that NS learners already have knowledge not only of an "enormous number of individual words but also know much more about how these combine or collocate" (Conzett, 2000: 74). Thus, it seems that NNS learners need to pay more attention to the "syntagmatic relations of collocations between lexical items" (Gitsaki, 1999, cited in Ying and O'Neill, 2009) to build their active lexicon.

Traditionally, language teachers tend to focus on teaching grammatical features as they consider them the main challenge for NNS learners; they may not notice that their overemphasis on teaching grammatical features rather than focusing on collocations would prevent their learners from advancing from their 'intermediate plateau'(Lewis, 2000:14). Therefore, it seems reasonable to focus on teaching collocations to NNS students (especially to those in the intermediate level) rather than focusing on teaching grammatical features (Hill, 2000).

6.2.2 Approaches to Teaching Collocations

It seems that learning collocations can be facilitated by teaching. Teachers can help facilitate the learning of collocations in a number of ways, as suggested by Hill (2000). First, teachers can teach collocations as they teach new words. Whenever new words are taught, it is better to teach their collocations as well. However, at lower levels, it would certainly be considered a 'learning burden' to try to cover too many aspects of the lexical information about new words at once (Nation, 2001).

Second, teachers can make learners aware of the vital role of collocations in language learning by asking them to notice two- or three-word expressions rather than looking for individual

words. Thus, noticing may lead to raising learners' awareness of collocations (see section 6.4). Moreover, teachers can extend their learners' collocational competence of words they already know. A learner with 2,000 words who is equipped with collocational competence is more communicatively competent than a learner with 2,000 words who does not have collocational competence (Lewis, 2000). This approach would be more usual in teaching collocations since the focus is on teaching collocations of already-known words rather than on teaching collocations of new words.

Even though some researchers (e.g., Conzett, 2000; Woolard, 2000) have called for the independent learning of collocations, teacher guidance is still needed. Woolard (2000) recommends equipping learners with search skills to enable them to discover significant collocations by themselves, in both the language they meet inside the classroom and in the language they encounter outside the classroom. Hill (2000) insisted on the teacher's role of guiding learners to be independent collectors of collocations. If learners are trained to "notice common collocations in the texts they meet, they will be able to select those collocations which are crucial to their particular needs" (Woolard, 2000:35).

Various resources that can help learners maximise their opportunities to acquire knowledge of collocations outside the classroom have been suggested by Woolard (2000) and Lewis (2000b). First, collocation dictionaries can provide useful information on collocations by exemplifying collocations in sentences. However, they are underused resources in language learning (Nesi, 2014; Boulton, 2008). Second, corpora and concordances provide much richer sources of collocations than dictionaries. Johns (1991a) has demonstrated the value of using concordances

in language learning, by developing his Data-Driven Learning (henceforth DDL; more details about this approach will be given in section 6.3).

Third, lexical notebooks in which learners record their collocations are also of great benefit (Woolard, 2000; Schmitt and Schmitt, 1995). Learners need to store and record their learned collocations so that they can be revisited and retained for whenever they need them. The notebook is not only "a decoding tool, but a resource which learners can use as encoding instrument to guide them in their own production of language" (Woolard, 2000: 44). Learners may need to be guided in how to organise their collocations in notebooks. Though it might be considered an old-fashioned approach, it can be used in a modern way to store learners' learned collocation in their mobile phones or laptops. In conclusion, it is probably true that learners need to be encouraged to learn collocations independently. However, the need for teacher guidance cannot be dismissed, especially in the DDL learning.

6.2.3 Selecting which Collocations to Teach

An important question that arises when a teacher aims to teach collocation is which collocations to teach. Language teachers have to avoid presenting all collocations found in a text.

Woolard (2000) suggests teaching collocations that are misused by learners. Teachers have to search for the problematic collocations that have not been produced correctly by learners in their production of language. Woolard's approach depends on raising learners' awareness about their misused collocations. It is equivalent to error-recognition and correction activities in which learners learn from their negative language samples (Granger and Tribble, 1998; Nesselhauf,

2004; Thornbury, 1999). Thornbury (1999: 122) points out that using learners' errors for awareness-raising or consciousness-raising purposes can customise the lesson, tailoring it to the specific problems learners have.

Another criterion for collocation selection that has been applied by Hill (2000) in his teaching of collocations is collocational strength. Collocations can be seen in a cline or spectrum of strength starting with unique collocations, strong collocations, medium-strength collocations, and finishing with the weak collocations. Unique collocations are the fixed collocations that may not be of interest to learners. Strong collocations are those that contain one word that collocates with few other words (e.g. *trenchant criticism*).These types of collocations are rare and considered obscure when compared to other types of collocations. Therefore, it is advisable not to replace "teaching obscure words [with] teaching obscure collocations" (Hill, 2000:60).The weak collocations are the ones that contain words that can be occur with many words and they are flexible (e.g. *red shirt, red car*).

The medium-strength collocations are the ones where learners may not be aware of their uses, as they resemble free expressions. For example, learners may know the word *hold* and *conversation*, but may not know that they can make a collocation *hold a conversation* where *hold* does not have its usual concrete meaning (Hill, 2000:64).Thus, it would probably be noted that the main learning target for most language learners should not be the strong or weak collocations but the medium-strength collocations.

A third method of selection is to teach collocates for synonymous words, as recommended by a number of researchers (Hill, 2001; Woolard,2000) who thought that it would have a positive

effect on collocation learning. Woolard (2000) suggested using concordances to help define the difference between synonymous verbs (*treat* and *repair*). Learners are presented with concordances for the verbs *treat* and *repair* and asked to look at the sentences to define the difference between them. Although presenting NNS learners with collocates of synonymous words might be effective, a negative effect has been found by Webb and Kajimoto (2011) when their Japanese learners were tested in their use of synonyms to increase their collocation knowledge.

Fourthly, frequency has been considered the main criterion for selecting which collocations need to be taught via the corpus-based approach. It has been suggested that teaching the most frequent collocates of the most frequent node words would be of greatest benefit to language learners (Nation, 2001, 2008; Webb and Kagimoto, 2011). In their design of awareness-raising activities for first year PhD Engineering students, Jones and Durrant (2010) selected the most frequent nine node words that occurred in their compiled corpus of Engineering research articles. They added two other criteria: the selected words should occur in the AWL (Coxhead, 2000) or in the students' reading texts. The final two criteria were also reported by Nation (2001, 2008) who claimed that only most frequent 2000 word families followed by Coxhead's (2000) AWL and words that fulfil a need should be explicitly taught.

Different criteria can be applied to choose which collocations to teach either by considering their problematic collocation through their misuse or by discovering which collocations are most encountered by ESP students in their textbooks. However, it should be remembered that in the ESP context, learners' needs should be the main concern (Dudley-Evans and St. John, 1998).

6.2.4 Formulaic Language Processing and the Teaching of Collocations

One of the main questions that has to be considered in teaching collocations is whether to teach collocations holistically (as unanalysed wholes) or analytically (in parts).

It has been suggested by Wray (1999, 2000, 2002) that formulaic language (including collocations in this study) is best processed when reading, writing, etc. Via a holistic approach rather than in an analytic approach. Wray's holistic approach is in agreement with Sinclair's idiom principle (1991): in this principle, learners ideally should bring about the selection of two or more words together, based on their previous and regular occurrence together. That is, when learners store ready-made frameworks in their memory, they can easily use them later and avoid the labour of generating a novel one.

The main advantage of applying the holistic approach is its economy and speed in reducing the time of recognition and production of already stored words (Wray, 2002; Nation, 2001). On the other hand, the main disadvantage of chunking is storage. "If chunks are stored in long-term memory, then there will be a lot of items to store" (Nation, 2001: 321). Another disadvantage of chunking is that the parts stored in chunks will not be available for creative combination with other words (Nation, 2001).

Wray's analytical approach "entails the interaction of words and morphemes with grammatical rules, to create, and decode novel or potentially novel linguistic material" (2002: 14). This

approach is similar to Sinclair's (1991) open-choice principle in which learners would create and encode novel and creative sentences whenever they need them; it is the same kind of creative model as is assumed in the Chomskian account of language processing. The advantage of the analytic approach is its flexibility for novel expressions (for more details of Sinclair's principles see section 5.2.1).

The link between these two approaches has been clearly established by Sinclair (1991) who proposes that, ideally, when reading, "The first mode to be applied is the idiom principle, since most of the text will be interpretable by this principle. Whenever there is a good reason, the interpretive process switches to the open-choice principle, and quickly back again. Lexical choices which are unexpected in their environment will presumably occasion a switch" (1991: 114). Thus, it can be concluded that both approaches are connected and that the switch between them is based on the reader's existing store of formulaic chunks, including collocations.

To apply this now to teaching, the holistic approach might be preferred for teaching collocations as it is believed to be more economical for later processing, but the analytic approach is thought to be more effectively applied with L2 learners who need to raise their awareness about collocations. Thus, it is better to make them aware of parts of the most frequent collocations first. Then, they can be introduced to collocations in chunks.

6.3 Corpus-based Approaches in Teaching Collocations: Data-Driven Learning (DDL)

Corpora have had a great impact on language learning and teaching. The corpus-based approach in which language learners are exposed to a set of concordance lines to investigate language features has enormously contributed to enhance language learning (Gavioli, 2005; O'Keeffe and McCarthy, 2010; Boulton et al., 2012). Hence, concordancing is considered a valuable tool for both teachers and learners in language pedagogy (Johns, 1986; Aston, 1995; Gavioli and Aston, 2001; Gabrielatos, 2005). The use of concordances in language teaching is mainly related to DDL and was first advocated and developed by Johns (1986, 1991 a, 1991b). According to Johns and King (1991: iii), DDL is defined as:

"The use in the classroom of computer-generated concordances to get students to explore the regularities of patterning in the target language, and the development of activities and exercises based on concordance output".

In a corpus-based DDL classroom, the language learners are generally provided with concordance data to enrich their 'language awareness' (Hawkins, 1984; Van Lier, 1995) and/or to lead to 'consciousness-raising' (Ellis, 1992; Rutherford, 1987; Sharwood-Smith, 1990). Learners are encouraged to be engaged in discovery learning and to build their autonomy since language is presented in a way that allows learners to discover new knowledge for themselves, rather than being spoon-fed. The discovery learning is conducted by providing authentic language examples, rather than examples created by teachers.

In this respect, corpus-based DDL can be categorised as a form of inductive learning in which students work on concordance output to generalise language regularities and patterns for

themselves rather than by receiving explicit explanations from teachers deductively (Boulton, 2009, 2010). However, it should be noted that the corpus-based DDL approach is different from other inductive learning approaches. Some distinctive features of DDL are summarised in what follows.

First, language input is presented in the form of concordance lines, which are authentic language samples extracted from pedagogically useful corpora. Concordance lines are usually presented in the KWIC format in which words, phrases, or combinations of words are clearly displayed in the middle of concordance lines (Kennedy, 1998; Kettemann, 1996; Sinclair, 2003) and can be read vertically (Boulton, 2009). An example of concordance lines resulting from searching for *source* in the RC is given in Figure (6-1).

Figure 6-1: A KWIC format example of concordance lines for *source*

1 …JADE framework is an open ***source*** project distributed by…
2 …is to implement an open ***source*** Real-Time Operating System…
3 …look at a couple of open ***source*** operating systems is followed…
4 …this includes many open ***source*** systems, highlighting the…
5 …this is a free open ***source*** piece of software that has…

As can be seen in Figure (6-1), when learners cast their eyes down the middle column of the concordance lines, they will gradually recognise that the word *source* is always preceded by *open*. In this way, concordance data presented in the KWIC format makes it easy for learners to see what words occur immediately before and after the keyword.

Second, corpus-based DDL is a new way of language learning, in which learners are encouraged to work on concordance data to discover language patterns and use them. Students are mainly required to play the role of the linguistic researcher or language detective instead of being passive recipients of knowledge from the teacher. As Johns put it, "research is too serious to be left to researchers" (1991a: 2) and this is why every student should become 'a Sherlock Holmes' (1997: 101).

Third, DDL involves a strong form of consciousness-raising or awareness-raising that can be particularly useful in drawing learners' attention to particular language features and developing their inductive learning strategies as a language-learning tool. O'Sullivan (2007:277) provides an impressive list of cognitive skills that DDL may be supposed to promote, many of which presumably also apply to paper-based materials: "predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying".

By applying these skills, learners will not only develop their linguistic skills but also their cognitive and meta-cognitive skills, which will lead to greater autonomy and better language learning skills in the long term (Boulton, 2009, 2010). Johns (1991b) argues that the development of these skills will help learners learn how to observe any type of language data and make useful generalisations, within and beyond the classroom. Therefore, the corpus-based DDL approach has been described as process rather than product-oriented, learner rather than language-centred, meaning rather than form-focused (Bernardini, 2001).

Finally, yet more importantly, the teacher's role in the DDL classroom is different from the traditional authoritative language input, as teachers act "as research director and research collaborator rather than transmitter of knowledge"(Johns,1986:14).Teachers prepare concordance-based material in response to language problems raised by learners. Thus, students are encouraged to raise their problems either in the classroom or during consultation time outside the classroom.

On the other hand, a number of researchers have pointed out some barriers to the use of corpus-based DDL (for detailed discussion, see Chambers, 2007; Farr, 2008; Boulton, 2010; Boulton, 2012). These barriers are mainly related to the implementation of DDL rather than to the nature of this approach. Boulton (2010) discussed three main fears. First, it is assumed that DDL can best be applied with advanced learners as recommended by Johns (1991a). Indeed, Boulton's (2008) survey of 39 empirical DDL studies found that only four studies were applied to lower-level learners. However, the results of these studies do provide positive evidence from the use of DDL with lower-level learners (Tian, 2005; Yoon and Hirvela, 2004). Second, DDL has been described as a waste of time and effort since it requires the use of specialist resources and extra training for both teachers and learners.

Third, technological considerations have been viewed as one of the main barriers to the introduction of DDL. Some teachers may have 'technophobia' and lack the ICT skills to use DDL with their learners or they may be afraid that their learners are better in their ICT skills. Moreover, teachers may not have regular access to computer laboratories (Tian, 2005) while other teachers may feel uncomfortable teaching in computer laboratories, for a variety of reasons (Farr, 2008).

To overcome these barriers, Gabrielatos (2005) recommends ordinary teachers and learners using DDL in ordinary classrooms by using paper-based materials prepared by the teachers in advance(soft version of using DDL: this version will be discussed in detail below). A number of papers show learners using paper-based materials successfully as a reference source (Boulton, 2008, 2009, 2012) as well as for learning different aspects of General English language (e.g., Allan, 2006; Koosha and Jafarpour, 2006) and for ESP (Boulton, 2012; Boulton et al., 2012).

6.3.1 Main Approaches to DDL

Two main approaches were recommended by Leech (1997) when using concordances in language teaching: the soft version with 'paper-based materials' (Boulton, 2009, 2010) and the hard version employing 'hands-on concordancing' (Boulton, 2009, 2010). The soft version involves teacher-designed and selected concordance materials in the form of printouts whereas the hard version involves learners conducting autonomous or independent concordancing themselves by directly accessing a concordance program using computers, CDs, or web-based online tools.

In the soft version, the teacher has access to a corpus and the relevant software, prints out concordance samples from the corpus, and designs tasks and activities (Gabrielatos, 2005; Boulton, 2009, 2010). Learners are introduced to these corpus-based materials in paper form and have to examine concordance lines to be able to complete the given tasks (Bernardini, 2004; Granger and Tribble, 1998; Tribble and Jones, 1990; Cresswell, 2007).

On the other hand, in the hard version, learners have direct access to a corpus and have to use their skills to investigate the corpus. Thus, the teaching burden will be less. The tasks and activities in this version can be presented in three ways, as suggested by Aston (1995): they can be created by the teacher (Tognini-Bonelli, 2001), incorporated into CALL programs (Hughes, 1997; Milton, 1998), or selected by the learners, with or without the instructor's involvement and management (Bernardini, 2002).

It is clear that the use of soft DDL is more effective than using "full-blown hands-on concordancing" (Boulton, 2010) since it reduces the learning burden and technological difficulties. Learners are allowed to gain insights into selected data and learn to interpret limited set of data before they engage in the full discovery process.

6.3.2 DDL Awareness-Raising Studies

The DDL approach can be particularly useful in drawing learners' attention to specific language features and developing their inductive learning strategies as a language-learning tool. A number of studies have been conducted to raise learners' awareness of different linguistic aspects using DDL as their main approach (Thurstun and Candlin, 1998; Kübler and Foucou, 2003; Kheirzadeh and Marandi, 2014).

Kheirzadeh and Marandi (2014) used the hard version of the DDL to raise their EFL Iranian students' awareness of the benefits of using concordancing in the learning of collocations and to discover which type of collocations are frequently searched by their EFL students. After introducing their 27 Iranian students to the tools and benefits of corpus in the learning of

collocations in the first two sessions, they trained their students to search for the collocations they think they needed most in their study as well as for other sets of collocations given by their teachers in the next five sessions.

Using the *Compleat Lexical Tutor*, the students were asked to undertake a small research about the collocations they felt that they needed most and to write down the results and samples of their findings. They were also asked to write down their comments about the pros and cons of the use of corpus in their learning of collocations. Moreover, five students were interviewed on the same issue. The results showed that the students were completely satisfied with their use of concordance in their learning of verb and noun collocations, which were the most frequently searched collocations. They realised that using concordancing is useful for learning collocations and in recognition of different uses of verb noun collocations as well as their different patterns used.

Kübler and Foucou (2003) also applied the hard version of DDL in their teaching of CS verbs to French speakers to describe verbs and their syntactic differences between English and French and to raise their learners' awareness about these variations. Using contrastive corpora – specific CS English corpus, English and its French equivalent corpora, and general English corpus – three types of verbs were identified: highly technical verbs, general verbs with specialised uses in CS, and general verbs. When the three types of verbs were compared with their French equivalents, differences in their syntactic structures between French and English were identified. The first two types of verbs were considered more problematic than the third type. Thus, they were searched for their equivalents in French and were identified for their syntactic structures.

As a result, students were able to observe the different verb structures from contrastive concordance samples and were able to look for their equivalences in the parallel corpus. The description of different verbs structures was useful in designing gap-filling exercises. It can be concluded that DDL can be used to raise learners' awareness of different linguistic aspects and their syntactic structures. Thus, it has been chosen to be the main approach in designing awareness-raising activities in the current study. In the following section, I will review research on the Depth of Processing theory, which will highlight the learning processes. These processes need to be considered in designing the awareness –raising activities.

6.4 Cognition and L2 Vocabulary Learning: Depth of Processing

The Depth of Processing theory, which has been applied in Applied Linguistics, can be applied in learning new words as well as their collocations. It has been analysed in a series of processes. Nation (2001) identified three main processes – noticing, retrieval, and generative processing – that are involved in learning a new word. Stahl (1985 mentioned in Nation, 2001) proposed similar components of processing but under different terms: association, comprehension, and generation. Laufer and Hulstijn (2001) proposed a different set of processes: 'need', 'search', and 'evaluation'.

In Nation and Stahl's views, the first process is finding out the basic form-meaning connection of a word. During 'noticing' the learner views the item on which s/he is focusing his or her attention as separate from the message of which it forms part (Nation, 2001:64), whereas

‘association’ refers to the end product of the process of association of form and meaning rather than attempting to explain how it takes place.

‘Noticing’ has been considered an essential step in language learning to make learners aware of the meaning of new words and their collocations (Nation, 2001).Teachers can have a direct influence on ‘noticing’ by using different techniques in listening and reading tasks, such as pre-teaching, highlighting the target words by using underlining, italics, or bold letters, and glossing the word, which will result in raising learners’ consciousness of the required words and their collocates. Teachers have to select interesting ways to encourage learners’ noticing by keeping their motivation high, since “motivation enables noticing” (Nation, 2001:63).

The second major process is the ‘retrieval’ of what has been learned about a lexical item. After learners are introduced to information about new words through teacher explanation, dictionary use, or self-guessing, learners need to repeatedly retrieve what they know about the learned words when they hear or see them again. Thus, repetition of the learned words is important to ease the retrieval of them later. However, Nation (2001:67) points out two major factors that may affect the process of retrieval: the learner’s vocabulary size and the length of time that the memory of a meeting with a word lasts. Nation notes, “The larger the vocabulary size, the greater the quantity of language that needs to be processed in order to meet the words to be learned again” (2001: 67).

Nation (2001:72) proposes the serialisation of stories as a way to encourage retrieval of the taught items, since vocabulary tends to be repeated in long stories. In oral activities, retrieval is

encouraged by making it necessary for learners to use input words: hence, Nation (2001: 72) proposes the ‘strip story’, a method initially proposed by Gibson (1975 mentioned in Nation, 2001). In this task, each learner learns a sentence from a paragraph by heart. Learners have to cooperate to put the text together. No writing is allowed so retrieval is required to complete the task.

The final process under the Depth of Processing is ‘Generative Processing’ in Nation’s terminology (2001), which corresponds to Stahl’s ‘Generation’. This term refers to the novel production of already taught lexical items in ways different from before. Nation (2001:73-74) proposes a number of ways for the promotion of generative processing, such as the presentation of a word in a different context in serialised stories, asking learners to retell a story, and encouraging learners to negotiate the written text and reconstruct its parts rather than repeating it, thus creating an opportunity for them to use taught vocabulary generatively. These three processes could be also applied in learning collocations.

Laufer and Hulstijn (2001) proposed another way of breaking the Depth of Processing into more concrete concepts. These concepts are ‘need’, ‘search’, and ‘evaluation’. The first concept is motivational (‘need’ to achieve by finding out) and the other two concepts are purely cognitive processes. ‘Search’ is the search for the meaning of a word or the form that expresses a certain concept. The ‘evaluation’ concept involves the comparison of the possible interpretations of a word so that the interpretation most appropriate to context will be selected. Laufer and Hulstijn (2001: 15) state, that all things being equal, “the higher the cumulative degree of these processes (called ‘involvement load’), the better the retention of the words learned”.

These attempts to analyse learning into distinct processes are useful for pedagogical purposes because they make clear claims about the features a vocabulary-learning task should have. Most importantly, noticing and motivation seem essential for successful vocabulary learning. For this reason, awareness- raising activities in this study were mainly focused on noticing collocations and their patterns.

6.5 Taxonomies of Awareness-raising Activities

Raising learners' awareness about language features has been viewed as the main first step to facilitate their learning. Noticing is the starting level in which learners' attentions are directed toward specific features of language including collocations (Schmidt, 1992; Nation, 2001). On the basis of this noticing, learners may develop the second "deep level of cognitive awareness by employing various cognitive strategies for deep processing of the noticed features in the input, thus having a greater chance of internalising them" (Ying and O'Neill, 2009:183). Since noticing is the first step in which learners are exposed to language features, it should be applied in language teaching to raise learners' awareness of new lexical or grammatical features in general. Schmidt and Frota (1986 cited in Ying and O'Neill, 2009) claim, "Those who notice most learn most". Thus, I need to see how to apply this idea in DDL teaching of collocations.

When available literature was reviewed, we found that few taxonomies have been developed of activities for raising learner awareness of different linguistic features. Dave and Jane Willis' consciousness-raising taxonomy (cited in Lewis, 1997: 53), which results in "an increased awareness of and sensitivity to language", consisted of seven stages, as follows:

1-Students search to identify a pattern or usage and the forms associated with it.

2-Students classify according to similarities and differences.

3-Students are asked to check a generalisation about language against more data.

4-Students are encouraged to find similarities and differences between patterns in English and those of their own language.

5-Students manipulate language designed to reveal underlying patterns.

6-Students recall and reconstruct parts of a text, chosen to highlight a significant feature.

7-Students are trained to use reference works. (Adapted from Lewis, 1997: 53).

Observing Dave and Jane Willis' steps of raising-awareness, it seems that these can be related to the corpus-based DDL approach developed by Johns (1991b) in which learners are required to start with research then practice and, finally, apply this same process to be able to learn new linguistic features. Learners begin by looking at concordance lines for the key terms or grammatical feature under investigation, in our case collocations, trying to think of their meaning or use. In the next stage, learners familiarise themselves with the patterns of language surrounding the key terms or grammatical features. Then, they practice key terms and grammatical features by themselves without referring to the concordance lines.

Finally, they produce their own writing by using the key terms or grammatical features investigated. Thurstun and Candlin (1998) have applied this approach in their development of corpus-based activities to raise their learners' awareness about the use of rhetorical function words in academic writing (detailed information about this study will be provided in section 6.6.2).

Both Dave and Jane Willis' taxonomy and Johns' (1991) DDL approach loosely follow the deep processing learning theory developed by Nation (2001). Dave and Jane Willis' first three steps are equivalent to the noticing stage, the fourth and fifth steps are equivalent to the retrieval stage, and the final two steps are equivalent to the generative production stage. Similarly, Johns' (1991b) approach is equivalent to the three stages of deep processing. What he calls research is equivalent to noticing, practice to the retrieval stage, and improvisation to generative production.

By contrast, Ying and O'Neill (2009) developed their 'AWARE' process-oriented approach (see Figure (6-2) below) and conducted a study with two purposes: first, to investigate Chinese students' perspectives and practices in relation to collocation awareness-raising through the adaption of AWARE and, second, to discover about the difficulties and problems encountered during their use of this approach. 20 adult participants at intermediate level of language proficiency were interviewed before and after the language programme and their reflective journals were analysed. The study followed the steps of the AWARE model as follows.

Figure 6-2: The steps of the AWARE model adapted from Ying and O'Neill (2009:183).

A: Awareness raising of important language features, in particular collocations (helping learners notice collocations in the weekly theme-based reading or in any other source of input)

W: Why should we learn collocation? (Helping learners see the rationale for/meaning of learning what they learn)

A: Acquiring noticed collocations using various strategies (learners making selective use of a repertoire of learning strategies that suit their individual learning style to promote effective learning of collocations)

R: Reflection on learning process and content (learners thinking about their learning processes and making necessary adjustments for better learning)

E: Exhibiting what has been learned(learners making a weekly oral report in class on the theme under focus by using as many as possible of the collocations they have noticed and learned)

After applying their 'AWARE' approach in an ESL course of five months, 20 adult learners were interviewed and their reflective journals were analysed. The majority of the students were able to manage their collocation learning independently and were able to adjust their learning based on what they thought worked well or what did not. Even though a number of learning problems were encountered in different stages of their learning, such as the inability to judge what exactly needed to be noticed and to decide what and how to reflect on their learning, the majority of learners were able to overcome these problems independently and make necessary changes overtime.

Learners were positive about the focus on the learning of collocation and felt that learning collocations is of great significance for them to improve their language proficiency. Based on their results, Ying and O'Neill (2009) advise language teachers to apply their 'AWARE' approach in their teaching of collocations to help their students who are in their intermediate level of proficiency to learn this aspect of language independently and effectively.

Even though their 'AWARE' approach highlighted an effective method for guiding learners in their learning of collocations, similar steps can be adopted in designing awareness-raising activities by first directing students' attention to notice collocations in the 'A' step and then by highlighting the importance of learning collocations to learners in step 'W'. The final three steps would be better applied in a classroom setting. However, it would be more useful in training students in their self-training rather than in teaching collocations. Since the main aim of the current study is to design a sample of awareness-raising activities for NNS CS students to be

taught by teacher, the first three stages of Dave and Jane Willis' consciousness-raising taxonomy, which were equivalents to Nation's noticing stage, were adopted.

6.6 Collocation Activities

6.6.1 Traditional Collocation Activities in EFL Textbooks

Collocations have been presented in EFL textbooks in various ways. Hill et al. (2000) suggested two main types of activities for teaching collocations: reading text or using dictionaries. In the first type, learners search for collocations from their reading texts either individually or in groups. In the second type, learners are provided with a set of words to search for their collocations using a collocation dictionary. Multiple exercises have been suggested for the dictionary-based activities; 'correct the wrong word', 'find opposites or synonyms', 'odd word out', 'short paragraphs', and 'arrange words into groups' (Hill et al., 2000). No matter which type of activities are designed and adopted by teachers, what matters most is selecting activities that encourage learners to notice collocations in ways that maximise the chance of input being retained as long-term intake.

To investigate the effectiveness of the most frequent types of activities to teach verb-noun collocations, Boers et al.(2014)located six types of verb-noun collocation activities from their manual checking of 11 pedagogic materials (for more information about these materials see Boers et al., 2014: 8). Four types of verb-noun collocation activities were categorised as frequent since they were located in most of the materials. These activities were 'Connect', 'Choose and insert the verb', 'Indicate the right verb', and 'Choose and insert the collocation'.

Boers et al. (2014) mentioned advantages and disadvantages of the four selected types of activities when evaluated for their effectiveness in raising students' awareness of collocations. The first three types were considered unsuitable for raising learners' awareness about collocations, as learners are required to establish appropriate matches between sets of verbs and nouns. These activities conflict the psycholinguistic view of the way collocations should be presented (see previous section 6.2.4 discussing the psycholinguistic view of learning collocations).

On the other hand, the fourth type 'Choose and insert the collocation' is quite different in its presentation from the previously mentioned types since collocations are presented as chunks (Boers et al., 2014).Thus, it appears more in agreement with the psycholinguistic view of presenting and processing collocations. However, it should be noted that their activities were not corpus-based DDL. Their empirical results reveal that 'Choose and insert the collocation' activity which presented collocations in chunks was more beneficial to language learners than the other three types of activities that present collocations in parts.

The final two additional formats of activities 'Correct the wrong collocations' and 'Odd one out' , which were similar to Hill et al. (2000) activities seem "less geared towards the retention of new, correct collocations" (Boers et al., 2014: 17). The main disadvantage of these two types of activities was that they direct learners' attention, in the first instance, to what is not to be remembered.

To summarise, most of the traditional types of activities presented collocations in parts rather than in chunks except 'Choose and insert the collocations'. They were proved ineffective in raising learners' awareness of collocations. However, what about corpus-based DDL activities? Would they be effective in raising learners' awareness of collocations? To answer these questions, types of corpus-based awareness-raising activities designed by a number of researchers will be reviewed in the following section.

6.6.2 Corpus-based Awareness-raising Activities

A number of researchers have designed corpus-based activities using the soft version of DDL to raise learners' awareness of certain linguistic features (Tribble, 1990; Tribble and Jones, 1990; Johns, 1991a, 1991b; Hyland, 1998, 2003; Thompson and Tribble, 2001; Yoon and Hirvela, 2004).

Jones and Durrant (2010) designed a set of awareness-raising activities for first year PhD Engineering students to direct their attention to the use of the most frequent academic words. For this purpose, they compiled a discipline-specific corpus of 11,624,741 words from Engineering and Science research articles. First, they examined the 50 most frequent keywords in their corpus to extract a few words to be included in the awareness-raising activities. Nine words (*average, behaviour, consequently, higher, positive, presented, response, shown* and *study*) were selected according to three criteria: words that occur frequently in corpus data across disciplines, words that frequently occurred in students' sample text, and words that occur frequently in all of the selected sources, corpus data across disciplines, students sample texts, and in the AWL (Coxhead, 2000) or in two of these sources.

They set out a number of questions to be answered after learners were introduced to concordance lines. The primary focus was on examining and analysing KWIC in detail by looking at the right and left context of the KWIC. Even though their awareness-raising activities were aimed to be focused on discipline-specific words, they were designed to focus on the most frequent academic words that were thought to be more pedagogically useful. In addition, they claimed that it would be more useful to introduce PhD learners who had no experience of using concordances to first study familiar words and then move to their discipline-specific words and their collocations. Thus, it would be more effective and less threatening to display familiar words to learners in their first encounter with corpus-based activities rather than displaying discipline-specific words.

Jiang (2009) developed a set of self-designed activities to improve awareness and productive use of L2 collocations of Chinese secondary school students. Four main types of activities were designed following Nation's (2001) Depth of Processing theory (for more details see section 6.4) and were given to the Chinese students after reading a passage: 'Note down the good expression', 'Use the right expression', 'Enhance your collocation awareness', and 'Retell the story'. In the first activity, students were directed to notice collocations in chunks. In the second and third activities, students were asked to use the collocations and to complete tasks about the recognised collocations so that they can be retrieved correctly. In the final activity, they were asked to re-tell the story using the recognised collocations.

Jiang (2009) designed a set of awareness-raising activities as a result of her investigation of the use in a corpus and materials of the most frequent six words located in the Chinese Learner English Corpus (CELC; Gui and Yang, 2003). She first compared their uses with the Freiburg-

LOB corpus of British English (FLOB; Hundt, and Siemund, 1998) to gain a better understanding of Chinese learners' collocation knowledge and uses and then checked the usage and coverage of these six words in three sets of teaching materials that are taught to Chinese students. Her self-designed activities were positively evaluated by teachers and students.

Thurstun and Candlin (1998) designed their set of corpus-based activities to introduce the most frequent and significant academic words to both NNS and NS students who were unfamiliar with their uses in an academic context. They argued that the principal reason for using corpus-based materials to teach academic vocabulary is not only to help learners to guess the meaning and use of unknown words from context but also to direct their attention to the central importance of collocational relationships associated with the keywords. They first selected the most frequent 150 academic words from Nations' UWL (1990) and grouped these words into categories according to their rhetorical functions. Six categories were developed: stating the topic of your writing, referring to the research literature, reporting the research of others, expressing your opinions tentatively, explaining the procedure taken in a study, and drawing conclusions.

Following Johns' (1991b) approach, various activities were designed: first, a sample of concordance lines was introduced to learners to notice keywords and answer a set of questions related to the keywords; then another set of activities were given to practice keywords. Finally, students were asked to produce their own written sentences using the keywords. Another set of activities was designed using the problem-solving approach. Two types of gap-filling activities were designed for this purpose: concordance lines in which a single word is missing and a set of concordance lines in which two or three words are missing. Even though their aims were different from Jiang (2009), both studies applied the Depth of Processing theory in their

awareness-raising activities. Jones and Durrant (2010), on the other hand, applied the first process 'noticing' only. It can be noted that Stahl (1985 mentioned in Nation, 2001) Depth of Processing theory is pedagogically useful.

However, few studies have been conducted to design online corpus-based materials for ESP learners (specifically for Computer Science students). Chang and Kuo (2011) designed their online corpus-based materials from 60 research articles from CS to improve their Chinese graduate learners' understanding of the rhetorical moves, move patterns, and specific vocabulary used in research articles. Their online materials focused on raising students' awareness about the information structure and language use of each section of the research articles, from abstract to conclusion. The language features covered were the tenses, modals, and reporting verbs. Other writing resources were incorporated in their website: online dictionaries, collocations builder, and the concordancer. The purpose of these tools was to facilitate NNS students' writing process and writing development, as L2 writing research has revealed that in the process of composing and revising L2 writers may need to deal with lexico-grammatical problems (Shei and Pain, 2000; Chang and Kuo, 2011).

Their findings revealed that 80% of the learners were satisfied with the online materials provided and with the learning tasks given as they were designed to fulfil their discipline specific needs. This finding has been confirmed by Lee and Swales (2006:71) who pointed out that "the closer the participants could come to their discipline-specific written discourses, the more engaged with the texts they became and the more time they were willing to spend on them". Their study demonstrates the value of online EAP coursework in promoting active learning with research-supported materials that are based on real-world language use data. Most importantly, the supportive writing tools provided to the Chinese learners confirm learners' need for raising

awareness of certain linguistic features in their writing and their sources, specifically collocations.

From the previously mentioned studies, it can be summarised that most of the corpus-based studies conducted on teaching materials have focused first on the teaching of EAP academic words, rather than focusing on ESP vocabulary. Thus, it can be concluded that it would be better to first introduce ESP learners to EAP vocabulary and collocations (GAC and GCSC in the CJT; for more details see section 5.4.1) as they are more likely to be familiar with academic vocabulary and then extend their knowledge to discipline-specific vocabulary and collocations (SCSC in the CJT; for more details see section 5.4.1).

This seems in accordance with previous researchers' (Kennedy and Bolitho, 1984; Baker, 1988; Li and Pemberton, 1994) recommendation that ESP students' need for academic vocabulary is greater than their need for discipline- specific vocabulary. However, Kübler and Foucou (2003) recommended teaching both discipline–specific vocabulary as well as academic vocabulary in their teaching of CS vocabulary, since NNS learners may have not been exposed to discipline-specific vocabulary; thus they need to learn both to be able to write competently and idiomatically in CS (for detailed information about these two types of vocabulary, see section 5.2.1). So far, two factors have been thought to be important in selecting which words to be taught first: the problematic words and the most frequent words either from students' reading textbooks or from a corpus data.

6.7 The Awareness-raising Activities Designed in our Study

When designing corpus-based materials and DDL activities, three main factors need to be considered: the corpus, the learning context, and the students' proficiency. Since the main aim of designing corpus-based activities in this thesis is to raise NNS students' awareness of the use of collocations and their patterns in an ESP context, specifically in CS, the RC that was compiled for this study was used for designing the awareness-raising activities (see section 3.5 for detailed information about the RC). NNS students' corpus was used for activities, which include comparison and contrasting.

Even though the size of the RC may not be considered as large as it should be, it would be sufficient for designing awareness-raising activities for CS learners. Aston (1997) and Flowerdew (2001) have recommended working with small corpora for pedagogical purposes, as they are potentially more fully analysable, easier to become familiar with, easier to interpret, and more clearly patterned. Similarly, Tribble and Jones (1990: 71) suggested, "small collections of text (less than 50,000 words) are often best for classroom research as they do not take too long to process".

Regarding the learning context, since CS students in Saudi Arabia have never been exposed to any kind of corpus-based DDL activities, a decision was made in designing raising-awareness activities to employ the soft version, where the teacher designed the tasks according to learners' proficiency levels. These activities are designed for first year NNS CS postgraduate students who will be continuing their MSc degrees by research.

Following the soft version of DDL, the corpus-based materials would be presented in teacher-designed worksheets (Gabrielatos, 2005; Boulton, 2010). The materials are not for self-study, mainly because Saudi students are unfamiliar with computer-based technologies in language teaching and, therefore, concordancing.

In what follows, a few concordance lines (6 to 15 lines) are provided to illustrate each activity (Jones and Durrant, 2010; Thurstun and Candlin, 1998; Barlow and Burdine, 2006). Concordance examples in these activities take the form of cut-off sentences, which can help students focus on keywords and their co-occurring or adjacent words and which may make the target collocation patterns and use more salient. An example of cut-off concordances for the collocation *source code* is presented in Figure (6-3).

Figure 6-3: An example of cut-off concordance lines for the collocation *source code***.**

1- …knowledge of the source ***code*** or better user…
2- …two million lines of source ***code***, and evaluated the…
3- …We have provided C++ source ***code***, but it is straightforward…
4 …we provide the source ***code*** for Computing the proposed…

6.7.1 Criteria for Selecting Collocations for Awareness-raising Activities

Three criteria were applied for selecting collocations for awareness-raising activities. First, problematicity of collocations for NNS students was considered the main criterion for selection (Woolard, 2000; Gaskell and Cobb, 2004). The 24 shared N collocations were all examined to

discover the problematic collocations for NNS students (see Appendix F for the full list of the 24 NNS N collocations). Collocations were considered problematic if they were used differently from the RC (over- or underused) by NNS. For example, if an N collocation was over/underused by NNS students, it was considered a problematic collocation. When these criteria applied, 10 of the 24 shared N collocations were found to be problematic. Eight N collocations were overused and two were underused. These collocations fall under the three categories found in the CJT: GAC, GCSC, and SCSC (see CJT in Appendix G for more information). They are shown in Table 6-1 below.

Second, SCSCs and other collocations that fall under two types of collocations were excluded (see the results of the CJT in Appendix M). SCSCs were excluded for two main reasons. It has been confirmed that ESP students need to learn about academic vocabulary as well as their specific-discipline vocabulary. Thus, the focus on teaching GAC and GCSC would be useful. Second, it was confirmed by one of the CS experts that focusing on teaching GAC and GCSC would be more interesting and useful to CS postgraduate students. As can be noticed from Table 6-1, only two N collocations were SCSC (*layer application* and *class method*) and thus were excluded. Two other N collocations (*network traffic* and *design system*) that were categorised differently by CS experts were excluded. Only six N collocations were left to be used in the awareness-raising activities.

Third, only two or more patterns collocations were included in the awareness-raising activities. Applying this criterion, there were two N collocations (*code following* and *resources available*) that used similar number of patterns by both NNS and RC, thus they were excluded. Only four N

collocations remained (*code source*, *data type*, *data access*, and *data user*) that follow the three criteria; see Table 6-1 below.

Table 6-1: Three criteria applied in selecting N collocations for awareness-raising activities.

No.	**NNS N collocations**	**Significant Over/under use**	**CJT Result**	**No. of patterns in RC**	**No.of patterns in NNS**	**Three criteria applied**
1.	code following	overuse	GCSC	1	1	×
2.	code source	underuse	GCSC	2	1	√
3.	data type	overuse	GCSC	2	3	√
4.	data access	overuse	GAC	3	2	√
5.	data user	overuse	GAC	3	5	√
6.	resources available	overuse	GAC	2	2	×
7.	design system	overuse	GCSC/GAC			×
8.	layer application	overuse	SCSC			×
9.	method class	underuse	SCSC			×
10.	network traffic	overuse	GCSC/SCSC			×

6.7.2 Main Types of Awareness-raising Activities that were Designed

Three awareness-raising activities were designed following Dave and Jane Willis' consciousness-raising (CR) taxonomy (cited in Lewis, 1997: 53), all with focus on the first four steps of their CR taxonomy (for more details about the CR taxonomy see section 6.5).

In the first step, where students are required to search to identify a pattern or usage, NNS CS students would be asked to search for collocations of a specific keyword in a set of concordance lines from the RC. In the second step, in which students are asked to clarify similarities and differences of the recognised patterns or usage, NNS CS students' attention would be directed to the collocation patterns that occurred in the RC so that they can notice similarities and differences among the recognised patterns. In the third step, in which students are asked to check generalisations about what they identified against other data, NNS CS students would be asked to compare collocation patterns recognised in the RC with NNS students' use. In this activity, students would be provided with a set of concordance lines from NNS students' use of the same collocation, so that they will be able to find similarities and differences in their uses as well as making generalisations.

No attempts have been made to design any production activities since students might presumably produce discipline-specific sentences, which would probably be difficult for the language teacher to comment on in terms of the content. Three types of awareness-raising activities were designed for each of the aforementioned steps; they are described fully in the following three sections.

6.7.2.1. Noticing Collocation

In this type of activity, each of the four selected collocations was presented individually, each with a set of concordance lines from the RC displaying the collocation. Adapting Jones and Durrant's (2010) approach of raising students' awareness about a certain word, in which they asked students to look at the right and the left context of each word, students would be asked to recognise the collocate of the keyword highlighted in the concordance lines and then answered the set of questions about the use of the collocation. For example, the keyword *data* was highlighted in the following concordances to be clearly searched for its collocates, as shown in Figure (6-4).

Students would be expected to be able to recognise the collocation *data type* from the given concordance lines as they were all clearly displayed. Students would be asked to focus on the word *data* and to look for their left and right noun-phrase context so that they would not be distracted by other words. Moreover, after they noticed the collocation *data type*, they would be able to recognise the various ways to expand this collocation by adding prepositions such as 'data with type', and 'type of (the) data'. Thus, students would be encouraged to recognise the collocation as well as its various extended versions

Figure 6-4: An example of a collocation noticing activity.

The following exercise will help you notice the kinds of words and phrases that are often found around 'data' (either on its left or on its right) in Computer Science writing. Spend some time analysing the concordance-lines of this word and answer the following questions:

A. The word 'data' is a noun. Look at the words to the right of 'data'. Which words are more frequently used?
B. Can you identify the part of speech of these words?

C. Which words and phrases go to the left of 'data'? Which go to its right?

1 …and a column for each ***data***type used at least once…

2…has a parameter of the ***data*** type, and it equals…

3 …classes constitutes an abstract ***data***type encapsulating methods…

4…and the old ***data*** with type A is allocated…

5…According to the type of ***data*** available for training...

6…the particular type of ***data*** sought…

7…where the type of the ***data*** the session on the main stack…

<extracted from the reference corpus>

6.7.2.2. Noticing and Identifying Patterns of a Collocation

Pattern recognition activities were designed following the principle suggested by Gabrielatos (2005), stating that intense language exposure can help learners formulate intuitions about language use. That is, focused language exposure through pattern-recognition activities can be useful for language learners in countries where the target language is not widely spoken(e.g. Saudi Arabia)because they do not have many opportunities to be exposed to sufficient real language use in context, which is essential for developing the ability to recognise language patterns. Barlow and Burdine (2006:4) refer to these activities as "pattern recognition" and "concordances-based research".

After students have recognised a collocation, they would be asked to find the patterns used in the concordance lines and complete tables with frequent patterns used in the reference corpus. Students would notice various patterns used by CS experts, thus, they could provide a written record of different ways of presenting the same collocation.

Figure 6-5: An example of the activity 'noticing and identifying patterns of the collocation '*data type''*.

Look at the concordance lines of the first activity (noticing collocation) and try to answer the following questions:

A. How many patterns did you find for 'data type'?
B. In the following table, write down the patterns and how many times you found each pattern.

Patterns for 'data type'	Number of occurrences (frequency)

After learners have checked the concordance lines and identified the collocation, it would be easy for them to classify the different patterns used by expert writers. Thus, they would be asked to categorise the patterns of the collocation and to count their number of occurrences in this task.

6.7.2.3. Comparing and Contrasting Patterns between the NNS Students' corpus and the Reference Corpus

Students would be asked to compare collocation patterns between NNS concordance lines and the patterns already identified from the RC in the previous activity. Through their comparison of various patterns used by NNS students' corpus and the RC they should discover the frequent and infrequent patterns used.

Figure 6-6: An example of the activity ‘comparing collocation patterns between the NNS student corpus and the reference corpus’.

Look at the concordance lines of set (A) that are taken from non-native speakers students’ corpus. Spend some time analysing the words and phrases that go together with ‘data type’. Then answer the questions below.

Set (A)

1…structure can help create a ***data*** type definition for documents…

2…system, double type ***data*** values sent by the…

3…submitting the wrong type of ***data*** into a document…

4…the multiplexing of type of data. The different data…

5…on the Ethernet type of ***data*** transmission**...**

6…is the only type of ***data*** traffic used in this…

7 …case where this type of ***data*** is used for just two…

8…are a type of unwanted ***data*** available on web pages…

9 …another type of unwanted ***data*** that need to be removed…

<extracted from NNS corpus>

A. How many patterns are used by non-native speakers’ students for ‘data type’?
B. Do the non-native speakers’ students use any of the same words and phrases you found earlier, when you looked at ‘data type’ in the previous activity?
C. In the following table, write down the patterns and how many times you found each pattern.

Patterns for ‘data type’	Number of occurrences (frequency)
Pattern 1	
Pattern 2	
Pattern 3	

D. Compare between the patterns you identified for non-native speakers with those found in the reference corpus in the previous activity.

As students had already been exposed to the collocation ‘data type’ and its patterns, they would be introduced here to examples from the NNS student corpus so that they can compare

similarities and differences in the use of the collocation patterns between the corpora. Expert writers' use of collocation patterns was considered the most crucial and thus was first introduced. In this task, students would be directed to pay attention to the differences and similarities between NNS students' use and experts' use of patterns to make their decisions about which patterns were most frequently used and which patterns were accepted by experts. Similar sets of activities were designed for the other three collocations: They are all included in Appendix N.

6.8 Conclusion

This Chapter presented the main issues related to teaching collocations, the corpus-based DDL approach and collocation activities. Next, it described in detail how the corpus-based awareness-raising activities for Saudi postgraduate students have been designed. Activities were designed in the soft version of DDL using the RC as the main source of examples for the first two awareness-raising activities, noticing collocation, and noticing and identifying patterns of collocation. The NNS students' corpus was used only for comparison purposes in the final activity: comparing and contrasting patterns between NNS students' corpora and the RC.

Chapter 7 Conclusion

7.1 Scope of the Present Thesis

Research presented in this thesis focused on the use of academic collocations and patterns in the writing of CS postgraduate students. Academic collocations, which have been widely investigated in student corpora in the EAP context, have been ignored in the ESP context. The present thesis aimed to fill this gap in the ESP context through exploring the use of academic collocations and patterns in the writing of CS postgraduate students and comparing this with experts' writing. In addition, a sample of awareness-raising activities was designed for raising NNS students' awareness of the use and patterns of some problematic academic collocations. Thus, three main studies have been described in this thesis: the use of academic collocations by non-expert CS postgraduate students (presented in Chapter 4), factors underlying over/underuse of collocations (presented in Chapter 5), and awareness-raising activities (presented in Chapter 6).

7.2 Major findings

The major findings answering the seven main research questions can be summarised under the three studies covered in this thesis:

7.2.1 Study 1: The use of academic collocations by non-expert CS postgraduate students

Three research questions were answered in this study:

RQ1. What are the most common academic collocations used by Computer Science students in their MSc dissertations?

After locating the most frequent members of the 100 most frequent AWL families in the students' corpora, a short list of the most frequent members of the 88 word families (see Table 4-1for details) for each student corpora was inserted into *ConGram* to locate their collocations. Collocations were located applying MI of 3, t.score of 2, and span of three words from the left and the right of the node words. The results reveal that both NNS and NS students tend to use noun collocations more than verb collocations (as displayed in Table 4-5). This finding seems to be in agreement with Halliday (1966) and Coxhead and Byrd (2007) who claim that science discourse is characterised by the use of nominalisations and thus can be described as more noun centric than verb centric. Surprisingly, both NNS and NS use only few verb collocations significantly.

RQ2: To what extent do native and non-native postgraduate CS students make greater or less use of academic collocations in their writing in comparison with the reference corpus?

Both NNS and NS students were found to overuse academic noun collocations in their writing of dissertations, compared to experts' use, when the 100 most frequent noun and verb collocations from each students' corpora were tested for their significance. N collocations were similarly overused by both NNS and NS students. NNS significantly overused 52% of the 100 most

frequent noun collocations, while NS significantly overused 78% of the 100 most frequent noun collocations. This result contrasts previous research findings that confirmed NNS usually overuse a limited set of collocations and do not use collocations like NS (Foster, 2001; Granger, 1998; Howarth, 1998a; Durrant and Schmitt, 2009). Thus, it can be inferred that NNS tend not to find difficulty in using certain noun collocations in their ESP context.

However, this result could be explained on a number of grounds other than language ability. One is genre variations: that is, the writing style in dissertations differs from that of the writing in research articles. Secondly, the lower level of lexical variation could be explained by the larger number of words in each text in the sample. Thus, lexical variation, including number of different collocations used, is likely to be lower in dissertations than in research articles. Hence, the chance of repeating the same collocations in dissertations would be more likely than in research articles.

RQ3: To what extent do native and non-native postgraduate CS students differ in their use of the shared set of academic noun collocations?

When the 30 shared academic noun collocations used in both NNS and NS students' corpora were compared a great number of these collocations were overused by both groups of students, in comparison with the expert corpus, while only few collocations were used in significantly different frequencies. A number of factors were potentially thought to explain the variations found in the data answering RQs 1-3: specific collocation patterns used, genre, topic and sub-discipline specificity.

7.2.2 Study 2: Factors underlying the non-experts' over/underuse of noun collocations

The aforementioned factors were investigated in detail in the second study (see Chapter 5) where patterns of the 30 shared noun collocations used by both NS and NNS students were first identified and then CS experts were interviewed and asked to complete the categorisation judgement task. Three research questions were answered in this study:

RQ4.To what extent can the relative collocation pattern frequency between the NNS and NS corpora, on the one hand, and the RC corpus on the other, explain collocations' over/underuse in the NNS and NS corpora?

Overall, the over/underused patterns hypothesis regarding the number of patterns of each collocation used in the different corpora was only supported in the comparison between NS and the RC where t (23) =-1.683, p=0.05.There was no significant difference between NNS and the RC as the paired test gave t *(16)* =0.169, p>.05. However, some individual collocations were overused by NNS or NS students because of their use of more patterns than the patterns used in the expert writers' corpus.

RQ5. To what extent do the shared collocations differ in their patterns?

After identifying, the erroneous occurrences of the 30 shared noun collocations and re-testing their significance, six collocations were excluded, as they were non-significant (see section 5.3.1.3 for more details). The remaining 24 shared noun collocations were identified for their patterns. They were used with different patterns among the corpora. Variation in the patterns

used could explain partially the over/underuse of some of the collocations. Both NNS and NS students used different range of patterns in their use of collocations as they compared to the reference corpus. The most recognisable patterns were N+N, Adj+N, N+PRP+N, and N+ADJ+N: these patterns matched those classified by Hunston and Francis (2000).

RQ6a. What are the factors behind students' over/underuse of academic collocations according to CS experts' views?

RQ6b. What are the CS experts' views about the reasons underlying the use of specific collocation patterns in the data?

A number of factors could explain the over/underuse of the 24 shared noun collocations according to the CS experts: genre, topic, discipline-specificity, and writers' personal style. Computer Scientists confirmed the effect of genre on the overuse of some collocations. They confirmed that some N collocations were overused by students in their writing of dissertations, as students need to write in detail. For example, the collocation *following code* was clearly confirmed to be overused by students rather than by experts since students are required to develop their software or applications and thus need to use a large number of codes, while CS experts tend to mention only the development of applications without the need to reference their codes. Topic was also found to play a role in the overuse of some collocations. For example, the collocations *method class* was noted by one of the CS experts to be related to topics specific to Java.

In addition, their categorisation judgment task revealed that there are some discipline-specific collocations, e.g., *layer application* and *network traffic*, which were categorised as collocations

specific to the IS sub-discipline of CS, while other N collocations were found to be more general in their academic use across other disciplines e.g. *data access*.

7.2.3 Study 3: Academic Collocations Awareness-raising activities

RQ7: What kind of teaching materials are needed to raise NNS students' awareness of the use of academic collocations?

Reviewing the available literature on corpus-based activities (Tribble,1990; Tribble and Jones, 1990; Johns,1991a,1991b; Hyland,1998, 2000; Yoon and Hirvela, 2004; Gabrielatos, 2005; Jones and Durrant, 2010) and identifying the problematic over/underused collocations for NNS students from the previous two studies (Study1 and 2), a sample of three awareness-raising activities was designed to be applied with NNS students: noticing collocation, noticing and identifying patterns of a collocation, and comparing and contrasting patterns between NNS students corpus and the RC. These activities were designed with an aim to be applied in future research in order to raise CS postgraduate NNS students' awareness about the use of some problematic N collocations and their most frequent patterns used.

7.3 Implications

This section will discuss implications of the results presented above from two standpoints: linguistic theory and language teaching.

7.3.1 Theoretical Implications

It has been found in the literature that NNS tend not to use collocations like NS and that their use is limited to a certain set of collocations(Foster, 2001; Granger, 1998; Howarth, 1998a; Durrant and Schmitt, 2009).This is true in an EAP context, but not in an ESP context. The result in this thesis contradicts the previous research findings. NNS Computer Science students were found to overuse N collocations (compared with expert writers), as did NS students, and underused few N collocations. However, Siyanova and Schmitt (2008) also found that their NNS, who were advanced students of English, were similar in their use of adj+N collocations to their NS. Their NNS students in an EAP context seem to be similar to their NS students in their use of N collocations. These conflicts in findings mean that further research is needed to investigate the use of academic collocations by different levels of NNS students and both non-expert and expert NS.

Moreover, the overuse of academic N collocations by NNS and NS students when they were compared to experts' use could be explained by genre variations and discipline-specificity. This finding is in line with Hyland and Tse's (2007) claims that some academic words have discipline-specific collocations and thus they should not be all included in Coxhead's (2000) AWL. Various patterns identified for the 24 shared N collocations in the students' corpora (see Chapter 5) confirmed some of the idiom principle features cited by Sinclair (1991). Different internal insertions were added and different word order was found in some collocations.

7.3.2 Pedagogical Implications

It has been confirmed that teaching collocations is essential in EAP context (Hill, 2000; Lewis, 2000a; Conzett, 2000) as they are considered an important component of language knowledge for learners' oral and written production. Using corpus-based research (e.g., Durrant, 2009; Gardner and Davies, 2007; Shin and Nation, 2008) can, thus, inform collocational teaching practice through extracting the most frequent collocations and those comprising highly frequent words. It can also be applied in raising learners' awareness of collocations and their patterns. The three samples of awareness-raising collocations activities in Chapter 6 were designed to implement the DDL approach in raising NNS learners' awareness of collocations use and patterns.

The present section discussed theoretical and pedagogical implications of the findings presented in this thesis. Despite these implications, the studies presented are limited in a number of ways. These limitations will be considered in detail in the next section.

7.4 Limitations

The research presented in this thesis is limited in a number of ways. First, it only looked at the most frequent lexical collocations (N collocations and V collocations) and ignored grammatical collocations. Second, the students' and the reference corpora were limited in their coverage to only three sub-disciplines of CS (AI, SE, and IS). This limitation could not be avoided since the available dissertations were all included in one of these three sub-disciplines. It would be better if more of the CS sub-disciplines were covered. If this had been the case, the findings of this thesis could have been more generalisable.

On the other hand, the size of the students' and the reference corpora are probably not large enough to allow for the examination of all academic collocations in CS. This limitation could not be avoided since there were not enough NS dissertations available. The NNS corpus was compiled so that it would be of a size equal to the NS corpus. The RC size was limited in number (600, 269 words) due to time restrictions and the need to POS tagging and to check them in all corpora.

Another limitation in relation to the RC is that it contained only research articles. No concern was given to include textbooks since their discourse is different from MSc dissertations. Carter (1998) and Römer (2004) found that the distribution and patterns of language features between reference corproa and textbooks are different and thus they will contain different phraseology from the ones used in academic writing. The research articles included in the RC appeared only in the years 2011 and 2012. It would be better if the corpus had consisted of articles published in more than two years.

Third, not all of the 24 shared N collocations among corpora were analysed in detail (see Chapter 5). Two main reasons hindered the analysis. First, few CS experts agreed to participate in the study. Second, some N collocations were more explicable than others were. Some N collocations were relevant to the selected CS sub-disciplines while others were more specific in their uses.

7.5 Suggestions for Future Research

Based on the findings and limitations discussed above, three lines might be suggested for future research.

First, it would be useful in the future to carry out a study where some of the CS postgraduate NNS and NS were interviewed. Thus, their views about the over/underuse of N collocations could be investigated and possibly more learnt about the underlying factors.

Second, it would be useful to investigate students' use of collocations at different levels of proficiency, following Laufer and Waldman's (2011) procedure. For example, a comparison between undergraduate and postgraduate students' use of academic collocations could be carried out to validate our results on non-expert postgraduate overuse of academic N collocations compared with experts.

The final direction for future research would be to trial the sample of awareness-raising activities that I designed with NNS postgraduate students to test the effectiveness of these activities in raising NNS awareness of the use of some problematic N collocations and obtain opinions about their value and interest.

References

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, 235-247.

Allan, R. (2006). Data-driven learning and vocabulary: Investigating the use of concordances with advanced learners of English. *Centre for Language and Communication Studies, Occasional Paper*, 66. Dublin: Trinity College Dublin.

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

Anderson, R.C., &Freebody, P. (1981). Vocabulary knowledge. In J.T. Guthrie (Ed.), *Comprehension and teaching: Research Reviews* (pp.77-117). Newark: International Reading Association.

Arnaud, P.J.L. (1984). The lexical richness of L2 written production and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley and D.K. Stevenson (Eds.): *Practice and problems in language testing*. University of Essex occasional papers, 29, 14-28.

Aston, G. (1995). Corpora in language pedagogy: matching theory and practice. In G. Cook and B. Seidlhofer (Eds.), *Principles and Practice in Applied Linguistics* (pp.257-270). Oxford: Oxford University Press.

Aston, G. (1997). Enriching the learning environment: corpora in ELT. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and Language Corpora* (51-64). London: Longman.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria, *Literary and Linguistic Computing*, 7, 1-16.

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

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4, 91-105.

Barfield, A. & Gyllstad, H. (2009). *Researching Collocations in another Language: Multiple Interpretations*, Basingstoke: Palgrave Macmillan.

Barlow, M. & Burdine, S. (2006). *Phrasal verbs: American English*. Houston: Athelstan (Corpus LAB).

Barnbrook, G., Oliver, M., & Krishnamurthy, R. (2013). *Collocations: applications and implications*. Palgrave Macmillan: New York.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Narr.

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6, 253-279.

Becka, J.V. (1972). The lexical composition of specialized texts and its quantitative aspect, *Prague Studies in Mathematical Linguistics*, 4, 47-64.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI Combinatory Dictionary of English: A guide to word combinations (revised edition)*. Amsterdam: John Benjamins.

Bereiter, C., & Scardmalia, M. (1987). *The Psychology of Written Composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bernardini, S. (2001). Corpora in the classroom: An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp.15-36). Amsterdam: John Benjamins.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann and G. Marko (Eds.) *Teaching and learning by doing corpus analysis.* Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000 (pp.165-182). Amsterdam: Rodopi.

Bernardini, S. (2004). Corpora in the classroom: an overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp.15-36). Amsterdam: John Benjamins.

Bhatia, V.K. (1993). *Analysing genre: Language use in professional settings*. London, UK: Longman.

Biber, D. (1988). Variation *across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design, *Literary and Linguistic Computing*, 8 (4), 243-257.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181-190.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3), 371-405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press.

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

Boers, F., Demecheleer, M., Coxhead, A., and Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18*(1), 54-74.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, H. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research,* 10, 245-261.

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

Boulton, A. (2008). Looking for empirical evidence of data-driven learning at lower levels. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus Linguistics, Computer Tools, and Applications – State of the Art* (pp.581-598). Frankfurt: Peter Lang.

Boulton, A. (2009). Data-driven learning: reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics,* 35(1), 81-106.

Boulton, A. (2010). Data-Driven Learning: Taking the Computer Out of the Equation. *Language Learning*, 60(3), 534-572.

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet. *Corpus-informed Research and Learning in ESP: Issues and Applications* (pp.261-292). John Benjamins.

Boulton, A., Carter-Thomas, S., &Rowley-Jolivet, E. (2012). Issues in corpus-informed research and learning in ESP. In A. Boulton, S. Carter-Thomas and E. Rowley-Jolivet (Eds.). *Corpus-informed Research and Learning in ESP: Issues and Applications* (pp.1-16). John Benjamins.

Bunton, D. (2002). Generic Moves in PhD Thesis Introduction. In. J. Flowerdew (Ed.), *Academic Discourse* (pp.57-75). Harlow, UK: Longman.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. University of Sydney Papers in *TESOL*, 5, 31-64.

Cacchiani, S. (1984). Lexico-functional categories and complex collocations: The case of Intensifiers. In U. Romer and R., Schulze. *Exploring the Lexis-Grammar Interface*. (pp.229-246). Amsterdam: John Benjamin.

Campion, M., & Elley, W. (1971). *An Academic Vocabulary List*. Wellington: New Zealand Council for Educational Research.

Carter, R. (1987). *Vocabulary: Applied Linguistic Perspectives.* London: Unwin Hyman.

Carter, R. 1998. Orders of reality: CANCODe, communication, and culture. *ELT journal* 52(1), 43-56.

Carter, R., & McCarthy, M.J. (1995). Grammar and the Spoken Language, *Applied Linguistics*, 16(2), 41-58.

Carter, R., & McCarthy, M.J. (2001). Size is not everything: spoken English, corpus and the classroom, *TESOL Quarterly*, 35(2), 337-340.

Carter, R., & McCarthy, M.J. (2006). *Cambridge Grammar of English.* Cambridge: Cambridge University Press.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda and J. Santana (Eds.), *Corpora in the Foreign Language Classroom* (pp.3-16). Amsterdam: Rodopi.

Chang, CH. & Kuo, CH. (2011). A corpus-based approach to online materials development for writing research articles. *English for Specific Purpose*, 30, 222-234.

Chen, Q., & Ge, C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes*, 26, 502-514.

Cheng, W., Greaves, C., Sinclair, J.M., & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*, 30(2), 236-252.

Cheng, W., Greaves, C. and Warren, M. (2006). From n-gram to skip-gram to ConcGram. *International Journal of Corpus Linguistics*, 11, 411-433.

Chi Man-Lai, A., Wong Pui-Yiu, K. & Wong Chau-ping, M. (1994). Collocational problems amongst ESL learners: A corpus-based study. In L. Flowerdew and A.K.K. Tong, *Entering Text* (pp.157-165). Hong Kong: Language Centre, Hong Kong University of Science and Technology. Retrieved 14 May 2011, from http://repository.ust.hk/ir/bitstream/1783.1-1088/2/entertext04.pdf.

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

Chung, T. (2009). The newspaper word list: A specialised vocabulary for reading newspapers. *JALT Journal*, 31(2), 159-182.

Chung, T.M., & Nation, I.S.P. (2004). Identifying technical vocabulary. *System*, 32, 251-263.

Clear, J. (1993). From Firth principles: Computational tools for the study of collocation. In M. Baker, G. Francis and E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp.271-292). Amsterdam: John Benjamins.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *Canadian Modern Language Review*, 59, 393-424.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards and B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp.15–38). Amsterdam: John Benjamins.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data analysis: Complementary research strategies.* Thousand Oaks, CA: Sage.

Cohen, A. D., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1988). Reading English for specialized purposes: Discourse analysis and the use of student informants. In P. L. Carrell, J. Devine and D. Eskey (Eds.), *Interactive Approaches to Second Language Reading* (pp.152-167). Cambridge: Cambridge University Press.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp.70-86). England: Language Teaching Publications.

Cooper, A., & Bikowski, D. (2007). Writing at the graduate level: What tasks do professors actually require? *Journal of English for Academic Purposes*, 6(3), 206-221.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23, 397-423.

Cowan, J.R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, 8, 389-400.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics, 2*(3), 223-235.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly*, 34(2), 213-238.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16, 129-147.

Coxhead, A. & Byrd, P. (2012). Collocations and Academic Word List: The strong, the weak and the lonely. In I. Moskowich and B. Crespo (Eds.) *Encoding the Past, Decoding the Future: Corpora in the 21st Century* (pp.1-20). Cambridge: Cambridge Scholars Publishing.

Coxhead, A., & Hirsh, D. (2007). A pilot science word list for EAP. *Revue Francaise de Linguistique Applique´,* XII(2) 65-78.

Coxhead, A. & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew and M. Peacock (Eds.). *Research Perspectives on English for Academic Purposes* (pp.252-267). Cambridge: Cambridge University Press.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, 16(2), 35-52.

Cresswell, A. (2007). Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda and J. Santana (Eds.), *Corpora in the Foreign Language Classroom* (pp.267-287). Amsterdam: Rodopi.

Dai, Z., & Ding, Y. (2010). Effectiveness of text memorization in EFL learning of Chinese students. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp.71−87). New York: Continuum.

Davies, M. (2004). BYU-BNC: The British National Corpus. *Available at* http://corpus.byu.edu/bnc

Davies, M. (2012). Corpus of Contemporary American English (1990–2012). *Brigham Young University, USA. [Online]*. *Available at:* http://corpus.byu.edu/coca.

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

Dong, Y.R. (1998). Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and advisors from two U.S. institutions. *English for Specific Purposes*, 17, 269-390.

Dornyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies.* Oxford: Oxford University Press.

Dresher, R. (1934). Training in mathematics vocabulary. *Educational Research Bulletin*, 13, 201-204.

Dudley-Evans, T. & St. John, M.J. (1998). *Developments in ESP: A Multi-Disciplinary Approach*. Cambridge: Cambridge University Press.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2), 125-155.

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

Durrant, P. & Schmitt, and N. (2009). To what extent do native and non-native writers make use of collocations*? IRAL*, 47, 157-177.

Ellis, N. C. (1997). Vocabulary acquisition: Word structure, collocation, word-class, and meaning. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.122-139). Cambridge: Cambridge University Press.

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

Ellis, R. (1992). Grammar teaching-practice or consciousness-raising? In R. Ellis (Ed.) *Second Language Acquisition and Second Language Pedagogy* (pp.232-241). Clevedon: Multilingual Matters.

Evert, S. (2004). Computational approaches to collocations. Retrieved 14 October 2011, from www.collocations.de.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (extended manuscript of Chapter 58), (pp.1-53). Berlin: Mouton de Gruyter.

Evison, J. (2010). What are the basics of analysing a corpus? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.122-135). London: Routledge.

Farooqui, A. (2010). *Teaching Technical Vocabulary in ESP Courses: Students and ESP/Subject Teachers' Perspectives*. Unpublished MA dissertation, University of Essex.

Farrell, P. (1990). A lexical analysis of the English of electronics and a study of semi-technical vocabulary (ERIC Document Reproduction Service ED 332 551). *CLCS Occasional Paper* No. 25. Dublin: Trinity College.

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness,* 17(1), 25-43.

Firth, J. R. (1957/1968). A synopsis of linguistic theory 1930-55. In F. R. Palmer (Ed.), *Selected Papers of J.R. Firth, 1952-59* (pp.168-205). Harlow: Longmans.

Fitschen, A., & Gupta, P. (2008). Lemmatising and morphological tagging. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp.552-564). Berlin: Mouton de Gruyter.

Flowerdew, J. (2002). Genre in the classroom: A linguistic approach. In A. M. Johns (Ed.). *Genre in the Classroom: Multiple Perspectives* (pp.91-104). London, UK: Lawrence Erlbaum associates.

Flowerdew, J., & Peacock, M. (2001). *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University press.

Flowerdew, L. (1998). Corpus Linguistics Techniques Applied to Textlingistics. System, 26,541-552.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, and R. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp.363-379). Amsterdam: John Benjamins.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. Upton (Eds.), *Discourse in the professions* (pp. 11–33). Amsterdam: Benjamins.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus based methodologies. *English for Specific Purposes, 24*, 321–332.

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

Francis, G., Hunston, S. & Manning, E. (1996). *Collins Cobuild Grammar Patterns 1: Verbs*. London: Harper Collins.

Francis, G., Hunston, S. & Manning, E. (1998). *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: Harper Collins.

Francis, W. N., & Kucera, H. (1979). *A Standard Corpus of Present-day Edited American English for use with Digital Computers*. Providence, RI: Department of Linguistics, Brown University.

Gabrielatos, C. (2005). Corpora and language teaching: just a fling or wedding bells? *TESL-EJ*, 8(4), 1-35.

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

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301-319.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning.* Amsterdam: Benjamins.

Gavioli, L., and Aston, G. (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal*, 55(3), 238-246.

Geisler, C. (1994). *Academic Literacy and the Nature of Expertise: Reading, Writing and Knowing in Academic Philosophy.* Hillsdale, NJ: Lawrence Erlbum Associates.

Ghadessy, P. (1979). Frequency counts, word lists, material preparation: A new approach. *English Teaching Forum*, 17, 24-27.

Gledhill, C. (2000a). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19, 115-135.

Gledhill, C. (2000b). *Collocations in Science Writing* (Vol. 22). Gunter NarrVerlag.

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

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

Granger, S. (2004). Computer learner corpus research: current status and future prospects. *Language and Computers*, 52(1), 123-145.

Granger, S and Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on Computer* (pp.199-209). London: Longman.

Greaves, C. (2005). *Introduction to ConcGram*. Tuscan Word International Workshop. Certosa di Pontignano, Tuscany, Italy. 25-29 June 2005.

Greenbaum, S. (1988). Some verb-intensifier collocations in American and British English. In S. Greenbaum (Ed.), *Good English and the Grammarian* (pp.113-124). London: Longman.

Gries, Stefan Th. (2010). Useful statistics for corpus linguistics. In A. Sánchez and M. Almela (Eds.), *A Mosaic of Corpus Linguistics: Selected Approaches* (pp.269-291). Frankfurt am Main: Peter Lang.

Groom, N. (2005). Pattern and meaning across genres and disciplines: an exploratory study. *Journal of English for Academic Purposes* 4, 257-277.

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

Gyllstad, H. (2007). *Testing English Collocations: Developing Receptive Tests for use with Advanced Swedish Learners*. Lund University.

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

Halliday, M.A.K. (1998). Things and relations: regrammaticising experience as technical knowledge. In J.R. Martin and R. Veel (Eds.), *Reading Science: Critical and Functional Perspectives on Discourses of Science* (pp.185–235). London: Routledge.

Halliday, M.A K., and Martin J. (1993). *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Halliday, M.A.K. and Sinclair, J.(1966).Lexis as a Linguistic Level, in C.E. Bazell, J.C. Catford, M. A.K. Halliday and R.H. Robins(Eds), *In Memory of J. R.Firth (PP.148-162).*London: Longman.

Handl, S. (2009). Towards collocational webs for presenting collocations in learners' dictionaries. In A. Barfield and H. Gyllstad (Eds.), *Researching Collocations in another Language: Multiple Interpretations* (pp.69-85). Basingstoke: Palgrave Macmillan.

Harwood, N. (2005). "I hoped to counteract the memory problem, but I made no impact whatsoever": Discussing methods in Computing Science using I. *English for Specific Purposes*, 24, 243-267.

Harwood, N. (2006). (In) appropriate personal pronoun use in political science: a qualitative study and a proposed heuristic for future research. *Written communication*, 23, 424-450.

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

Haswell, R. (1991). *Gaining Ground in College Writing: Tales of Development and Interpretation.* Dallas: Southern Methodist University Press.

Hawkins, E. (1984). *Awareness of Language: An Introduction*. Cambridge: Cambridge University Press.

Herniksen, B &Stæhr, L.S. (2009). Commentary of Part IV: Processes in the Development of L2 Collocational Knowledge – A Challenge for Language Learners, Researchers and Teachers. In A. Barfield and H. Gyllstad (Eds.) *Researching Collocations in Another Language: Multiple Interpretations* (pp.224-231). Basingstoke: Palgrave Macmillan.

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp.47-69). England: Language Teaching Publications.

Hill, J., Lewis, M., & Lewis, M. (2000). Strategies, activities, and exercises. In L. Lewis (Ed.) *Teaching Collocation: Further Developments in the Lexical Approach* (pp.88-116). England: Language Teaching Publications.

Hill, J., & Morgan, L. (1997). *LTP Dictionary of Selected Collocations*. Hove: LTP.

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

Hoey, M. (2003). Why grammar is beyond belief. In J.P. Noppen, C. Den and I. Tudor (Eds.), *Beyond: New Perspectives in Language, Literature and ELT* (pp.183-196). Ghent: Academic Press.

Hoey, M. (2004). The Textual Priming of Lexis. In S. Bernardinin, G. Aston and D. Stewart (Eds.), *Corpora and Language Learners* (pp.21-41). Amsterdam: John Benjamins.

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

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

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

Howarth, A.P. (1998b). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp.161-186). Oxford: Oxford University Press.

Hudson, R. (1984). *Word Grammar*. Oxford: Blackwell.

Hughes, G. (1997). Developing a computing infrastructure for corpus-based teaching. In A.Wichmann, S.Fligelstone, T. McEnery and G.Knowles (Eds.), *Teaching and Language Corpora* (pp.292-307). New York: Addison Wesley Longman.

Hundt, M., Sand, S., and Siemund, R. (1998).*Manual of Information to Accompany the Freiburg-LOB Corpus of British English (FLOB)*. Germany: English seminar, Albert-Ludwigs-Universitat Freiburg. Available at: http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM.

Hunston, S. (2002a). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2002 b). Pattern grammar, language teaching, and linguistic variation: applications of a corpus-driven grammar. In R. Reppen, M. Susan, Fitzmaurice and D. Biber (Eds.) *Using Corpora to Explore Linguistic Variation) pp.167*-186). Amsterdam: John Benjamins.

Hunston, S. (2008).Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, 13, 271-295.

Hunston, S. (2010). How can a corpus be used to explore patterns? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.152-166). London: Routledge.

Hunston, S. & Francis, G. (1996). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Hunston, S. & Francis, G. (1998). Verbs observed: a corpus-driven pedagogic grammar. *Applied Linguistics*, 19 (1), 45-72.

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

Hutchinson, T., & Waters, A. (1987). *English for Specific Purpose: A Learning-centred Approach.* Cambridge: Cambridge University Press.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

Hyland, K. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes*, 21(4), 385-395.

Hyland, K. (2003). *Second Language Writing.* Cambridge: Cambridge University Press.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor, MI: The University of Michigan Press.

Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. London: Routledge.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate

Writing. *International Journal of Applied Linguistics*, 18, 1, 41-62.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2), 235-253.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly,* 30, 693-722.

Jiang, J. (2009). Designing pedagogic materials to improve awareness and productive use of L2 collocations. In A. Barfield and H. Gyllstad, *Researching Collocations in Another Language* (pp.99-113). England: Palgrave Macmillan.

Johns, A. M., Bawarshi, A., Coe, R. M., Hyland, K., Paltridge, B., Reiff, M. J., & Tardy, C. (2006). Crossing the boundaries of genre studies: Commentaries by experts. *Journal of Second Language Writing*, 15(3), 234-249.

Johns, T. (1986). The concordance: A language learner's research toll. *System,* 14, 151-162.

Johns, T. (1991a). Should you be persuaded. *ELR Journal,* 4, 1-16.

Johns, T. (1991b). From printout to handout. *ELR Journal,* 4, 27-46.

Johns, T & King, P. (Eds.) (1991). *English Language Research Journal Vol.4: Classroom Concordancing.* Birmingham: The University of Birmingham.

Jones, M. & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe and M. McCarthy (Eds.). *The Routledge Handbook of Corpus Linguistics* (pp.387- 398). New York: Routledge.

Jones, S., & Sinclair, J.M. (1974). English lexical collocations. A Study in Computational linguistics. *Cahiers de Lexicologie,* 24, 15-61.

Jordan, R.R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers.* Cambridge: Cambridge University Press.

Kaszubski, P. (2000). *Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-based Approach.* Adam Mickiewicz University, Poznán.

Kennedy, C. and Bolitho, R. (1984). *English for Specific Purposes*. Macmillan.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kettemann, B. (1996). *Concordancing in English Language Teaching.* Available online at http://www-gewi.kfunigraz.c.at/ed/project/concord.html.

Kheirzadeh, S. & Marandi, S.S. (2014). Concordancing as a Tool in Learning Collocations: The Case of Iranian EFL Learners. *Procedia-Social and Behavioral Sciences,* 98, 940-949.

Koester, A. (2010). Building Small specialised corpora. In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.66-79). London: Routledge.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *Elia*, 7, 79-102.

Koosha, M., & Jafarpour, A. (2006). Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL Journal Quarterly*, 8(4), 192-209.

Kübler, N., & Foucou, P. Y. (2003). Teaching English verbs with bilingual corpora: Examples in the computer science area. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, 185-206.

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English.* Providence, RI: Brown University Press.

Lam, J. (2001). A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography. *Research Reports*, Vol. 3. Language Centre, Hong Kong University of Science & Technology.

Laufer, B. (1997). What's in a word that makes it hard or easy? Some intra lexical factors affecting the difficulty of vocabulary acquisition. In N. Schmitt and M. McCarthy (Eds.),

Vocabulary: Description, Acquisition and Pedagogy (pp.140-155). Cambridge: Cambridge University Press.

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics,* 22, 1-26.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647-672.

Lee, D., & Swales, J. M. (2006). A corpus-based course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56-75.

Leech, G. (1997).Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and Language Corpora* (pp.1-23). New York: Addison Wesley Longman.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* London: Longman.

Lewis, M. (1997). *Implementing the Lexical Approach*. Hove, UK: LTP.

Lewis, M. (Ed.) (2000a). *Teaching Collocation: Further Developments in the Lexical Approach*. England: Language Teaching Publications.

Lewis, M. (2000b). Materials and resources for teaching collocation. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp.186-204). England: Language Teaching Publications.

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.): *Teaching Collocation: Further Developments in the Lexical Approach* (pp.10-27). LTP.

Li, S.-L., & Pemberton, R. (1994). An investigation of students' knowledge of academic and sub technical vocabulary. In L. Flowerdew and A. K. K. Tong (Eds.), *Entering Text* (pp.183-196). Hong Kong: The Hong Kong University of Science & Technology.

Li, Y., & Qian, D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System*, 38, 402-411.

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Lynn, R.W. (1973). Preparing word lists a suggested method. *RELC Journal*, 4(1), 25-32.

Ma, K. C. (1993).Small-corpora Concordacning in ESL Teaching and Learning, *Hong Kong Papers in Linguistics and Language Teaching*,16,11-30.

Mackey, A., & Gass, S.M. (2005). *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, D.D., & Richards, B.J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85-104.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, 19, 63-86.

Martı´nez, I., Beck, S., & Panza, C. (2009). Academic vocabulary in agriculture research articles. *English for Specific Purposes*, 28, 183-198.

Mason, O. & Hunston, S. (2004). The automatic recognition of verb patterns: a feasibility study. *International Journal of Corpus Linguistics*, 9, 253-270.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McCathy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. & O'Dell, F. (2005). *English Collocations in Use: How Words Work Together for Fluent and Natural English: Self-study and Classroom Use*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics, Method, Theory and Practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An Advanced Resource Book.* London: Routledge.

Meara, P. (1984). The study of lexis in interlanguage. In A. Davies, C. Criper & A.R.P. Howatt (Eds.), *Interlanguage* (pp.225-235). Edinburgh: Edinburgh University Press.

Mel'çuk, I. (1998). 'Collocations and lexical functions' in A.P.Cowie (ed.):*Phraseology.Theroy, Analysis and Applications*. Oxford: Clarendon Press, PP.23-53.

Meunier, F., & Granger, S. (2008). *Phraseology in Foreign Language Learning and Teaching.* John Benjamins Publishing Company: Amsterdam.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook* (2nd Ed.). Thousand Oaks, CA: Sage.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on Computer* (pp.186-198). London: Longman.

Minshall, D. E. (2013). *A Computer Science Word List*. Unpublished MA dissertation, University of Swansea. Available at DE Minshall www.baleap.org.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.40-63). Cambridge: Cambridge University Press.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25(2), 235-256.

Nagy, W.E, Anderson, R.C., Schommer, M., Scott, J.A. & Stallman, A. (1989). Morphological families in the internal Lexicon. *Reading Research Quarterly*, 24, 263-282.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary.* New York: Newbury House.

Nation, I.S.P. (2001). *Learning Vocabulary in another Language.* Cambridge: Cambridge University Press.

Nation, I.S.P. (2008). *Teaching Vocabulary: Strategies and Techniques*. Heinle.

Nation, I.S.P., & Heatley, A. (2007). Range. *Retrieved from www.vuw.ac.nz/lals/staff/paulnation/RANGE32.zip*.

Nation, I.S.P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.6-19). Cambridge: Cambridge University Press.

Nattinger, J.R., & DeCarrico, J.S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nelson, M. (2010). Building a written corpus: what are the basics? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.53-65). London: Routledge.

Nesi, H. (2008) Corpora and English for Academic Purposes published In LSP: Interfacing Language with other Realms: *Proceedings of the 6th Languages for Specific Purposes International Seminar.* Held 9-10 April 2008 at Universiti Teknologi Malaysia. Johor Bahru: Malaysia.

Nesi, H. (2014). Dictionary use by English language learners. *Language Teaching*, 47, 38-55.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242.

Nesselhuf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.) *How to use corpora in language teaching* (pp.125-152). Amsterdam: John Benjamins.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

O'Keeffe, (2007). The Pragmatics of Corpus Linguistics, keynote paper presented at the fourth Corpus Linguistics Conference held at the University of Birmingham, Birmingham.

O'Keeffe, A. & McCarthy, M. (2010). *Routledge Handbook of Corpus Linguistics.* London: Routledge.

O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3), 269-286.

Paltridge, B. (2001). *Genre and the Language Learning Classroom*. Ann Arbor, MI: University of Michigan Press.

Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Native like selection and native like fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and Communication* (pp.191-225). London: Longman.

Peacock, M. (2012). High-frequency collocations of nouns in research articles across eight disciplines. *Ibérica*, 23, 29-46.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Reppen, R. (2010). Building a corpus: What are the key considerations? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.31-37). London: Routledge.

Richards, B.J. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language*, 14, 201-209.

Römer, U. 2004. Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In Aston, G., Bernardini, S., and Stewart, D. (eds.) *Corpora and Language learners.* Amsterdam: John Benjamins, pp. 151-168.

Rutherford, W. (1987). *Second Language Grammar: Learning and Teaching*. London: Longman.

Salager, F. (1983). The Lexis of fundamental medical English: classifactory framework and rhetorical function (a statistical approach), *Reading in a Foreign Language*, 1, 54-64.

Saldana, J. (2009). *The Coding Manual for Qualitative Researchers.* London, UK: Sage.

Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes*, 21, 1-17.

Samraj, B. (2008). A discourse analysis of Masters' theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes,* 24, 55-67.

Santos, M. G. (2002). Examining the vocabulary skills of language minority students in community college study context. *TESOL.*

Schmidt, R. (1992). Awareness in second language acquisition, *Annual Review of Applied Linguistics*, 13, 206-226.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp.1-22). Amsterdam: John Benjamins.

Schmitt, N. & Schmitt, D. (1995). Vocabulary notebooks: Theoretical underpinnings and practical suggestions. *ELT Journal*, 49, 133-43.

Scott, M. (1996). *WordSmith Tools*. Oxford: Oxford University Press.

Scott, M. (2010). What can corpus software do? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics* (pp.136-151). London: Routledge.

Scott, M., & Johns, T. (1993). *Microconcord.* Oxford, UK: Oxford University Press.

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, 14, 369-385.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Springer: New York.

Sharwood-Smith, M. (1990). Consciousness-raising and second language learner. *Applied Linguistics*, 11(2), 159-168.

Shaw, P. (1991). Science research students' composing processes. *English for Specific Purposes*, 10, 189-206.

Shei, C. C., & Pain, H. (2000). An ESL Writer's Collocational Aid. *Computer Assisted Language Learning*, 13, 167-182.

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

Shopen, T. (1985). *Language Typology and Syntactic Description: Grammatical categories and the lexicon* (Vol. 3). Cambridge University Press.

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

Sinclair, J. (1966). Beginning the study of lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday and R.H. Robins (Eds.), *In memory of J.R. Firth* (pp.410-429). London: Longmans.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus*, 9(1), 75-106.

Sinclair, J. (2003). *Reading Concordances: An Introduction.* Harlow: Longman.

Sinclair, J. (2004). *Trust the Text*. London: Routledge.

Sinclair, J., Fox, G., Bullon, S., & Manning, E. (1995). *Collins COBUILD English Dictionary* (Eds.). London and Glasgow: Harper Collins.

Sinclair, J., Jones, S., & Daley, R. (2004). *English Collocation Studies: The OSTI Report.* Continuum: London.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review,* 64(3), 429-458.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Spack, R. (1988). Initiating ESL Students Into the Academic Discourse Community: How Far Should We Go? *TESOL Quarterly*, 22(1), 29-51.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate Writing. *Journal of English for Academic Purposes*, 8, 207-223.

Strevens, P. (1973). Technical, technological, and scientific English. *ELT Journal*, 27, 223-234.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 23-55.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7, 215-244.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, 25(2), 34-50.

Swales, J.M. (1981). *Aspects of Article Introductions*. Birmingham, UK: Language Studies Unit, University of Aston.

Swales, J.M. (1985). English language papers and authors first language: preliminary explorations. *Scientometrics*, 8(1-2), 91-101.

Swales, J. M. (1990). *Genre Analysis.* Cambridge: Cambridge University Press.

Tardy, C.M. (2009). *Building Genre Knowledge: L2 Writing.* West Lafayette, IN: Parlor Press.

Thompson, P. (2001). *A Pedagogically-motivated Corpus-based Examination of PhD Theses: Macrostructure, Citation Practices and Uses of Modal Verbs*. University of Reading.

Thompson, P. and Tribble, C. (2001). Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning and Technology*, 5, 174-92.

Thornbury, S. (1999). *How to Teach grammar*. London: Longman.

Thornbury, S. (2002). *How to Teach vocabulary*. London: Longman.

Thurstun, J., & Candlin, C.N. (1997). *Exploring Academic English – A Workbook for Student Essay Writing*. Sydney: NCELTR.

Thurstun, J., & Candlin, C.N. (1998). Concordancing and the teaching of the vocabulary of Academic English. *English for Specific Purpose*, 17(3), 267-280.

Tian, S. (2005). The impact of learning tasks and learner proficiency on the effectiveness of data-driven learning. *Journal of Pan-Pacific Association of Applied Linguistics*, 9(2), 263-275.

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

Tono, Y. (2003). Learner corpora: design, development and applications. In D., Archer, P., Rayson, A., Wilson and T., McEnery (Eds.) Proceedings *of the Corpus Linguistics Conference. Technical Papers 16*. Lancaster University: University Centre for Computer Corpus Research on Language, 800-809.

Tribble, C. (1990). Concordancing and an EAP Writing Programme. *CAELL Journal*, 1(2), 10-15.

Tribble, C., & Jones, G. (1990). *Concordances in the Classroom.* London: Longman.

Tribble, S.)2010). What are concordances and how are they used? In A. O'Keeffe and M. McCarthy (Eds.) *Routledge Handbook of Corpus Linguistics) pp.167*-183). London: Routledge.

Van Lier, L. (1995). *Introducing Language Awareness*. London: Penguin English.

Vongpumivitch, V., Huang, J., & Chang, Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28(1), 33-41.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27, 442-458.

Wang, K., & Nation, P. (2004). Word meaning in academic English: Homography in the academic word list. *Applied Linguistics*, 25(3), 291-314.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal*, 32, 201-232.

Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, 6(1), 18-35.

Ward, J. (2009). A basic engineering English word list for less proficient foundation Engineering undergraduates. *English for Specific Purposes*, 28, 170-182.

Webb, S., & Kajimoto, E. (2011). Learning Collocations: Do the Number of Collocates, Position of the Node Word, and Synonymy Affect Learning? *Applied Linguistics*, 32(3), 259-276.

Wesche, M., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53, 13-40.

West, M. (1953). *A General Service List of English Words*. London, England: Longman, Green.

Widdowson, HG. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3-25.

Williams, G.C. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151-171.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, 35(3), 451-482.

Woolard, G. (2000). Collocation-encouraging learner independence. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp.28-46). England: Language Teaching Publications.

Worthington, D., & Nation, I.S.P. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal*, 27, 1-11.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32, 213-231.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4), 463-489.

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). Conclusion: Navigating L2 collocation research. In A. Barfield and H. Gyllstad (Eds.), *Researching Collocations in Another Language: Multiple Interpretations* (pp.232-244). Basingstoke: Palgrave Macmillan.

Xue, Guoyi, & Nation, I.S.P. (1984). A University Word List. *Language Learning and Communication*, 3(2), 215-229.

Yang, H. Z. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing*, 1(2), 93-103.

Ying, Y & O'Neill, M. (2009). Collocation learning through an 'AWARE' approach: learner perspectives and learning process. In A. Barfield and H. Gyllstad (Eds.), *Researching Collocations in Another Language* (pp.181-193). England: Palgrave Macmillan.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2. *Journal of Second Language Writing*, 13(4), 257-283.

Yuldashev, A. Fernandez, J., & Thorne, S. (2013). Second Language Learners' Contiguous and Discontiguous Multi-Word Unit Use Over Time. *The Modern Language Journal*, 97(1), 31-45.

Appendices

Appendix A: List of the Research Articles Constituting the Reference Corpus

A-Software Engineering Journals: 19 Articles

1-IEEE Transaction on Visualisation and Computer Graphics (7)

Chen, J., Cai, H., Auchus, A. P., & Laidlaw, D. H. (2012). Effects of Stereo and Screen Size on the Legibility of Three-Dimensional Streamtube Visualization. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2130-2139.

Gasteiger, R., Lehmann, D. J., van Pelt, R., Janiga, G., Beuing, O., Vilanova, A., & Preim, B. (2012). Automatic Detection and Visualization of Qualitative Hemodynamic Characteristics in Cerebral Aneurysms. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2178-2187.

Obermaier, H., & Joy, K. I. (2012). Derived metric tensors for flow surface visualization. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2149-2158.

Reich, W., & Scheuermann, G. (2012). Analysis of streamline separation at infinity using time-discrete Markov chains. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2140-2148.

Schindler, B., Fuchs, R., Barp, S., Waser, J., Pobitzer, A., Carnecky, R., & Peikert, R. (2012). Lagrangian coherent structures for design analysis of revolving doors. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2159-2168.

Treib, M., Burger, K., Reichl, F., Meneveau, C., Szalay, A., & Westermann, R. (2012). Turbulence visualization at the terascale on desktop PCs. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2169-2177.

Wenger, S., Ament, M., Guthe, S., Lorenz, D., Tillmann, A., Weiskopf, D., & Magnor, M. (2012). Visualization of astronomical nebulae via distributed multi-GPU compressed sensing tomography. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(12), 2188-2197.

2-ACM Transactions on Graphics (TOG) (6)

Akhter, I., Simon, T., Khan, S., Matthews, I., & Sheikh, Y. (2012). Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, *31*(2), 17.

Aliaga, D. G., Yeung, Y. H., Law, A., Sajadi, B., & Majumder, A. (2012). Fast high-resolution appearance editing using superimposed projections. *ACM Transactions on Graphics (TOG)*, *31*(2), 13.

Berthouzoz, F., & Fattal, R. (2012). Resolution enhancement by vibrating displays. *ACM Transactions on Graphics (TOG)*, *31*(2), 15.

Boyd, L., & Bridson, R. (2012). MultiFLIP for energetic two-phase fluid simulation. *ACM Transactions on Graphics (TOG)*, *31*(2), 16.

Seol, Y., Lewis, J. P., Seo, J., Choi, B., Anjyo, K., & Noh, J. (2012). Spacetime expression cloning for blendshapes. *ACM Transactions on Graphics (TOG)*, *31*(2), 14.

Tevs, A., Berner, A., Wand, M., Ihrke, I., Bokeloh, M., Kerber, J., & Seidel, H. P. (2012). Animation cartography – intrinsic reconstruction of shape and motion. *ACM Transactions on Graphics (TOG)*, *31*(2), 12.

3-**ACM Transactions on Software Engineering and Methodology (TOSEM) (6)**

Fantechi, A., Gnesi, S., Lapadula, A., Mazzanti, F., Pugliese, R., & Tiezzi, F. (2012). A logical verification methodology for service-oriented computing. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 16.

Jennings, P., Ghosh, A. P., & Basu, S. (2012). A two-phase approximation for model checking probabilistic unbounded until properties of probabilistic systems. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 18.

Kästner, C., Apel, S., Thüm, T., & Saake, G. (2012). Type checking annotation-based product lines. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 14.

Qi, D., Roychoudhury, A., Liang, Z., & Vaswani, K. (2012). DARWIN: An approach to debugging evolving programs. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 19.

Shonle, M., Griswold, W. G., & Lerner, S. (2012). A framework for the checking and refactoring of crosscutting concepts. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 15.

Strecker, J., & Memon, A. M. (2012). Accounting for defect characteristics in evaluations of testing techniques. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *21*(3), 17.

B-Artificial Intelligence Journals: 26 Articles

1-IEEE Transactions on Pattern Analysis and Machine Intelligence (10)

Dai, J., Feng, J., & Zhou, J. (2012). Robust and efficient ridge-based palmprint matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1618-1632.

Del Bue, A., Xavier, J., Agapito, L., & Paladini, M. (2012). Bilinear modeling via augmented lagrange multipliers (balm). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1496-1508.

Jeong, Y., Nister, D., Steedly, D., Szeliski, R., & Kweon, I. S. (2012). Pushing the envelope of modern methods for bundle adjustment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1605-1617.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1549-1562.

Maier-Hein, L., Franz, A. M., dos Santos, T. R., Schmidt, M., Fangerau, M., Meinzer, H., & Fitzpatrick, J. M. (2012). Convergent iterative closest-point algorithm to accomodate anisotropic and inhomogenous localization error. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1520-1532.

Ouzounis, G. K., Pesaresi, M., & Soille, P. (2012). Differential area profiles: Decomposition properties and efficient computation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1533-1548.

Qiu, P., & Mukherjee, P. S. (2012). Edge structure preserving 3D image denoising by local surface approximation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1457-1468.

Skibbe, H., Reisert, M., Schmidt, T., Brox, T., Ronneberger, O., & Burkhardt, H. (2012). Fast rotation invariant 3D feature computation utilizing efficient local neighborhood operators. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1563-1575.

Wang, Q. F., Yin, F., & Liu, C. L. (2012). Handwritten Chinese text recognition by integrating multiple contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1469-1481.

Wu, T. P., Yeung, S. K., Jia, J., Tang, C. K., & Medioni, G. (2012). A closed-form solution to tensor voting: Theory and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(8), 1482-1495.

2-International Journal of Intelligent Systems (6)

Anderson, D. T., Ros, M., Keller, J. M., Cuéllar, M. P., Popescu, M., Delgado, M., & Vila, A. (2012). Similarity measure for anomaly detection and comparing human behaviors. *International Journal of Intelligent Systems*, *27*(8), 733-756.

Chassy, P., Calmès, M. D., & Prade, H. (2012). Making sense as a process emerging from perception–memory interaction: A model. *International Journal of Intelligent Systems*, *27*(8), 757-775.

Lloret, E., Llorens, H., Moreda, P., Saquete, E., & Palomar, M. (2011). Text summarization contribution to semantic question answering: New approaches for finding answers on the web. *International Journal of Intelligent Systems*, *26*(12), 1125-1152.

Maturo, A. (2011). Fuzzy measures and coherent join measures. *International Journal of Intelligent Systems*, *26*(12), 1196-1205.

Prade, H., & Richard, G. (2011). Cataloguing/analogizing: A nonmonotonic view. *International Journal of Intelligent Systems*, *26*(12), 1176-1195.

Qi, X., Barrett, S., & Chang, R. (2011). A noise-resilient collaborative learning approach to content-based image retrieval. *International Journal of Intelligent Systems*, *26*(12), 1153-1175.

3-IEEE Transactions on Evolutionary Computation (10)

Arabas, J. (2012). Approximating the genetic diversity of populations in the quasi-equilibrium state. *Evolutionary Computation, IEEE Transactions on*, *16*(5), 632-644.

Arias-Montano, A., Coello Coello, C. A., & Mezura Montes, E. (2012). Multiobjective evolutionary algorithms in aeronautical and aerospace engineering. *Evolutionary Computation, IEEE Transactions on*, *16*(5), 662-694.

Chiong, R., & Kirley, M. (2012). Effects of iterated interactions in multiplayer spatial evolutionary games. *Evolutionary Computation, IEEE Transactions on*, *16*(4), 537-555.

Howard, G., Gale, E., Bull, L., de Lacy Costello, B., & Adamatzky, A. (2012). Evolution of plastic learning in spiking networks via memristive connections. *Evolutionary Computation, IEEE Transactions on*, *16*(5), 711-729.

Joó, A. M., Ekart, A., & Neirotti, J. P. (2012). Genetic algorithms for discovery of matrix multiplication methods. *IEEE transactions on evolutionary computation*, *16*(5), 749-751.

Kohl, N., & Miikkulainen, R. (2012). An integrated neuroevolutionary approach to reactive control and high-level strategy. *Evolutionary Computation, IEEE Transactions on*, *16*(4), 472-488.

Naznin, F., Sarker, R., & Essam, D. (2012). Progressive alignment method using genetic algorithm for multiple sequence alignment. *Evolutionary Computation, IEEE Transactions on*, *16*(5), 615-631.

Neshatian, K., Zhang, M., & Andreae, P. (2012). A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming. *IEEE transactions on evolutionary computation*, *16*(5), 645-661.

Qu, B. Y., Suganthan, P. N., & Liang, J. J. (2012). Differential evolution with neighborhood mutation for multimodal optimization. *IEEE transactions on evolutionary computation*, *16*(5), 601-614.

Schutze, O., Esquivel, X., Lara, A., & Coello Coello, C. A. (2012). Using the averaged hausdorff distance as a performance measure in evolutionary multiobjective optimization. *Evolutionary Computation, IEEE Transactions on*, *16*(4), 504-522.

C- Information Systems Journals: 18 Articles

1-MIS Quarterly (7)

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: theory and evidence. *MIS Quarterly*, *36*(2), 395-426.

Lee, Y., Chen, A. N., & Ilie, V. (2012). Can Online Wait Be Managed? The Effect of Filler Interfaces and Presentation Modes on Perceived Waiting Time Online. *MIS Quarterly*, *36*(2), 365-394.

Liu, C. Z., Kemerer, C. F., Slaughter, S. A., & Smith, M. D. (2012). Standards competition in the presence of digital conversion technology: An empirical analysis of the flash memory card market. *MIS Quarterly*, *36*(3), 921-942.

Oestreicher-Singer, G., & Sundararajan, A. (2012). Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*, *36*(1), 65-83.

Rivard, S., & Lapointe, L. (2012). Information technology implementers' responses to user resistance: nature and effects. *MIS Quarterly*, *36*(3), 897-920.

Sun, H. (2012). Understanding user revisions when using information system features: adaptive system use and triggers. *MIS Quarterly*, *36*(2), 453-478.

VanderMeer, D., Dutta, K., & Datta, A. (2012). A cost-based database request distribution technique for online e-commerce applications. *MIS Quarterly*, *36*(2), 479-507.

2-Enterprise Information Systems (4)

Fu, C., Zhang, G., Yang, J., & Liu, X. (2011). Study on the contract characteristics of Internet architecture. *Enterprise Information Systems*, *5*(4), 495-513.

Ma, J., Wang, K., & Xu, L. (2011). Modelling and analysis of workflow for lean supply chains. *Enterprise Information Systems*, *5*(4), 423-447.

Sun, Y., & Bhattacherjee, A. (2011). Multi-level analysis in information systems research: the case of enterprise resource planning system usage in China. *Enterprise Information Systems*, *5*(4), 469-494.

Zdravković, M., Panetto, H., Trajanović, M., & Aubry, A. (2011). An approach for formalising the supply chain operations. *Enterprise Information Systems*, *5*(4), 401-421.

3-ACM Transactions on Information Systems (7)

Altingovde, I. S., Ozcan, R., & Ulusoy, Ö. (2012). Static index pruning in web search engines: Combining term and document popularities with query views. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 2.

Bhatia, S., & Mitra, P. (2012). Summarizing figures, tables, and algorithms in scientific publications to augment search results. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 3.

Broschart, A., & Schenkel, R. (2012). High-performance processing of text queries with tunable pruned term and term pair indexes. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 5.

Carterette, B. A. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 4.

Chapelle, O., Joachims, T., Radlinski, F., & Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 6.

Fariña, A., Brisaboa, N. R., Navarro, G., Claude, F., Places, Á. S., & Rodríguez, E. (2012). Word-based self-indexes for natural language text. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 1.

Pal, A., Harper, F. M., & Konstan, J. A. (2012). Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems (TOIS)*, *30*(2), 10.

Appendix B:The 100 Most Frequent AWL Nouns and Verbs in Each Student Corpus

No.	Most frequent N in NNS corpus	Most frequent N in NS corpus	Most frequent V in NNS corpus	Most frequent V in NS corpus
1.	networkNNW	Data_NNW	implemented_VVN	implemented_VVN
2.	networksNNY	projectNNW	implement_VVI	implement_VVI
3.	NetworkingNNW	projectsNNY	implementing_VVG	Implementing_VVG
4.	dataNNK	codeNNW	implements_VVZ	implement_VVB
5.	dataNNW	codesNNY	implement_VVB	implements_VVZ
6.	routerNNW	taskNNW	implemented_VVD	implemented_VVD
7.	routersNNY	tasksNNY	required_VVN	created_VVN
8.	routeNNW	processNNW	requires_VVZ	create_VVI
9.	routesNNY	processingNNW	require_VVB	creates_VVZ
10.	routersNNK	processorNNW	require_VVI	Create_VVB
11.	processNNW	processesNNY	required_VVD	created_VVD
12.	processingNNW	processorsNNY	generated_VVN	required_VVN
13.	processesNNY	methodNNW	generate_VVI	requires_VVZ
14.	methodNNW	methodsNNY	generates_VVZ	require_VVB
15.	methodsNNY	DesignNNW	generated_VVD	required_VVD
16.	methodsNNK	designerNNW	generate_VVB	display_VVI
17.	fileNNW	designingNNW	created_VVN	displayed_VVN
18.	filesNNY	functionNNW	create_VVI	displays_VVZ
19.	linkNNW	functionsNNY	create_VVB	displaying_VVG
20.	linksNNY	functioningNNW	creates_VVZ	display_VVB
21.	linkingNNW	sectionNNW	created_VVD	displayed_VVD
22.	featuresNNY	sectionsNNY	defined_VVN	selected_VVN
23.	featureNNW	fileNNW	define_VVI	select_VVI
24.	DocumentNNW	filesNNY	defines_VVZ	selecting_VVG
25.	documentsNNY	accessNNW	define_VVB	selects_VVZ

26.	documentationNNW	accessingNNW	defined_VVD	select_VVB
27.	designNNW	accessibilityNNW	achieved_VVN	selected_VVD
28.	designsNNY	networkNNW	achieve_VVI	defined_VVN
29.	DesignersNNY	networksNNY	achieve_VVB	define_VVI
30.	designerNNW	networkingNNW	achieves_VVZ	defines_VVZ
31.	projectNNW	documentNNW	achieved_VVD	define_VVB
32.	projectsNNY	documentsNNY	extract_VVI	defined_VVD
33.	protocolNNW	documentationNNW	extracted_VVN	found_VVN
34.	protocolsNNY	environmentNNW	extracting_VVG	found_VVD
35.	approachNNW	environmentsNNY	extract_VVB	ensure_VVI
36.	approachesNNY	featuresNNY	extracts_VVZ	ensures_VVZ
37.	functionNNW	featureNNW	extracted_VVD	Ensure_VVB
38.	functionsNNY	issuesNNY	obtained_VVN	ensured_VVD
39.	functioningNNW	issueNNW	obtain_VVI	ensured_VVN
40.	simulationNNW	componentsNNY	obtaining_VVG	calling_VVG
41.	SimulationsNNY	componentNNW	obtains_VVZ	calls_VVZ
42.	elementNNW	deviceNNW	Obtain_VVB	call_VVI
43.	elementsNNY	devicesNNY	obtained_VVD	call_VVB
44.	channelNNW	frameworkNNW	illustrates_VVZ	identify_VVI
45.	channelsNNY	frameworksNNY	illustrated_VVN	identifying_VVG
46.	codeNNW	modeNNW	illustrate_VVB	identify_VVB
47.	codesNNY	modesNNY	illustrate_VVI	achieved_VVN
48.	techniquesNNY	approachNNW	illustrated_VVD	achieve_VVI
49.	techniqueNNW	approachesNNY	identify_VVI	achieves_VVZ
50.	taskNNW	sourceNNW	identifying_VVG	achieve_VVB
51.	tasksNNY	sourcesNNY	identify_VVB	achieved_VVD
52.	outputNNW	areaNNW	assigned_VVN	generated_VVN
53.	outputsNNY	areasNNY	assign_VVI	generate_VVI
54.	resourcesNNY	textNNW	assigning_VVG	generates_VVZ
55.	resourceNNW	textsNNY	assigns_VVZ	generated_VVD

56.	devicesNNY	textingNNW	select_VVI	Generate_VVB
57.	deviceNNW	elementNNW	selected_VVN	involves_VVZ
58.	scenarioNNW	elementsNNY	selecting_VVG	involved_VVD
59.	scenariosNNY	locationNNW	selects_VVZ	involve_VVI
60.	securityNNW	locationsNNY	consists_VVZ	involved_VVN
61.	accessNNW	inputNNW	consist_VVB	involve_VVB
62.	layerNNW	inputsNNY	consisting_VVG	occurs_VVZ
63.	layersNNY	inputtingNNW	consisted_VVN	occur_VVI
64.	mechanismNNW	attributeNNW	displayed_VVD	occurred_VVN
65.	mechanismsNNY	attributesNNY	displayed_VVN	occurred_VVD
66.	sourceNNW	researchNNW	display_VVI	occur_VVB
67.	inputNNW	researchersNNY	Display_VVB	occurring_VVG
68.	inputsNNY	researchesNNY	displays_VVZ	occure_VVI
69.	errorNNW	errorNNW	displaying_VVG	reoccurs_VVZ
70.	errorsNNY	errorsNNY	occurs_VVZ	detect_VVI
71.	structureNNW	computerNNW	occur_VVI	detected_VVN
72.	structuresNNY	computersNNY	occurred_VVD	detecting_VVG
73.	communicationNNW	formatNNW	occur_VVB	detects_VVZ
74.	communicationsNNY	formatsNNY	occurring_VVG	detected_VVD
75.	transmitterNNW	formattingNNW	occurred_VVN	demonstrates_VVZ
76.	transmittersNNY	routeNNW	enables_VVZ	demonstrate_VVI
77.	TransmissionNNW	routersNNY	enable_VVB	demonstrated_VVN
78.	transmissionsNNY	routesNNY	enabled_VVN	demonstrate_VVB
79.	parametersNNY	variableNNW	enable_VVI	demonstrated_VVD
80.	parameterNNW	variablesNNY	enabled_VVD	removed_VVN
81.	technologyNNW	siteNNW	ensure_VVI	remove_VVB
82.	areaNNW	sitesNNY	ensures_VVZ	remove_VVI
83.	areasNNY	linkNNW	ensure_VVB	removes_VVZ
84.	SectionNNW	linksNNY	ensured_VVN	removed_VVD
85.	SectionsNNY	linkingNNW	maintain_VVI	indicates_VVZ

86.	textNNW	linkersNNY	maintaining_VVG	indicate_VVI
87.	textsNNY	structureNNW	maintained_VVN	indicate_VVB
88.	componentsNNY	structuresNNY	maintains_VVZ	indicated_VVD
89.	componentNNW	outputNNW	maintain_VVB	indicated_VVN
90.	callNNW	outputsNNY	maintained_VVD	specify_VVI
91.	callsNNY	resourcesNNY	affect_VVI	Specifying_VVG
92.	researchNNW	ResourceNNW	affected_VVN	specify_VVB
93.	researchersNNY	analysisNNW	affects_VVZ	extract_VVI
94.	researchesNNY	formulaNNW	affecting_VVG	extracted_VVN
95.	researcherNNW	formulaeNNY	affect_VVB	extracting_VVG
96.	schemeNNW	formulasNNY	indicates_VVZ	extract_VVB
97.	schemesNNY	formulationNNW	indicate_VVI	extracts_VVZ
98.	siteNNW	simulationNNW	indicated_VVN	enable_VVI
99.	sitesNNY	simulationsNNY	indicate_VVB	enables_VVZ
100.	capacityNNW	NormalisationNNW	indicated_VVD	enabled_VVD

Appendix C: The Verbs and Nouns Selected for Insertion in *ConcGram* as Potential Collocation Nodes

No.	NNS N list	No.	NS N list	No.	NNS V list	No.	NS V list
1.	networkNNW	1.	dataNNK	1.	implemented_VVN	1.	implemented_VVN
	networksNNY	2.	projectNNW		implement_VVI		implement_VVI
	NetworkingNNW		projectsNNY		implementing_VVG		Implementing_VVG
2.	dataNNK	3.	codeNNW		implements_VVZ		implement_VVB
	dataNNW		codesNNY		implement_VVB		implements_VVZ
3.	routerNNW	4.	taskNNW		implemented_VVD		implemented_VVD
	routersNNY		tasksNNY	2.	required_VVN	2.	created_VVN
	routeNNW	5.	processNNW		requires_VVZ		create_VVI
	routesNNY		processingNNW		require_VVB		creates_VVZ
	routersNNK		processorNNW		require_VVI		Create_VVB
4.	processNNW		processesNNY		required_VVD		created_VVD
	processingNNW		processorsNNY	3.	generated_VVN	3.	required_VVN
	processesNNY	6.	methodNNW		generate_VVI		requires_VVZ
5.	methodNNW		methodsNNY		generates_VVZ		require_VVB
	methodsNNY	7.	DesignNNW		generated_VVD		required_VVD
	methodsNNK		designerNNW		generate_VVB	4.	display_VVI
6.	fileNNW		designingNNW	4.	created_VVN		displayed_VVN
	filesNNY	8.	functionNNW		create_VVI		displays_VVZ

7.	linkNNW		functionsNNY		create_VVB		displaying_VVG
	linksNNY		functioningNNW		creates_VVZ		display_VVB
	linkingNNW	9.	sectionNNW		created_VVD		displayed_VVD
8.	featuresNNY		sectionsNNY	5.	defined_VVN	5.	selected_VVN
	featureNNW	10.	fileNNW		define_VVI		select_VVI
9.	DocumentNNW		filesNNY		defines_VVZ		selecting_VVG
	documentsNNY	11.	accessNNW		define_VVB		selects_VVZ
	documentationNNW		accessingNNW		defined_VVD		select_VVB
10.	designNNW		accessibilityNNW	6.	achieved_VVN		selected_VVD
	designsNNY	12.	networkNNW		achieve_VVI	6.	defined_VVN
	DesignersNNY		networksNNY		achieve_VVB		define_VVI
	designerNNW		networkingNNW		achieves_VVZ		defines_VVZ
11.	projectNNW	13.	documentNNW		achieved_VVD		define_VVB
	projectsNNY		documentsNNY	7.	extract_VVI		defined_VVD
12.	protocolNNW		documentationNNW		extracted_VVN	7.	found_VVN
	protocolsNNY	14.	environmentNNW		extracting_VVG		found_VVD
13.	approachNNW		environmentsNNY		extract_VVB	8.	ensure_VVI
	approachesNNY	15.	featuresNNY		extracts_VVZ		ensures_VVZ
14.	functionNNW		featureNNW		extracted_VVD		Ensure_VVB
	functionsNNY	16.	issuesNNY	8.	obtained_VVN		ensured_VVD
	functioningNNW		issueNNW		obtain_VVI		ensured_VVN

15.	simulationNNW	17.	componentsNNY		obtaining_VVG	9.	calling_VVG
	SimulationsNNY		componentNNW		obtains_VVZ		calls_VVZ
16.	elementNNW	18.	deviceNNW		Obtain_VVB		call_VVI
	elementsNNY		devicesNNY		obtained_VVD		call_VVB
17.	channelNNW	19.	frameworkNNW	9.	illustrates_VVZ	10.	identify_VVI
	channelsNNY		frameworksNNY		illustrated_VVN		identifying_VVG
18.	codeNNW	20.	modeNNW		illustrate_VVB		identify_VVB
	codesNNY		modesNNY		illustrate_VVI	11.	achieved_VVN
19.	techniquesNNY	21.	approachNNW		illustrated_VVD		achieve_VVI
	techniqueNNW		approachesNNY	10.	identify_VVI		achieves_VVZ
20.	taskNNW	22.	sourceNNW		identifying_VVG		achieve_VVB
	tasksNNY		sourcesNNY		identify_VVB		achieved_VVD
21.	outputNNW	23.	areaNNW	11.	assigned_VVN	12.	generated_VVN
	outputsNNY		areasNNY		assign_VVI		generate_VVI
22.	resourcesNNY	24.	textNNW		assigning_VVG		generates_VVZ
	resourceNNW		textsNNY		assigns_VVZ		generated_VVD
23.	devicesNNY		textingNNW	12.	select_VVI		Generate_VVB
	deviceNNW	25.	elementNNW		selected_VVN	13.	involves_VVZ
24.	scenarioNNW		elementsNNY		selecting_VVG		involved_VVD
	scenariosNNY	26.	locationNNW		selects_VVZ		involve_VVI
25.	securityNNW		locationsNNY	13.	consists_VVZ		involved_VVN
26.	accessNNW	27.	inputNNW		consist_VVB		involve_VVB

27.	layerNNW		inputsNNY		consisting_VVG	14.	occurs_VVZ
	layersNNY		inputtingNNW		consisted_VVN		occur_VVI
28.	mechanismNNW	28.	attributeNNW	14.	displayed_VVD		occurred_VVN
	mechanismsNNY		attributesNNY		displayed_VVN		occurred_VVD
29.	sourceNNW	29.	researchNNW		display_VVI		occur_VVB
30.	inputNNW		researchersNNY		Display_VVB		occurring_VVG
	inputsNNY		researchesNNY		displays_VVZ		occure_VVI
31.	errorNNW	30.	errorNNW		displaying_VVG		reoccurs_VVZ
	errorsNNY		errorsNNY	15.	occurs_VVZ	15.	detect_VVI
32.	structureNNW	31.	computerNNW		occur_VVI		detected_VVN
	structuresNNY		computersNNY		occurred_VVD		detecting_VVG
33.	communicationNNW	32.	formatNNW		occur_VVB		detects_VVZ
	communicationsNNY		formatsNNY		occurring_VVG		detected_VVD
34.	transmitterNNW		formattingNNW		occurred_VVN	16.	demonstrates_VVZ
	transmittersNNY	33.	routeNNW	16.	enables_VVZ		demonstrate_VVI
	TransmissionNNW		routersNNY		enable_VVB		demonstrated_VVN
	transmissionsNNY		routesNNY		enabled_VVN		demonstrate_VVB
35.	parametersNNY	34.	variableNNW		enable_VVI		demonstrated_VVD
	parameterNNW		variablesNNY		enabled_VVD	17.	removed_VVN
36.	technologyNNW	35.	siteNNW	17.	ensure_VVI		remove_VVB
37.	areaNNW		sitesNNY		ensures_VVZ		remove_VVI
	areasNNY	36.	linkNNW		ensure_VVB		removes_VVZ

38.	SectionNNW		linksNNY		ensured_VVN		removed_VVD
	SectionsNNY		linkingNNW	18.	maintain_VVI	18.	indicates_VVZ
39.	textNNW		linkersNNY		maintaining_VVG		indicate_VVI
	textsNNY	37.	structureNNW		maintained_VVN		indicate_VVB
40.	componentsNNY		structuresNNY		maintains_VVZ		indicated_VVD
	componentNNW	38.	outputNNW		maintain_VVB		indicated_VVN
41.	callNNW		outputsNNY		maintained_VVD	19.	specify_VVI
	callsNNY	39.	resourcesNNY	19.	affect_VVI		Specifying_VVG
42.	researchNNW		ResourceNNW		affected_VVN		specify_VVB
	researchersNNY	40.	analysisNNW		affects_VVZ	20.	extract_VVI
	researchesNNY	41.	formulaNNW		affecting_VVG		extracted_VVN
	researcherNNW		formulaeNNY		affect_VVB		extracting_VVG
43.	schemeNNW		formulasNNY	20.	indicates_VVZ		extract_VVB
	schemesNNY		formulationNNW		indicate_VVI		extracts_VVZ
44.	siteNNW	42.	simulationNNW		indicated_VVN	21.	enable_VVI
	sitesNNY		simulationsNNY		indicate_VVB		enables_VVZ
45.	capacityNNW	43.	NormalisationNNW		indicated_VVD		enabled_VVD
46.	sequenceNNW	44.	layerNNW				enable_VVB
	sequencesNNY		layersNNY				enabled_VVN
47.	priorityNNW	45.	versionNNW			22.	conducted_VVN
48.	summaryNN		versionsN				conducting_VVG

	W		NY				
49.	domainNNW	46.	phaseNNW				conduct_VVI
	domainsNNY		phasesNNY				Conduct_VVB
50.	environmentNNW	47.	conceptNNW				conducted_VVD
	environmentsNNY		conceptsNNY			23.	consists_VVZ
51.	DetectionNNW	48.	parametersNNY				consisting_VVG
	detectorNNW		parameterNNW				consist_VVI
	detectionsNNY	49.	imagesNNY				consist_VVB
52.	computerNNW		imageNNW				consisted_VVN
	ComputersNNY	50.	targetNNW				consisted_VVD
53.	factorsNNY		targetsNNY			24.	affect_VVI
	factorNNW	51.	factorsNNY				affected_VVN
54.	issuesNNY		factorNNW				affecting_VVG
	issueNNW	52.	instanceNNW				affects_VVZ
55.	analysisNNW		instancesNNY				affect_VVB
56.	PhaseNNW	53.	rangeNNW				affected_VVD
	phasesNNY		rangesNNY			25.	converted_VVN
57.	allocationNNW	54.	optionNNW				convert_VVI
	allocationNNK		optionsNNY				converting_VVG
58.	rangeNNW	55.	techniquesNNY				converts_VVZ
	rangesNNY		techniqueNNW				convert_VVB
59.	labelNNW	56.	transferNN				converted_VVD

			W				
	labelsNNY		transfersNNY			26.	focus_VVI
60.	periodNNW		transferenceNNW				focuses_VVZ
	periodsNNY	57.	capacityNNW				focused_VVN
61.	impactNNW	58.	technologyNNW				focus_VVB
	impactsNNY	59.	benefitsNNY				focusing_VVG
62.	entityNNW		benefitNNW				focused_VVD
63.	utilizationNNW	60.	coreNNW				focussed_VVN
64.	ChapterNNW		coresNNY				focussed_VVD
	chaptersNNY	61.	procedureNNW				refocused_VVN
65.	goalNNW		proceduresNNY				
	goalsNNY	62.	utilizationNNW				
66.	extractionNNW						
	extractionsNNY						
	ExtractionsNPK						
67.	coreNNW						
68.	operationNNW						
	operationsNNY						
total	68 AWL families		62 AWL families		20 AWL families		26 AWL families

Appendix D: Chi-square Test for the 400 N and V Collocations from Both Student Corpora

No.	100NNS N collocations		NNS corpus frequency	RC frequency	Chi-square value	P value	Significant Over/underuse
1.	network_NNW	traffic_NNW	70	8	111.25	0.0001	Significant overuse
2.	simulation_NNW	results_NNY	56	5	93.483	0.0001	Significant overuse
3.	sites_NNY	Web_NNW	39	29	17.517	0.0001	Significant overuse
4.	error_NNW	rate_NNW	32	9	36.715	0.0001	Significant overuse
5.	site_NNW	Web_NNW	30	37	3.888	0.05	Significant underuse
6.	extraction_NNW	information_NNW	28	3	45.126	0.0001	Significant overuse
7.	allocation_NNW	dynamic_AJK	27	4	40.155	0.0001	Significant overuse
8.	data_NNK	sets_NNY	26	40	1.061	0.3	Non significant
9.	data_NNK	layer_NNW	26	21	10.137	0.002	Significant overuse
10.	data_NNK	different_AJK	26	15	16.586	0.0001	Significant overuse
11.	data_NNK	amount_NNW	26	10	24.37	0.0001	Significant overuse
12.	data_NNK	access_NNW	19	4	25.019	0.0001	Significant overuse
13.	design_NNW	system_NNW	19	13	9.694	0.003	Significant overuse
14.	source_NNW	open_AJK	19	33	0.228	0.66	Non significant
15.	techniques_NNY	different_AJK	19	10	13.434	0.001	Significant overuse
16.	data_NNK	training_NNW	18	70	6.644	0.01	Significant underuse
17.	data_NNK	source_NNW	18	25	1.374	0.25	Non significant
18.	data_NNK	user_NNW	18	9	13.419	0.001	Significant overuse
19.	data_NNK	information_NNW	18	8	14.991	0.0001	Significant overuse
20.	data_NNK	Web_NNW	18	7	16.73	0.0001	Significant overuse
21.	period_NNW	time_NNW	17	12	8.282	0.006	Significant overuse
22.	layer_NNW	application_N	15	10	13.64	0.009	Significant overuse

		NW					
23.	network_NNW	other_AJK	15	6	13.4	0.001	Significant overuse
24.	protocols_NNY	different_AJK	15	2	22.964	0.0001	Significant overuse
25.	scenarios_NNY	different_AJK	15	10	7.942	0.009	Significant overuse
26.	Section_NNW	previous_AJK	15	32	0.048	0.87	Non significant
27.	documents_NNY	web_NNW	14	8	9.032	0.005	Significant overuse
28.	features_NNY	frequency_NNW	14	4	15.923	0.0001	Significant overuse
29.	resources_NNY	available_AJK	14	5	13.849	0.0001	Significant overuse
30.	simulation_NNW	time_NNW	14	6	13.032	0.001	Significant overuse
31.	text_NNW	web_NNW	14	4	15.923	0.0001	Significant overuse
32.	approaches_NNY	different_AJK	13	15	2.132	0.162	Non significant
33.	data_NNK	size_NNW	13	12	3.881	0.057	Significant overuse
34.	data_NNK	time_NNW	13	9	6.519	0.021	Significant overuse
35.	data_NNK	other_AJK	13	6	10.465	0.002	Significant overuse
36.	feature_NNW	selection_NNW	13	27	0.015	0.993	Non significant
37.	processing_NNW	language_NNW	13	5	12.185	0.002	Significant overuse
38.	data_NNK	type_NNW	12	7	7.555	0.012	Significant overuse
39.	data_NNK	applications_NNY	12	5	10.559	0.003	Significant overuse
40.	environment_NNW	development_NNW	12	6	8.946	0.005	Significant overuse
41.	methods_NNY	Class_NNW	12	35	1.313	0.282	Non significant
42.	range_NNW	wide_AJK	12	8	6.354	0.016	Significant overuse
43.	Sections_NNY	following_AJK	12	19	0.391	0.569	Non significant
44.	analysis_NNW	Results_NNY	11	16	0.651	0.42	Non significant
45.	analysis_NNW	performance_NNW	11	8	5.118	0.029	Significant overuse
46.	code_NNW	source_NNW	11	70	14.325	0.0001	Significant underuse
47.	components_NN	set_NNW	11	4	10.744	0.002	Significant overuse

	Y						
48.	data_NNK	set_NNW	11	40	3.217	0.08	Non significant
49.	data_NNK	input_NNW	11	11	2.72	0.114	Non significant
50.	methods_NNY	other_AJK	11	39	2.928	0.103	Non significant
51.	methods_NNY	new_AJK	11	5	8.979	0.006	Significant overuse
52.	period_NNW	sample_NNW	11	6	7.482	0.009	Significant overuse
53.	processing_NNW	natural_AJK	11	5	8.979	0.006	Significant overuse
54.	Section_NNW	following_AJK	11	16	0.651	0.42	Non significant
55.	structure_NNW	tree_NNW	11	8	5.118	0.029	Significant overuse
56.	approach_NNW	new_AJK	10	16	0.298	0.678	Non significant
57.	code_NNW	following_AJK	10	5	7.455	0.011	Significant overuse
58.	documents_NNY	relevant_AJK	10	38	3.415	0.071	Non significant
59.	features_NNY	other_AJK	10	7	4.933	0.037	Significant overuse
60.	network_NNW	node_NNW	10	5	7.455	0.011	Significant overuse
61.	process_NNW	time_NNW	10	17	0.159	0.687	Non significant
62.	processing_NNW	time_NNW	10	11	1.904	0.172	Non significant
63.	data_NNK	structure_NNW	9	29	1.617	0.233	Non significant
64.	data_NNK	process_NNW	9	12	0.842	0.362	Non significant
65.	features_NNY	different_AJK	9	15	0.18	0.669	Non significant
66.	goal_NNW	main_AJK	9	9	2.226	0.141	Non significant
67.	methods_NNY	different_AJK	9	33	2.712	0.108	Non significant
68.	networks_NNY	neural_AJK	9	4	7.496	0.014	Significant overuse
69.	parameters_NNY	values_NNY	9	24	0.56	0.575	Non significant
70.	project_NNW	management_NNW	9	9	2.226	0.141	Non significant
71.	resources_NNY	system_NNW	9	5	5.997	0.021	Significant overuse
72.	approach_NNW	linguistic_AJK	8	10	0.984	0.326	Non significant
73.	computer_NNW	systems_NNY	8	10	0.984	0.326	Non significant

74.	data_NNK	collection_NNW	8	26	1.493	0.277	Non significant
75.	data_NNK	video_NNW	8	10	0.984	0.326	Non significant
76.	data_NNK	systems_NNY	8	5	4.622	0.04	Significant overuse
77.	design_NNW	implementation_NNW	8	6	3.543	0.085	Non significant
78.	error_NNW	squared_AJK	8	4	5.964	0.027	Significant overuse
79.	function_NNW	cost_NNW	8	5	4.622	0.04	Significant overuse
80.	parameters_NNY	different_AJK	8	11	0.645	0.468	Non significant
81.	sequence_NNW	video_NNW	8	7	2.675	0.108	Non significant
82.	simulation_NNW	end_NNW	8	5	4.622	0.04	Significant overuse
83.	tasks_NNY	different_AJK	8	4	5.964	0.027	Significant overuse
84.	code_NNW	number_NNW	7	4	4.516	0.051	Significant overuse
85.	components_NNY	different_AJK	7	8	1.184	0.283	Non significant
86.	components_NNY	frequency_NNW	7	5	3.349	0.121	Non significant
87.	data_NNK	available_AJK	7	16	0.092	0.829	Non significant
88.	data_NNK	control_NNW	7	8	1.184	0.283	Non significant
89.	data_NNK	way_NNW	7	5	3.349	0.121	Non significant
90.	Document_NNW	query_NNW	7	33	4.554	0.047	Significant underuse
91.	function_NNW	system_NNW	7	5	3.349	0.121	Non significant
92.	functions_NNY	system_NNW	7	11	0.2	0.6	Non significant
93.	input_NNW	algorithm_NNW	7	4	4.5	0.05	Significant overuse
94.	method_NNW	class_NNW	7	32	4.1	0.04	Significant underuse
95.	parameter_NNW	values_NNY	7	13	0.02	1	Non significant
96.	parameters_NNY	other_AJK	7	9	0.76	0.4	Non significant
97.	process_NNW	model_NNW	7	4	4.5	0.05	Significant overuse
98.	project_NNW	system_NNW	7	7	1.7	0.2	Non significant
99.	allocation_NNW	resource_NNW	6	7	0.94	0.3	Non significant

10	allocation_NNW	process_NNW	6	5	2.2	0.19	Non significant
No.	**100NS N collocations**		**NS corpus frequency**	**RC frequency**	**Chi-square value**	**P value**	**Over/underuse significant**
1.	code_NNW	source_NNW	128	70	90.4	0.0001	Significant overuse
2.	data_NNK	test_NNW	50	34	26.9	0.0001	Significant overuse
3.	Design_NNW	system_NNW	50	13	61.6	0.0001	Significant overuse
4.	environment_NNW	development_NNW	50	6	80.6	0.0001	Significant overuse
5.	computer_NNW	vision_NNW	48	13	57.9	0.0001	Significant overuse
6.	process_NNW	development_NNW	46	8	66.8	0.0001	Significant overuse
7.	source_NNW	open_AJK	44	33	20.4	0.0001	Significant overuse
8.	data_NNK	database_NNW	42	7	61.9	0.0001	Significant overuse
9.	data_NNK	raw_AJK	42	6	64.8	0.0001	Significant overuse
10.	layer_NNW	application_NNW	42	10	53.8	0.0001	Significant overuse
11.	code_NNW	following_AJK	40	5	63.8	0.0001	Significant overuse
12.	data_NNK	structures_NNY	39	36	9.5	0.003	Significant overuse
13.	data_NNK	user_NNW	36	9	45.2	0.0001	Significant overuse
14.	data_NNK	type_NNW	36	7	50.3	0.0001	Significant overuse
15.	code_NNW	lines_NNY	34	52	1.7	0.2	
16.	data_NNK	storage_NNW	34	5	52	0.0001	Significant overuse
17.	data_NNK	structure_NN	32	29	10.5	0.001	Significant overuse

	W					
18.	section_NNW previous_AJK	32	32	8.4	0.004	Significant overuse
19.	task_NNW time_NNW	32	5	48.1	0.0001	Significant overuse
20.	computer_NNW science_NNW	30	14	24.8	0.0001	Significant overuse
21.	data_NNK time_NNW	30	9	34.2	0.0001	Significant overuse
22.	method_NNW class_NNW	30	32	6.7	0.013	Significant underuse
23.	data_NNK real_AJK	28	4	43.2	0.0001	Significant overuse
24.	document_NNW time_NNW	28	5	40.3	0.0001	Significant overuse
25.	data_NNK layer_NNW	26	21	10.6	0.001	Significant overuse
26.	data_NNK input_NNW	24	11	20.1	0.0001	Significant overuse
27.	data_NNK objects_NNY	24	11	20.1	0.0001	Significant overuse
28.	data_NNK amount_NNW	24	10	21.8	0.0001	Significant overuse
29.	data_NNK Web_NNW	24	7	27.8	0.0001	Significant overuse
30.	data_NNK Table_NNW	24	6	30.1	0.0001	Significant overuse
31.	file_NNW source_NNW	24	5	32.6	0.0001	Significant overuse
32.	data_NNK set_NNW	22	40	0.18	0.69	Non significant
33.	Design_NNW implementation_NNW	22	6	26.4	0.0001	Significant overuse
34.	site_NNW Web_NNW	22	37	0.5	0.49	Non significant
35.	area_NNW research_NNW	20	12	12.6	0.001	Significant overuse
36.	attribute_NNW value_NNW	20	4	27.6	0.0001	Significant overuse
37.	components_NN	20	8	18.8	0.0001	Significant overuse

	Y	different_AJK					
38.	layer_NNW	web_NNW	20	5	25.1	0.0001	Significant overuse
39.	location_NNW	users_NNY	20	4	27.6	0.0001	Significant overuse
40.	project_NNW	management_NNW	20	9	17	0.0001	Significant overuse
41.	resources_NNY	system_NNW	20	5	25	0.0001	Significant overuse
42.	section_NNW	following_AJK	20	16	8.3	0.006	Significant overuse
43.	tasks_NNY	number_NNW	20	5	25	0.0001	Significant overuse
44.	code_NNW	program_NNW	18	11	11	0.001	Significant overuse
45.	code_NNW	amount_NNW	18	6	19.2	0.0001	Significant overuse
46.	Design_NNW	systems_NNY	18	5	21.4	0.0001	Significant overuse
47.	resources_NNY	available_AJK	18	5	21.4	0.0001	Significant overuse
48.	section_NNW	model_NNW	18	19	4.1	0.05	Significant underuse
49.	data_NNK	sets_NNY	16	40	0.46	0.5	Non significant
50.	data_NNK	new_AJK	16	28	0.23	0.6	Non significant
51.	data_NNK	available_AJK	17	16	4.2	0.05	Significant oversue
52.	data_NNK	information_NNW	16	8	12.3	0.001	Significant overuse
53.	data_NNK	transfer_NNW	16	6	15.8	0.0001	Significant overuse
54.	framework_NNW	Eclipse_NNW	16	6	15.8	0.0001	Significant overuse
55.	framework_NNW	application_NNW	16	4	20	0.0001	Significant overuse
56.	project_NNW	different_AJK	16	7	14	0.0001	Significant overuse

57.	code_NNW example_NNW	14	4	16.4	0.0001	Significant overuse
58.	data_NNK different_AJK	14	15	3.1	0.11	Non significant
59.	data_NNK other_AJK	14	6	12.4	0.001	Significant overuse
60.	document_NNW user_NNW	14	6	12.4	0.001	Significant overuse
61.	environment_NNW Eclipse_NNW	14	4	16.4	0.0001	Significant overuse
62.	functions_NNY other_AJK	14	11	6	0.01	Significant overuse
63.	method_NNW following_AJK	14	4	16.4	0.0001	Significant overuse
64.	methods_NNY Class_NNW	14	35	0.4	0.6	Non significant
65.	methods_NNY different_AJK	14	33	0.2	0.7	Non significant
66.	section_NNW system_NNW	14	4	16.4	0.0001	Significant overuse
67.	structure_NNW system_NNW	14	4	16.4	0.0001	Significant overuse
68.	analysis_NNW Results_NNY	12	16	1.2	0.31	Non significant
69.	approach_NNW evolutionary_AJK	12	12	3.1	0.08	Non significant
70.	attributes_NNY number_NNW	12	36	1.3	0.2	Non significant
71.	data_NNK model_NNW	12	18	0.68	0.4	Non significant
72.	data_NNK results_NNY	12	11	3.8	0.07	Non significant
73.	data_NNK memory_NNW	12	10	4.6	0.04	Significant overuse
74.	data_NNK large_AJK	12	7	7.8	0.01	Significant overuse
75.	data_NNK functions_NNY	12	6	9.2	0.004	Significant overuse
76.	data_NNK	12	4	12.8	0.001	Significant overuse

		access_NNW					
77.	data_NNK	object_NNW	12	4	12.8	0.001	Significant overuse
78.	Design_NNW	architectural_AJK	12	47	4.2	0.05	Significant underuse
79.	environment_NNW	application_NNW	12	11	3.8	0.07	Non significant
80.	features_NNY	system_NNW	12	5	10.9	0.003	Significant overuse
81.	functions_NNY	different_AJK	12	12	3.1	0.08	Non significant
82.	functions_NNY	system_NNW	12	11	3.8	0.07	Non significant
83.	instances_NNY	different_AJK	12	9	5.5	0.03	Significant overuse
84.	project_NNW	software_NNW	12	5	10.9	0.003	Significant overuse
85.	source_NNW	information_NNW	12	14	2	0.2	Non significant
86.	source_NNW	system_NNW	12	4	12.8	0.001	Significant overuse
87.	attribute_NNW	name_NNW	10	5	7.7	0.01	Significant overuse
88.	code_NNW	number_NNW	10	4	9.4	0.004	Significant overuse
89.	data_NNK	analysis_NNW	10	16	0.36	0.5	Non significant
90.	data_NNK	communication_NNW	10	7	5.1	0.03	Significant overuse
91.	data_NNK	measurement_NNW	10	5	7.7	0.01	Significant overuse
92.	data_NNK	Additional_AJK	10	4	9.4	0.004	Significant overuse
93.	data_NNK	software_NNW	10	4	9.4	0.004	Significant overuse
94.	Design_NNW	detailed_AJK	10	13	1.16	0.2	Non significant

95.	Design_NNW software_NNW	10	4	9.4	0.004	Significant overuse
96.	features_NNY other_AJK	10	7	5.1	0.03	Significant overuse
97.	formats_NNY different_AJK	10	7	5.1	0.03	Significant overuse
98.	function_NNW simple_AJK	10	4	9.4	0.004	Significant overuse
99.	instance_NNW particular_AJK	10	6	6.3	0.01	Significant overuse
10(	network_NNW traffic_NNW	10	8	4.1	0.04	Significant overuse

No.	100NNS V collocations	NNS corpus frequency	Reference corpus frequency	Chi-square value	P value	Over/underuse significant
1.	obtained_VVN results_NNY	28	47	0.51	0.46	Non significant
2.	illustrates_VVZ Figure_NNW	16	18	2.8	0.1	Non significant
3.	extracted_VVN features_NNY	15	7	11.9	0.001	significant overuse
4.	illustrated_VVN Figure_NNW	11	12	2.14	0.18	Non significant
5.	required_VVN time_NNW	10	16	0.29	0.67	Non significant
6.	obtain_VVI information_NNW	9	9	2.2	0.14	Non significant
7.	obtained_VVN result_NNW	9	5	5.9	0.02	significant overuse
8.	achieve_VVI goal_NNW	8	4	5.9	0.02	significant overuse
9.	obtain_VVI results_NNY	6	8	0.56	0.57	Non significant
10.	achieve_VVI high_AJK	6	7	0.94	0.38	Non significant
11.	extracted_VVN data_NNK	6	4	3.1	0.09	Non significant
12.	achieved_VVN performance_NNW	5	9	0.3	1	Non significant
13.	defined_VVN time_NNW	5	5	1.2	0.3	Non significant
14.	defined_VVN Section_NNW	4	9	0.04	1	Non significant
15.	created_VVN object_NNW	4	7	0.04	1	Non significant
16.	affect_VVI performance_NNW	4	6	0.19	0.7	Non significant
17.	defined_VVN different_AJK	4	6	0.19	0.7	Non significant
18.	affected_VVN number_NNW	4	5	0.49	0.49	Non significant
19.	achieve_VVI system_NNW	4	4	0.98	0.45	Non significant
20.	implement_VVI easy_AJK	4	4	0.98	0.45	Non significant
21.	defined_VVN set_NNW	3	16	2.6	0.14	Non significant
22.	defined_VVN number_NNW	3	14	1.9	0.2	Non significant
23.	consists_VVZ set_NNW	3	11	0.9	0.4	Non significant
24.	defined_VVN model_NNW	3	11	0.9	0.4	Non significant
25.	defined_VVN ratio_NNW	3	11	0.9	0.4	Non significant
26.	obtained_VVN data_NNK	3	10	0.62	0.5	Non significant

27.	achieved_VVN results_NNY	3	8	0.18	0.76	Non significant
28.	illustrated_VVN Section_NNW	3	6	0	1	Non significant
29.	achieve_VVI performance_NNW	3	5	0.06	1	Non significant
30.	ensure_VVI system_NNW	3	5	0.06	1	Non significant
31.	occurs_VVZ problem_NNW	3	5	0.06	1	Non significant
32.	generate_VVI different_AJK	3	4	0.28	0.68	Non significant
33.	generated_VVN code_NNW	3	4	0.28	0.68	Non significant
34.	indicate_VVB performance_NNW	3	4	0.28	0.68	Non significant
35.	requires_VVZ approach_NNW	3	4	0.28	0.68	Non significant
36.	obtained_VVN values_NNY	2	14	3.1	0.1	Non significant
37.	consists_VVZ classes_NNY	2	10	1.5	0.3	Non significant
38.	requires_VVZ method_NNW	2	9	1.1	0.3	Non significant
39.	required_VVN information_NNW	2	8	0.8	0.5	Non significant
40.	defines_VVZ set_NNW	2	7	0.5	0.72	Non significant
41.	implemented_VVN Java_NPK	2	6	0.25	0.7	Non significant
42.	obtained_VVN different_AJK	2	6	0.25	0.7	Non significant
43.	select_VVI appropriate_AJK	2	6	0.25	0.7	Non significant
44.	defined_VVN method_NNW	2	5	0.07	1	Non significant
45.	generated_VVN different_AJK	2	5	0.07	1	Non significant
46.	obtain_VVI data_NNK	2	5	0.07	1	Non significant
47.	assigned_VVN set_NNW	2	4	Could not be performed(no.47-100)		
48.	create_VVI structure_NNW	2	4	Could not be performed		
49.	identify_VVI able_AJK	2	4	Could not be performed		
50.	obtained_VVN following_AJK	2	4			
51.	obtained_VVN information_NNW	2	4			
52.	occurred_VVD changes_NNY	2	4			
53.	require_VVB information_NNW	2	4			
54.	required_VVN additional_AJK	2	4			

55.	obtained_VVN	set_NNW	1	10			
56.	obtained_VVN	performance_NNW	1	9			
57.	generated_VVN	set_NNW	1	8			
58.	obtained_VVN	previous_AJK	1	8			
59.	define_VVB	following_AJK	1	7			
60.	obtain_VVI	possible_AJK	1	7			
61.	achieve_VVI	objectives_NNY	1	6			
62.	defined_VVN	ability_NNW	1	6			
63.	defined_VVN	class_NNW	1	6			
64.	defined_VVN	function_NNW	1	6			
65.	defined_VVN	object_NNW	1	6			
66.	defined_VVN	problem_NNW	1	6			
67.	defined_VVN	variables_NNY	1	6			
68.	illustrated_VVN	example_NNW	1	6			
69.	illustrates_VVZ	example_NNW	1	6			
70.	occurs_VVZ	term_NNW	1	6			
71.	requires_VVZ	knowledge_NNW	1	6			
72.	affect_VVB	factors_NNY	1	5			
73.	created_VVN	structure_NNW	1	5			
74.	creates_VVZ	new_AJK	1	5			
75.	defined_VVN	node_NNW	1	5			
76.	defined_VVN	task_NNW	1	5			
77.	generate_VVI	data_NNK	1	5			
78.	indicate_VVB	values_NNY	1	5			
79.	obtain_VVI	analysis_NNW	1	5			
80.	required_VVN	effort_NNW	1	5			
81.	requires_VVZ	problem_NNW	1	5			
82.	achieved_VVN	number_NNW	1	4			
83.	achieved_VVN	precision_NNW	1	4			

84.	achieved_VVN	rate_NNW	1	4			
85.	consists_VVZ	process_NNW	1	4			
86.	created_VVN	data_NNK	1	4			
87.	define_VVI	set_NNW	1	4			
88.	defined_VVN	way_NNW	1	4			
89.	generate_VVI	model_NNW	1	4			
90.	implemented_ VVN	analysis_NNW	1	4			
91.	implemented_ VVN	method_NNW	1	4			
92.	implemented_ VVN	systems_NNY	1	4			
93.	indicated_VVN	section_NNW	1	4			
94.	obtain_VVI	able_AJK	1	4			
95.	obtain_VVI	difficult_AJK	1	4			
96.	obtain_VVI	new_AJK	1	4			
97.	obtained_VVN	sample_NNW	1	4			
98.	occur_VVI	errors_NNY	1	4			
99.	required_VVN	test_NNW	1	4			
100.	ensures_VVZ	quality_NNW	1	2			

No.	**100NS V collocations**		**NS corpus frequency**	**Reference corpus frequency**	**Chi-square value**	**P value**	**Over/underuse significant**
1.	defined_VVN	Section_NNW	32	9	37	0.0001	Significant overuse
2.	ensure_VVI	system_NNW	26	5	36	0.0001	Significant overuse
3.	created_VVN	new_AJK	24	8	25	0.0001	Significant overuse
4.	created_VVN	object_NNW	16	7	14	0.0001	Significant overuse
5.	affect_VVI	performance_NNW	14	6	12.6	0.001	Significant overuse
6.	extracted_VVN	data_NNK	14	4	16.4	0.0001	Significant overuse

7.	defined_VVN	model_NNW	12	11	3.8	0.07	Non significant
8.	found_VVN	solution_NNW	12	11	3.8	0.07	Non significant
9.	required_VVN	information_NNW	10	8	4.1	0.04	Significant overuse
10.	creates_VVZ	new_AJK	10	5	7.7	0.01	Significant overuse
11.	required_VVN	work_NNW	10	5	7.7	0.01	Significant overuse
12.	implemented_VVN	method_NNW	10	4	9.4	0.004	Significant overuse
13.	demonstrates_VVZ	Section_NNW	10	4	9.4	0.004	Significant overuse
14.	defined_VVN	ratio_NNW	8	11	0.72	0.46	Non significant
15.	detect_VVI	able_AJK	8	6	3.7	0.08	Non significant
16.	found_VVN	algorithm_NNW	8	4	6.1	0.02	Significant overuse
17.	created_VVN	data_NNK	8	4	6.1	0.02	Significant overuse
18.	required_VVN	number_NNW	6	16	0.31	0.65	Non significant
19.	required_VVN	time_NNW	6	16	0.31	0.65	Non significant
20.	consists_VVZ	classes_NNY	6	10	0.17	0.7	Non significant
21.	requires_VVZ	method_NNW	6	9	0.34	0.5	Non significant
22.	removed_VVN	code_NNW	6	8	0.6	0.4	Non significant
23.	defines_VVZ	section_NNW	6	8	0.6	0.4	Non significant
24.	defined_VVN	problem_NNW	6	6	1.5	0.22	Non significant
25.	selected_VVN	points_NNY	4	11	0.26	0.7	Non significant
26.	consists_VVZ	set_NNW	4	11	0.26	0.7	Non significant
27.	achieved_VVN	performance_NNW	4	9	0.02	1	Non significant
28.	found_VVN	solutions_NNY	4	9	0.001	1	Non significant
29.	achieved_VVN	results_NNY	4	8			Non significant
30.	required_VVN	Application_NNW	4	6	0.22	0.7	Non significant
31.	defined_VVN	class_NNW	4	6	0.22	0.7	Non significant
32.	defined_VVN	function_NNW	4	6	0.22	0.7	Non significant
33.	selected_VVN	point_NNW	4	6	0.22	0.7	Non significant
34.	generate_VVI	test_NNW	4	6	0.22	0.7	Non significant
35.	generate_VVI	data_NNK	4	5	0.5	0.4	Non significant

36.	defined_VVN	task_NNW	4	5	0.5	0.4	Non significant
37.	identify_VVI	able_AJK	4	4	1	0.45	Non significant
38.	implemented_VVN	algorithm_NNW	4	4	1	0.45	Non significant
39.	occur_VVI	errors_NNY	4	4	1	0.45	Non significant
40.	achieve_VVI	goal_NNW	4	4	1	0.45	Non significant
41.	achieve_VVI	objective_NNW	4	4	1	0.45	Non significant
42.	focuses_VVZ	paper_NNW	4	4	1	0.45	Non significant
43.	implemented_VVN	systems_NNY	4	4	1	0.45	Non significant
44.	ensure_VVI	process_NNW	4	3	1.8	0.22	Non significant
45.	defined_VVN	number_NNW	2	14	3	0.1	Non significant
46.	indicate_VVB	results_NNY	2	12	2	0.16	Non significant
47.	demonstrate_VVB	results_NNY	2	10	1.4	..35	Non significant
48.	define_VVB	following_AJK	2	7	0.4	0.7	Non significant
49.	achieve_VVI	high_AJK	2	7	0.4	0.7	Non significant
50.	defined_VVN	Objects_NNY	2	7	0.4	0.7	Non significant
51.	found_VVN	set_NNW	2	7	0.4	0.7	Non significant
52.	created_VVN	activity_NNW	2	6	0.22	1	Non significant
53.	select_VVI	appropriate_AJK	2	6	0.22	1	Non significant
54.	conducted_VVD	experiments_NNY	2	6	0.22	1	Non significant
55.	implemented_VVN	Java_NPK	2	6	0.22	1	Non significant
56.	required_VVN	minimum_AJK	2	6	0.22	1	Non significant
57.	defined_VVN	object_NNW	2	6	0.22	1	Non significant
58.	achieve_VVI	objectives_NNY	2	6	0.22	1	Non significant
59.	defined_VVN	schema_NNW	2	6	0.22	1	Non significant
60.	conducted_VVN	Study_NNW	2	6	0.22	1	Non significant
61.	defined_VVN	variables_NNY	2	6	0.22	1	Non significant
62.	required_VVN	changes_NNY	2	5	0.05	1	Non significant

63.	generated_VVN different_AJK	2	5	0.05	1	Non significant
64.	defined_VVN features_NNY	2	5	0.05	1	Non significant
65.	defined_VVN node_NNW	2	5	0.05	1	Non significant
66.	achieve_VVI performance_NNW	2	5	0.05	1	Non significant
67.	found_VVN similar_AJK	2	5	0.05	1	Non significant
68.	created_VVN structure_NNW	2	5	0.05	1	Non significant
69.	conducted_VVN test_NNW	2	5	0.05	1	Non significant
70.	defined_VVN time_NNW	2	5	0.05	1	Non significant
71.	required_VVN additional_AJK	2	4	Could not be performed(no.71-100)		
72.	implemented_VVN analysis_NNW	2	4	Could not be performed		
73.	requires_VVZ approach_NNW	2	4	Could not be performed		
74.	involve_VVB changes_NNY	2	4			
75.	occurred_VVD changes_NNY	2	4			
76.	ensure_VVI consistency_NNW	2	4			
77.	generate_VVI different_AJK	2	4			
78.	generated_VVN event_NNW	2	4			
79.	consists_VVZ files_NNY	2	4			
80.	found_VVN files_NNY	2	4			
81.	require_VVB information_NNW	2	4			
82.	generated_VVN initial_AJK	2	4			
83.	found_VVN method_NNW	2	4			
84.	calls_VVZ methods_NNY	2	4			
85.	require_VVB methods_NNY	2	4			
86.	involves_VVZ model_NNW	2	4			
87.	required_VVN model_NNW	2	4			
88.	select_VVI points_NNY	2	4			
89.	achieved_VVN precision_NNW	2	4			
90.	consists_VVZ process_NNW	2	4			

91.	achieved_VVN	rate_NNW	2	4			
92.	specify_VVI	rules_NNY	2	4			
93.	achieved_VVN	score_NNW	2	4			
94.	demonstrated_ VVN	Section_NNW	2	4			
95.	define_VVI	set_NNW	2	4			
96.	required_VVN	test_NNW	2	4			
97.	removed_VVN	time_NNW	2	4			
98.	consisting_VV G	tuple_NNW	2	4			
99.	defined_VVN	way_NNW	2	4			
100.	defined_VVN	Section_NNW	32	9			

Appendix E: Dictionaries Check for the 49 N Collocations

collocations			
	Online CS dictionary(WHATIS.COM)	BBI dictionary	**GAC** **GCSC**
1-code following	Not found	Not found	ASK CS EXPERT
2-data input	yes		GCSC
3-data access	yes		GCSC
4-data user	yes		GCSC
5-data information	yes		GCSC
6-data type	Yes		GCSC
7-design system	yes		GCSC
8-environment development	Yes		GCSC
9-layer application	Yes		GCSC
10-network traffic	yes		GCSC
11-resources available	yes	yes	GAC
12-resources system	yes		GCSC
13-code source	yes		GCSC

14-data layer	yes		GCSC
15-data available	Not found	Not found	AskCS expert
16-previous section	Not found	Not found	AskCS expert
17-following section	Not found	Not found	Ask CS expert
18-Web site	yes	yes	GAC
19-Open source	Yes		GCSC
20-different components	Not found	Not found	Ask CSexpert
21-simulation results	Yes		GCSC
22-data structure	yes		GCSC
23-error rate	Yes	Yes	GAC
24-extraction information	yes		GCSC
25- dynamic allocation	Yes		GCSC
26-data training	yes		GCSC
27-data test	Yes	Yes	GAC
28-computer vision	Yes		GCSC

29-process development	Yes		GCSC
30-data database	yes		GCSC
31-data raw	yes	yes	GAC
32-design architectural	yes		GCSC
33-design implementation	Not found	Not found	askCSexpert
34-vulnerable files	Not found	Not found	Ask CS expert
35-function ranking	Not found	Not found	Ask CS expert
36-document ranking	Not found	Not found	askCS expert
37-document scope	Not found	Not found	askCSexpert
38- neutral files	Not found	Not found	Ask CS expert
39-document cohesion	Not found	Not found	Ask CS expert
40- data time	Yes	Yes	GAC
41- number code	Yes	yes	GAC
42-amount data	Not found	Not found	AskCS expert
43- other features	Not found	Not found	askCSexpert

44- other data	Not found	Not found	Ask CSexpert
45-data web	Not found	Not found	askCSexpert
46-method class	Not found	Not found	askCSexpert
47-model section	Not found	Not found	askCSexpert
48-data different	Not found	Not found	Ask CS expert
49-query document	Not found	Not found	askCSexpert

Appendix F: Re-test Significant Results of the 30 Shared N Collocations

NO.	NNS N collocations	NNS freq	RC freq	Chi-square value	P value	Significant Over/underuse
1.	Code following	8.00	3.00	7.641	.006	Significant overuse
2.	code source	11.00	68.00	13.491	.000	Significant underuse
3.	data access	19.00	3.00	27.723	.000	Significant overuse
4.	environment development	12.00	5.00			No need to retest
5.	design system	16.00	7.00	13.510	.000	Significant overuse
6.	data time	4.00	4.00	.989	.320	nonsignificant
7.	data amount	25.00	9.00	.559	.455	Nonsignificant
8.	data other	5.00	5.00			No need to retest
9.	data information	11.00	5.00	.034	.859	Non significant
10.	data layer	26.00	21.00	2.682	.101	Non significant
11.	data type	9.00	7.00	3.750	.053	Significant overuse
12.	data user	11.00	4.00	10.744	.001	Significant overuse
13.	features other	7.00	6.00	2.439	.118	Non significant
14.	layer application	13.00	9.00			No need to retest
15.	merhod class	1.00	13.00	4.343	.037	Significant useruse
16.	network traffic	55.00	8.00	82.235	.000	Significant overuse
17.	resources available	13.00	4.00	14.165	.000	Significant overuse
18.	resources system	4.00	4.00	.989	.320	Non significant
19.	Section previous	27.00	33.00			No need to retest
20.	Data Different	10.00	8.00	3.966	.046	Significant overuse
21.	data input	11.00	11.00	2.72	0.114	no need to retest
22.	data structure	9.00	29.00	1.617	0.233	no need to retest
23.	data available	7.00	16.00	0.092	0.829	Non significant

24.	Design implementation	8	6	3.543	0.085	Non significant
25.	section previous	15	32	0.048	0.87	no need to retest
26.	section following	12	19	0.391	0.569	no need to retest
27.	site Web	30	37			No need to retest
28.	source open	19	33	0.228	0.66	No need to retest
29.	components different	7	8	1.184	0.283	No need to retest
30.	data different	26	15	16.58	0.0001	Significant overuse

No.	NS N collocations	NS Freq	RC Freq	Chi-square value	P value	Significant Over/underuse
1.	code_following	36.00	3.00	62.345	.000	Significant overuse
2.	code_source	126.00	68.00	90.277	.000	Significant overuse
3.	data_access	10.00	3.00	11.411	.001	Significant overuse
4.	environment_ development	50.00	5.00	83.834	.000	Significant overuse
5.	designsystem	41.00	7.00	59.96	.000	Significant overuse
6.	datatime	20.00	4.00	27.649	.000	Significant overuse
7.	Data amount	24.00	9.00			No need to retest
8.	Data other					Nonsignificant
9.	datainformation	6.00	5.00	2.33	.127	Non significant
10.	Data layer	26.00	21.00			No need to retest
11.	datatype	26.00	7.00	31.475	.000	Significant overuse
12.	Data user	14.00	4.00	16.410	.000	Significant overuse
13.	Data web					Non significant
14.	Features other	10.00	7.00			No need to retest
15.	layerapplication	38.00	9.00	48,952	.000	Significant overuse
16.	Code number					Non significant
17.	method class	4.00	13.00	.677	.411	Non significant
18.	network traffic	9.00	8.00	3.092	.07	Non significant
19.	resources availble	12.00	4.00	12.844	.000	Significant overuse
20.	resources system	10.00	4.00	9.412	.002	Significant overuse
21.	Data input	22.00	9.00	20.347	.000	Significant overuse
22.	Data structure	30.00	26.00	10.83	.001	Significant overuse
23.	Data available	8.00	10.00	11.086	.29	Non significant
24.	Section previous	27.00	33.00	3.977	.046	Significant underuse
25.	Section following	20.00	14.00	10.348	.001	Significant overuse
26.	Different components	20.00	6.00	22.822	.000	Significant overuse

27.	Web site					No neede to retest
28.	Open source					No need to retest
29.	Data different					Excluded
30.	Design implemnation					Excluded
31.	Data different					Excluded since all RC concordance lines deleted
32.	Design implantation					Excluded since all RC concordance lines excluded

Appendix G:Categorsaition Judgment Task(CJT)

Categorisation Judgement for Phrases used in Computer Science

I am working on the use of academic phrases (words that usually occur together, e.g. ***data access, network traffic, layer application***) in the writing of Computer Science postgraduate students. I have compared the most frequent academic phrases used by Computer Science students with their use in expert writing (Computer Science journal articles). My results reveal that some phrases were used more/ less by students compared to the experts. I now wish to find out more about these phrases from Computer Science specialists. Thus, I would be grateful for your views about whether the phrases I have found in the students' writing are:

d- **General academic phrases** (these phrases can be found in Computer Science as well as in other academic disciplines, e.g. *available data, different components*)
e- **General Computer Science (CS) academic phrases** (these phrases can be found in Computer Science only, but in ANY discipline of Computer Science, e.g. data input)
f- **Specific Computer Science (CS) academic phrase** (these phrases can be found in Computer Science only, but can only be found in certain disciplines of Computer Science, e.g. *network traffic: in the sub disciplines of software engineering and information systems only*)

I would ask you to judge each phrase under one of these three categories by ticking the appropriate column. I would also be grateful for any additional comments or observations you have about these phrases which come to mind. Examples are given in the table below:

phrases	**General academic phrase**	**General CS academic phrase**	**Specific CS academic phrase**			**Comments**
			Artificial Intelligence	**Software Engineering**	**Information Systems**	
1-Available data	√					
2-Data input		√				
3-netwrok traffic				√	√	A very common phrase in some types of CS.

In the first example, the Computer Science specialist felt that the phrase *available data* can be found in ANY or ALL disciplines, not only Computer Science, and so s/he ticked the '**General Academic phrase'** box.

In the second example, the Computer Science specialist felt that the phrase *data input* is a phrase used in Computer Science only and can be used in ANY discipline of Computer Science, and so s/he ticked the '**General CS academic phrase'** box.

In the third example, the Computer Science specialist felt that the phrase *network traffic* is a phrase used only in SPECIFIC types of Computer Science, Software Engineering and Information Systems, but not in Artificial Intelligence. S/he has also added a comment, saying *network traffic* is very common in certain fields of Computer Science.

Many thanks for your help; it's greatly appreciated.

A-Adjacent Phrases: (these phrases occur usually next to each other)

Example: data input/input data

As shown in the following extracts from Computer Science students' writing:

1- "…first voting pass" is applied to process raw **input data** to detect structures and outliers.

2- …despite the fact that the **input data** are highly corrupted. Our algorithm is called …

Phrases with adjacent words	**General academic phrases**	**General CS academic phrases**	**Specific CS academic phrases**			
			Artificial Intelligence	**Software Engineering**	**Information System**	**Comments**
1-code following/ following code						
2-data input/ input data						
3-data access						
4-data user						
5-data information/ information data						
6-data type						
7-design system/ system design						
8-environment development/ development environment						
9-layer application/ application layer						
10-network traffic/ traffic network						

11-resources available/ available resources						
12-resources system/ system resources						
13-code source / source code						
14-data layer						
15-data available/ available data						
16-previous section						
17-following section						
18-Web site/Site Web						
19-Open source						
20-components different/ different components						
21-simulation results						
22-data structure/s						
23-error rate						
24-extraction information						
25-allocation dynamic/ dynamic allocation						

26-data training						
27-data test						
28-computer vision						
29-process development						
30-data database						
31-data raw/ raw data						
32-design architectural						
33-design implementation						
34-files vulnerable / vulnerable files						
35-function ranking						
36-document ranking						
37-document scope						
38-files neutral/ neutral files						
39-document cohesion						

B-Categorisation judgement for phrases with non-adjacent words

This task is focused on a different set of phrases that occur with non-adjacent words. That is, these two words occur together but they are separated by other words. In this task, each phrase is presented individually with two examples given. Please read the two examples carefully and then judge the phrase either as a general academic phrase, general Computer Science (CS) academic phrase or specific Computer Science academic phrase and tick the boxes as appropriate. I would be grateful for any additional comments or observations you have about these phrases which come to mind.

1-data …time/ time…data

As shown in the following extracts from Computer Science students' writing:

1- Using the unbalanced ***data*** and reduced the ***time*** for evaluation. Figs… provide the pseudo code…
2- Mozilla Firefox had 34 releases at the ***time*** of ***data*** collection developed over four years.

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

2-number…code

As shown in the following extracts from Computer Science students' writing:

1- The ***number***of lines of ***code*** written....
2- The ***number*** of low-level machine ***code***instructions.

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

3-amount …data

As shown in following extracts from Computer Science students' writing:

1- We computed the relative difference in the ***amount*** of transferred ***data*** between corresponding…
2- That is, query views increased the ***amount*** of ***data*** transferred in some cases**.**

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

4- other…features

As shown in the following extracts from Computer Science students' writing:

1- Snort has ***other*** important ***features*** such as pre-processors and…
2- ***Other*** improvements for AR ***features*** **i**nclude autoregressive frequency.

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

5-other …data

As shown in the following extracts from Computer Science students' writing:

1- Passwords and ***other***insecure ***data***Snort's pre-processor receives…
2- There are ***other*** types of ***data***traffic including database…

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

6-data …web

As shown in the following extracts from Computer Science students' writing:

1- …is a type of unwanted ***data***available on ***web*** pages.
2- Such pages displaying dynamic ***data*** are known as deep ***Web…***

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

7- method …class

As shown in the following extracts from Computer Science students' writing

1-…to the document using the ***method***of the DataSet ***class***.

2-…It uses the ***method*** from the Membership ***class…***

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

8- model …section

As shown in the following extracts from Computer Science students' writing:

1- In the complete touch event ***model***in ***section***…
2- … are defined in both earlier ***models***in ***section***…

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

9-data …different

As shown in the following extracts from Computer Science students' writing:

1- The length of the transmitted ***data*** can be ***different*** from…
2- Applications that can exchange ***data*** between their ***different***…

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

10-query …document

As shown in the following extracts from Computer Science students' writing:

1- The ***query*** and the specified ***document***in order to obtain the…
2- The ***query*** term matches the ***document***that contains the different…

General academic phrase	General CS academic phrase	Specific CS academic phrase			Comments
		Artificial Intelligence	**Software engineering**	**Information Systems**	

Thank you very much for your cooperation. Please feel free to contact me or my supervisors should you have any questions about my work.

Afnan Farooqui

PhD student

Department of Language and Linguistics

asfaro@essex.ac.uk

Supervisors:

Nigel Harwood nharwood@essex.ac.uk

Sophia Skoufaki sskouf@essex.ac.uk

Appendix H: Topics Checks for Some of the 49 N Collocations

Overused NNS Collocations	RC freq	NNS freq	NNS Dissertations	RC files
network traffic	8	70	1SE ,20IS,21IS,22IS ,6IS ,10IS,30IS	3 SE, 5IS,20SE
simulation results	5	56	15IS,17AI,20IS,22IS,6IS,7IS,10IS ,24IS,29IS,30IS	3 SE ,3AI,13IS
sites web	29	39	23AI,28SE,14SE,16SE	37 AI , 6 IS ,7IS
error rate	9	32	13IS,23AI,26IS,29IS	5 IS,33AI,30AI
extraction information	3	28	23AI,18AI	17 IS,5SE
allocation dynamic	4	27	13 IS,19IS	6IS,3AI,13IS
data layer	21	26	3SE	5SE,7SE
data different	15	26	11IS,3SE,6IS,4IS	13 SE, 1 SE ,38 AI
data amount	10	26	4SE,28SE,19AI	12SE,15AI,11IS
data access	4	19	3 SE	8 SE,13IS

Underused NNS collocations	RC freq	NNS freq	NNS dissertations	RC topics
site web	37	30	23AI,28SE,3SE,14SE,16SE	37AI, 1IS,3IS,6IS

data training	70	18	1SE,2AI,8AI,23AI	1AI,14AI,18AI,28AI ,36AI,38AI, 9IS,13IS,16IS, I SE,13SE,15SE
code source	70	11	7IS,16SE,14SE,17AI	2IS,4SE,5SE,10SE,1 5SE,16SE,19SE
document query	33	7	23AI,14SE,15SE,16SE	2IS,37AI,16IS
method class	32	7	8AI,2AI,17AI	19SE

overused NS collocations	**RC freq**	**NS freq**	**NS dissertations**	**Rc**
code source	70	128	2AI,3AI,4SE,5SE,7IS,10SE,14SE 15SE,16SE	2IS,4SE,5SE,10SE,1 5SE,16SE,19SE
data test	34	50	3AI,4SE,7IS,9IS,22SE,26SE	18AI ,15 SE
design system	13	50	3AI,4SE,11AI,12IS,14SE	2SE,18IS,6AI
environment development	6	50	4SE,22SE,3AI	16 SE,13SE
computer vision	13	48	4SE,9IS,13AI	15AI, 23AI,30AI
process development	8	46	2AI,4SE,5SE,7IS,10SE,12IS,13A I,19SE, 15SE,22SE	2 IS ,1SE,13SE,15SE,1 6SE, 18SE,19SE
source open	33	44	2AI,3AI,8AI,9IS,19SE,10SE,13A I,15SE,16SE,21IS,26SE	5 SE,4AI,9IS,

data database	7	42	9IS ,10AI	ISE,18SE,6SE
data raw	6	42	4SE,9IS,15SE	2AI,4SE,3AI,19SE
layer application	10	42	16 SE,22SE	1SE, 13IS,19SE,23AI

Appendix I: E-mail Template asking CS experts for Participation

Dear Dr. …,

I am a PhD student in the Department of Language and Linguistics. My research examines Computer Scientists' writing; specifically, it involves a comparison of Native and Non-Native students' use of academic phrases in their MSc dissertations with that of expert writers in journal articles. I have conducted my research looking at three Computer Science sub-disciplines: Artificial Intelligence, Software Engineering and Information Systems. My results reveal that students use certain academic phrases less/more than expert writers do. I would now like to find out more about these phrases from Computer Scientists like yourself; I am contacting you since you are a specialist in one of these sub-disciplines. I would therefore be very grateful if you could participate in my research, which would involve a categorisation judgment task and a follow-up interview.

The categorisation judgment task involves categorising 59 phrases found in the students' writing as'General academic phrases' (that is, phrases used in various academic disciplines), 'General Computer Science academic phrases' (that is, phrases used in all Computer Science sub-disciplines), or 'Specific Computer Science academic phrases' (that is, phrases used only in some Computer Science sub-disciplines). In the follow-up interview interviewees will talk in more detail about their answers in the categorization judgment task, The judgment task will last 15-20 minutes and the follow-up interview 50-80 minutes.

This project has been approved by my department's Ethics Officer.Of course participationis voluntary, and participants are free to withdraw from the study at any point in time. When I publish the findings of my research, participants' identities will not be revealed.

I would be most grateful for your participation.If you would like to participate or have any questions about thejudgmenttaskor follow-up interview,pleasedo not hesitate to contact me at asfaro@essex.ac.uk

Kind regards,

Afnan Farooqui

PhD Candidate
Language and Linguistics Department
University of Essex

Appendix J: Semi-structured In-depth Interviews with CS Experts.

Computer Science Informants Interview

This interview is related to the judgment task you have already kindly completed. First, I would like to ask you some general questions about the dissertation requirements and IELTS. Then, I will ask you more detailed questions about your comments on the judgement task.

A-General Questions:

First, I have some questions about CS dissertations and the English language requirements of the department…

1- I have checked the CS website and found that MSc dissertations should be between 50-60 pages. What is the minimum and maximum number of words for a dissertation in your department?
2- To what extent do you think that non-natives' level of language proficiency is near or similar to the native speaker students?

B- Collocation Use (detailed questions about the Judgment task)

Now I'd like to ask you some questions about the judgement task…

Having looked at your comments on the judgement task, you identified some of the phrases as general/ specific CS academic phrases and others as general academic phrases .

" if we just start by looking at the (…) ,you wrote….."

1-Are there any more comments you would like to make?

2-Why do you think it is considered a specific academic phrase for Computer Science only?

3-I looked up the meaning of (…) in two dictionaries and found (show meaning on prompt cards) that, according to the dictionaries, it can be classified as a general academic phrase. But I see you marked this as a specific academic phrase for Computer Science only. Can you please comment on this?

C- Detailed Questions about some of the 24 shared phrases (use and patterns)

My results reveal that some of these collocations were used differently by Non-Native & Native students compared to expert writers—by experts, I mean Computer Scientists writing journal articles. I would like to investigate this point in more detail:

Questions will be asked on prompt cards

prompt card (1): **Environment development/development environment:**

In the table below, the 'Expert' column refers to Computer Science journal articles;

The 'Non-Native Writers' column is non-native Computer Science student writers;

The 'Native Writers' column is native Computer Science student writers.

The 'Normalised Frequency' column shows the frequency with which the expert and student writers use the phrase '*development environment*' per 100,000 words.

The 'No. of users' column shows how many writers used this phrase. So for instance, 1 out of 63 writers from the experts writing journal articles used the phrase *development environment*.

Corpus	Development environment	
	Normalised frequency (NF)	No. of users
Expert writers(journal articles)	.83	1/63
Non-Native writers	3.9	8/29
Native writers	16.9	12/26

1-You can see from the table that both native and non-native students use *environment development* more than the expert computer scientists writing journal articles. Could you comment on that?

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?"

3-To what extent do you think that the dissertation or journal article **topic** might affect writers' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

topics:

NS 1 : Implementation of Game Agents in Unreal Tournaments)

NS13 : Mobile Phone Training for the Elderly People)

NNS14 : Advanced Web Application Programming)

NNS23 : Intelligent Web Search Using Named Entity Recognition)

RC 16SE: A logical verification methodology for service-oriented computing.

4-Are there any other comments you would like to add about this phrase?

Prompt card (2): Following code

In the table below, the 'Expert' column refers to Computer Science journal articles;

The 'Non-Native Writers' column is non-native Computer Science student writers;

The 'Native Writers' column is native Computer Science student writers.

The 'Normalised Frequency' column shows the frequency with which the expert and student writers use this phrase per 100,000 words.

The 'No. of users' column shows how many writers used this phrase. So for instance, 1 out of 63 writers from the experts writing journal articles used the phrase *the following code*.

	Patterns for ' following code'	
Corpus	The following code	
	Normalised frequency	No.of users
Expert writers (journal articles)	0.5	1/63
Non-Native writers	2.7	2/29
Native writers	8.8	5/26

1-You can see from the table that both native and non-native students use *the following code* more than the expert computer scientists writing journal articles. Could you comment on that?

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?"

3-To what extent do you think that the dissertation or journal article **topic** might affect writers' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

Topics :

NS 27: minimum spanning tree with uncertainty.

NS 28: Optimising for High-Performance Cache Utilisation

NNS 11: Optical Information System.

NNS 14 : Advanced Web Application Programming

RC 19 SE: DARWIN : an approach to debugging evolving programs.

4. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

Prompt card (3):method class

	Patterns for ‘ method class ’			
Corpus	Class method		Method prp class	
	Normalised Frequency	NO.of users	Normalised Frequency	NO.of users
Expert writers(journal articles)	0.5	1/63	1.5	3/63
Non-native writers			0.33	1/29
Native writers	1.3	1/26		

1-You can see from the table that native students use *class method* more than the expert computer scientists writing journal articles. Could you comment on that?

While non-native students use method prp class less than the expert computer scientists writing journal articles. Could you comment on that?

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?”

3-To what extent do you think that the dissertation or journal article **topic** might affect writers’ use of this academic phrase? I’d like to show you some of the students’ dissertation topics and the expert writers’ journal article topics and ask what you think:

-**Topics:**

NNS 18: Web Summarization Searches

NS 7: Mobile Development

RC 15 ,16 se(2^{nd} p) : a framework for the checking and refactoring of crosscutting concepts

(16 se): a logical verification methodology for service-oriented computing

19se (1) DARWIN:an approach to debugging evolving programs

18is(2&3) Information Technology Implementers’ responses to User Resistance: Nature and Effects.

4. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

Prompt card (4): source code

	patterns for 'source code'	
Corpus	Source code	
	normalised Frequency	No. of users
Expert writers (journal articles)	11	7/63
Non-native writers	3.6	4/29
Native writers	42.8	9/26

1-You can see from the table that native and non-native students use *source code different* than the expert computer scientists writing journal articles. Could you comment on that?

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?"

3-To what extent do you think that the dissertation or journal article **topic** might affect writers' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

Topics:

NS 2 : Intelligent system and robotics

NS 5 : Web Application Programming

NNS 14: Advanced Web Application Programming

NNS17: Intelligent Control of an Unmanned Aerial Vehicle

RC (15 se): a framework for the checking and refactoring of crosscutting concepts

(16 se): a logical verification methodology for service-oriented computing.

4. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. **native and non-native writers use language differently;**

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

Prompt card (5): data type

	Patterns for collocation 'data type '			
Corpus	Data type		Type of data	
	Normalised frequency	no.of users	Normalised frequency	No of users
RC	0.5	2/63	0.5	2/63
NNS	0.33	1/29	1.6	4/29
NS	7.4	6/26	1.3	2/26

1-You can see from the table that only non-native students use *data type more* than the expert computer scientists writing journal articles. Could you comment on that?

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?

3-To what extent do you think that the dissertation or journal article topic might affect writers' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

Topics:

NS 8:intelligent system and robotics

NS 1: Implementation of Game Agents in Unreal Tournament

4. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

Prompt card (6):Data input

	Patterns for ' input data '							
Corpus	Input data		Input (n)data		Data input		Input of raw data	
	Normalised Frequency	NO.of users	normalised frequency	No.of users	Normalised Frequency	No. Of users	normalised frequency .	No. of users
expert writers (journal articles	1.3	3/63	0.16	1/63				
Non-Native writers								
Native writers	7.5	5/26			0.68	1/26	0.68	1/26

1-"You can see from the table that only native students use *input data more* than the expert computer scientists writing journal articles. Could you comment on that?"

2- To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?

3-To what extent do you think that the dissertation or journal article topic might affect either students' or experts' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

-**Topics:**

NS 1: Implementation of Game Agents in Unreal Tournament

NS 3: The development of a negotiation system using software agents to attempt to resolve the irregularities associated with the transfer of Professional Football Players.(E-commerce technology)

RC(15 AI): Similarity measure for anomaly detection and comparing human behaviours.

(17AI) : Text summarization contribution to semantic question answering: New approaches for finding answers on the web.

4. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

Prompt card (7):Section following / Section previous

Corpus	Following section	
	Normalised Frequency	No. of users
experts	2.33	7/63
Non-native writers	0	
Native writers	6.8	4/26

Corpus	Previous section	
	Normalised frequency	No. of users
Experts writers	5.4	16/63
Non-native writers	0	
Native writers	9.1	6/26

1-"You can see from the tables that only native students use *following section / previous section more* than the expert computer scientists writing journal articles. Could you comment on that?"

2- To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?

3. Can you think of any other reasons which may explain why the native and non-native students, and also the experts use this phrase more or less frequently?

4. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

5-Are there any other comments you would like to add?

D- more questions about single occurrence patterns:

prompt card (8)

Corpus	Data access		Traffic N+prp network		System resources	
	Normalised frequency	No. of users	Normalised frequency	No. of users	Normalised Frequency	No.of users
Expert writers						
Non-native writers	5.9	1/29	0.99	1/29	1.3	3/29
Native writers	2.7	3/26	0.68	1/26	2.7	3/26

1-why do you think that these three patterns were occur only in students' corpora but not in the reference corpus?

2- To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts?

3-To what extent do you think that the dissertation or journal article **topic** might affect either students' or experts' use of this academic phrase? I'd like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think:

Topics for traffic n+prp network:

NNS1: Computer Security

NS 20 :Dimensioning the Mobile Backhaul

Topics for system resources:

NNS 23 : Intelligent Web Search using Named Entity Recognition

NNS 17: Intelligent control of an unmanned aerial vehicle

NS 3: The development of a negotiation system using software agents to attempt to resolve the irregularities associated with the transfer of Professional Football Players.

NS 19: Create a social networking website to rival Facebook.

4. Can you think of any other reasons which may explain why only the native and non-native students use this phrase?

5. Here are three factors that some people have said may explain the differences. What's your own view?

A. native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently

C. personal style: different writers write in different ways.

6-Are there any other comments you would like to add?

7- Are there any other comments you would like to add about nothing from the interview?

Thank you very much for your participation.

Appendix K: Transcription of a Computer Scientist's Interview

Speaker introducing the interview

S: As you mentioned in your e-mail, you are specialised in Bioelectronics; how is that related to CS?

P2: I actually worked with Bioelectronics for the last 20 years, but I also worked in other branches of CS. I taught Software Engineering in the US and Portugal and now I am teaching Artificial Intelligence. You can see that I wondered into the CS.

S: How long have you been teaching in CS?

P2: In this university 11 years, but I have worked more than 25 years.

A- <u>General Questions</u>

First, I have some questions about CS dissertations and the English language requirements of the department

S: I have checked the CS website and found that MSc dissertations should be between 50-60 pages. What is the minimum and maximum number of words for a dissertation in your department?

P2: Actually, we do not count them in number of words at all. Some MScs are more software-oriented and others are more research-oriented. But we count them in number of pages 60 to 80 pages, 1.5 spacing. We do recommend students to write about 10,000 to 18,000 words, but we do not check that; what is important is the technical content.

S: From your experience, to what extent do you think that non-natives' level of language proficiency is near or similar to the native speaker students?

P2: Aah, I think there are two answers. We need to differ between foreign students and UK students. Foreign students are getting better. As I can see from their writing, it is clearer than before, maybe the university has risen the standard of the training courses for them. While the British students seem to go in the other direction. They seem to be unconcerned, getting sloppy, less careful, less clear, maybe because of all their texting and talking in computer. But in my opinion, there is a clear drop of the quality of the UK students' writing. This is in general.

S: How can you describe the writing of Arabic students specifically?

P2: I think there are two main kinds: ones that do not have any problems and you cannot notice they are foreigners, but I have had some Arabic students who have difficulty in writing good English. But it depends on their original countries. For example, Kuwait and Saudi Arabia students have a good level of English proficiency. They tend to be very good in English

S: Can they be described as native-like in their proficiency?

P2: Yes, they do, normally. Arabic students are good in their English compared to students from other parts of the world. Chinese students are really weak.

B- Collocation Use (Detailed Questions about the Judgement Task)

Now, I would like to ask you some questions about the judgement task.

Having looked at your comments on the judgement task, you identified some of the phrases as general/specific CS academic phrases and others as general academic phrases.

1- If we just start by looking at the second one (data input), you marked it as GAP; do you not think it is more specific to CS than other disciplines?

P2: Not really. From my experience of 25 years, I worked with people from Sociology and Medicine. They use this phrase regularly, either as data input or input data. It is mostly used in statics. Having worked with people in Medical Science, they heavily use data input in their statics. But most people in CS might not recognise that.

2- What about data access?

P2: It is still used in Medical areas. It has a slightly different meaning, but it still widely used. It can be related to accessing data of patients' records, but in CS it is related to different kinds of information, to numbers and other CS-oriented information. The meaning is different but the term is quite common.

3- What about data type? [showing definition]

P2: It is the same thing, all of these come from the fact that I had worked with people in Medical Sciences including hospitals. All of these terms are common. Actually one area of collaboration between Medical Science is with data type, data mining, and extra systems for patient support.

4- What about design system? You marked it as a GCS phrase. Do you think it might occur in different disciplines?

P2: Not really. The two words come together only in CS. I know it might be used in other disciplines but it is very colloquial. But in CS, it is a very specific term. System design or design system is a topic we teach our students about. I actually have a book here…in Software Engineering. There is a lot about design system; by system, we mean information system. It is a big area in CS and everybody who works in CS should learn it. It has a very specific meaning

because of the word ‘design’. Design here means how the parts are going together. And system related to the whole picture of the design program.

5- What about network traffic? It seems very specific to CS rather than in other disciplines.

P2: This term used to be very specific to the IS, but nowadays it can be used in any discipline of CS. It is more commonly used than before. It was used in network specialisation, but since almost everything we do is related to network traffic it has become a common term.

S: So it can be used in any discipline?

P2: Exactly.

6- What about process development? You marked it as specific to SE; why is that?

P2: It is a technique students use when they learn to develop a large software. Having said that, in Chemical Engineering, Process Development … they can use that term as well … but with a different meaning. But in SE it has a very specific meaning.

1- **What about design implementation? I can see you marked this one as specific phrase to SE. Why is that?**

P2: Yes, the standard in SE method have four steps: requirements, design, design implementation, and testing. These two words actually always appear in this way: design implementation. Most students will have a chapter on this in their dissertations.

2- **What about vulnerable files? Why did you write the comment you are “not sure”?**

P2: I am not sure whether this phrase can be used as a general academic phrase or general Computer Science phrase. In Computer Science, it is mostly used in security, which is a very specific area.

3- **Did you find difficulty in categorising these phrases?**

P2: Actually, yes. There are few phrases that were confusing. The ones that are separated. The words around the phrases can affect the meaning.

10- OK, let us look at one of them, method…class. You marked it as GA phrase and then you explained its use in different discipline. Can you comment on that?

P2: There are two different meanings, we can clearly see that. In AI when we used method…class, it is closer to the colloquial meaning, which is category. But in software development it is a small part of a program. It is a task. There are very specific meanings according to the areas in which they are used. We understand what it means but for an outsider like you it will be not clear.

S: Yeah, this is the problem I faced and thus I sought your help. Let’s move on to the third part.

Explanation about what this part includes [Show him the first prompt card and explain what is in the table].

C- Detailed Questions about some of the 24 Shared Phrases (Use and Patterns)

My results reveal that some of these collocations were used differently by Non-Native AND Native students compared to expert writers – by experts, I mean Computer Scientists writing journal articles. I would like to investigate this point in more detail.

Questions will be asked on prompt cards.

Prompt card (1):Environment development/development environment

1- You can see from the table that both native and non-native students use *environment development* more than the expert Computer Scientists writing journal articles. Could you comment on that?

P2: Yes, I think this is very interesting. The term 'development environment' refers to a program that allows working on another program. I would assume that people who write in journals will not talk about software used in developing another program. They just mention that. But students in all levels, when they write, have to say which software they used. So this would be related to the writing required in the journal. Development environment will not be a bit of important information; people are not interested in which program you used to develop the software. Whereas for students it would. However, this explanation cannot explain the difference between NS and NNS writers. I would assume that NNS writers would have a tendency to try to think about dissertations to mention the name of IDE. For example, eclipse instead of mentioning development environment. This is really a guess. Between these two, I am not entirely sure.

S: To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?

P2: Yes, definitely. It is a kind of information students need to report in their writing of dissertations but in journals it is not that important.

3-To what extent do you think that the dissertation or journal article topic might affect writers' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NS1: Implementation of Game Agents in Unreal Tournaments

NS13: Mobile Phone Training for Elderly People

NNS14: Advanced Web Application Programming

NNS23: Intelligent Web Search Using Named Entity Recognition

RC16SE: A Logical Verification Methodology for Service-oriented Computing

P2: Yes, the first and third ones will definitely use the term. It involves programming the agents so the students will surely mention which program they used. But for the second and fourth, I do not think development environment will make a difference. By the way, we do not say development environment, we said IDE. It means integrated development environment.

4-Are there any other comments you would like to add about this phrase?
P2: No more.

Let's move on to the next card.

Prompt card (2): Following code

1-You can see from the table that both native and non-native students use *the following code* more than the expert Computer Scientists writing journal articles. Could you comment on that?

P2: This one is similar and more effective than the previous one. The 'following code' means that students are going to show real lines of programming. Lines of code … or lines of programming. This phrase is really used when students write about programming … but in journal articles, they will not list codes. Most journals will say explicitly "do not put these codes"; they do not want to see the codes. If you want to see the codes, you can make it available online so people can download it. In fact, in both projects nowadays, since there are a lot of codes used… for example, a project of three or four months can easily have hundreds of pages of codes. Obviously, looking at all these codes will be difficult; even now, we ask our students not to add them into the project: they are better keeping them on CDs. If we need them, we will look at them. So we expect students to use less and less codes in their writing. I am not surprised if we look at the numbers of this one compared to the previous one (development environment), but looking at the number of users I think it is not statistically significant.

Aah, from my reading of thousands and thousands of journal paper I have read since the 80's, I remember that I read only one article that used the codes in it. It was about a new program and thus code would be mentioned. But other journals use pseudo codes: we just use it in sentences to explain how and why we use it in programming. It is like a recipe.

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?

P2: I think I already explained that.

3-To what extent do you think that the dissertation or journal article topic might affect writers' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NS27: Minimum Spanning Tree with Uncertainty

NS28: Optimising for High-Performance Cache Utilisation

NNS11: Optical Information System

NNS14: Advanced Web Application Programming

RC19SE: DARWIN: An Approach to Debugging Evolving Programs

P2: It will definitely make a difference but I am surprised how codes will be used in these specific topics. For the second one, it will definitely occur because of the level of programming used there. It is mostly used in hardware. We can expect some codes … about Cache Utilisation ... but for number three I think no codes will be used. Number four? Possibly. I think it might be there.

4. Can you think of any other reasons that may explain why the native and non-native students and the experts use this phrase more or less frequently?

P2: Aah...I think we can speculate that NS and NNS students come from different cultures, so it is not only the language. I think that NNS students come from cultures that are less computer-oriented than here. As you can see, students from their teenage are computer-oriented. It makes part of their lives. So it is possible that they are used to technical and computer terms from a younger age. So when they get here (university) they tend to use these terms more. I think this is only a possible explanation. It is only speculation.

S: Do you think that applies to Arabic culture?

P2: Yes...I think Arabic students are less computer-oriented compared to other Asian students. In fact, I realise they like to spend their times in social gathering. They invited me to their gatherings when they celebrate; they like their social life. But European students, they tend to spend more hours working in their labs. It might be an exaggeration to some extent but I think there are some cultural factors.

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently

C. Personal style: different writers write in different ways.

P2: I think all these factors are valid, but it will not explain how these different terms are used in general. All of these are pervasive.

P2: Obviously, I realised that NNS students tend to use long sentences with many chopped parts. Whereas experienced NS use short sentences. This is just general comment, not applicable only to this term.

6- Are there any other comments you would like to add?

P2: No.

Prompt card (3): Method class

1-**You can see from the table that native students use *class method* more than the expert Computer Scientists writing journal articles. Could you comment on that? While non-native students use method prp class less than the expert Computer Scientists writing journal articles. Could you comment on that?**

P2: I think that 'method prp class' indicates that the class they are talking about is related to AI whereas I mentioned before it is used for category. Whereas the 'METHOD CLASS' will be used in Java, in a very specific program. In other Computer Science it is called function but in Java specifically it is called method class.

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?

P2: Yes, I think that would be the main reason. The level of information needed is different. In journals you will not find this term often, like in dissertations.

3-To what extent do you think that the dissertation or journal article <u>topic</u> might affect writers' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NNS18: Web Summarization Searches

NS7: Mobile Development

RC15, 16 SE (2nd P): A Framework for the Checking and Refactoring of Crosscutting Concepts

16SE: A Logical Verification Methodology for Service-oriented Computing

19SE(1) DARWIN: An Approach to Debugging Evolving Programs

18IS(2 and 3): Information Technology Implementers' responses to User Resistance: Nature and Effects

P2: There is one possibility. Looking at these, some of these topics will use Java, thus the term will occur.

4- Can you think of any other reasons that may explain why the native and non-native studentsand the experts use this phrase more or less frequently?

P2: No more reasons.

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: I think A is the most likely explanation. Foreigner students like to use prepositions more than NS. They like to use long phrases in their writing. You might be aware of that as you are from linguistics.

S: What makes you say that?

P2: From my reading of hundreds and hundreds of dissertations, I can easily see that.

S: What does that mean?

P2: They tend to use a lot of prepositions. Sometimes they are not aware that the whole thing can be put in a natural way. For example, my friend's car tyres. NNS writers may say the tyres in my friends' car or the tyres in the car of my friend. The construction is very different.

S: Does that affect the style?

P2: Yes, of course. Arabic students tend to write a hundred pages of detailed information in the same, unlike Chinese students who use very concise information. It might be that they use the same style of their first language.

Prompt card (4): Source code

1- You can see from the table that native and non-native students use *source code* different from the expert Computer Scientists writing journal articles. Could you comment on that?

P2: Something you may not be aware of is that this term is more British oriented. In the US, they tend to use a different term; they say this is the program, lines of program, but in the UK, we use source code and lines of code instead. So there is a difference between the countries as well. When I came to this country, I realised the difference of how many times they use source code and lines of coding compared to the US.

I am not surprised to see the big numbers. It is a British term. It is interesting that the expert writers use it more than the NNS. I think the same explanation can be implied here.

2- To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?

P2: I think it would and especially in the case of NNS writers as they use it less than the other two groups. It is more British-oriented. But you can see in journal articles they will say that they will not mention the codes: they only mention that they use source codes for their programming.

3- To what extent do you think that the dissertation or journal article <u>topic</u> might affect writers' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NS2: Intelligent System and Robotics

NS5: Web Application Programming

NNS14: Advanced Web Application Programming

NNS17: Intelligent Control of an Unmanned Aerial Vehicle

RC(15SE): A Framework for the Checking and Refactoring of Crosscutting Concepts

16SE: A Logical Verification Methodology for Service-oriented Computing

P2: For source code, I would not think it is a term used in any area of CS. Actually, in all sub-areas of CS there is programming somewhere, so source code will be there. I would not think it is related to specific topics.

4- Can you think of any other reasons that may explain why the native and non-native studentsand the experts use this phrase more or less frequently?

P2: None.

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: All these can be applied. But I think the reason I mentioned about the term; that is, it is more British-oriented. So it can add to the division between NS and NNS use of language. For B I do not think so. I think the way you picked up the term is what makes the difference.

For C, somewhat. If you are more UK-oriented than US-oriented it will probably affect the use of this term.

6- Are there any other comments you would like to add?

P2: No more.

Prompt Card (4): Web site

S: It only occurs in NNS.

P2: Web site is a really general term. My guess will be the choice of topics. It might worth looking at the curriculum we have ... up to the last five years, we have massive projects based on websites. But the level of skills required was not high enough for the degree of the MSc. Most of the NS students from the UK had already done that in their undergraduate level or before they come to the university. We discourage NNS students to do their projects on web site. We need to use web site as a means not as an end. For example, in some webbing of sites web site will be used, as they tend to use web site earlier.

I would think the NNS might like to choose projects related to web site; it might be more useful to where they come from, while here they assumed they already have it.

When we come to experts to NS users, I am actually surprised to see a little difference between these two: I would expect the difference to be bigger. In journals, in most areas, it is likely to be used. It is a general term.

S: What about the topics?

P2: The first two are dealing with web site, but for others I am surprised to find this term.

S: What about the three factors suggested by others?

P2: For the first one, yes ... the background of NNS students affects their choice of topics. B, I don’t think so. Without looking at numbers, this is not a factor.

In personal style, I think if we exclude the choice of projects, so let's leave that aside, then I could think that personal style can play a secondary role in that people use web site in more generic projects when they are trying to think of general terms.

Prompt card (5): Data type

1-You can see from the table that only non-native students use *data type* more than the expert Computer Scientists writing journal articles. Could you comment on that?

P2: I think that goes to what I was guessing a few minutes ago that NNS students tend to use more prepositions to make longer phrases rather than using the short ones. You can notice this forms their writing and you can spot that. This is not surprising, to put it in this way. NS use more data type instead of type of data. So, this confirms what I said.

From my experience, even without looking at names of students, you can notice whether he is NS or not English. NNS tend to use a lot of prepositions. Actually, we can notice two important things from their writing. First, triple use of preposition when we do not need them. Second, the use of articles; they put the definite article instead of the indefinite one.

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?
P2: I do not think that may explain it. Data type is a very general term in CS.

3-To what extent do you think that the dissertation or journal article topic might affect writers' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NS8: Intelligent System and Robotics

NS1: Implementation of Game Agents in Unreal Tournament

P2: No, it is not related to the topics.

4- Can you think of any other reasons that may explain why the native and non-native studentsand the experts use this phrase more or less frequently?

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: It is mainly A.

6-Are there any other comments you would like to add?

P2: No more.

Prompt card (6): Data input

1-You can see from the table that only native students use *input data* more than the expert Computer Scientists writing journal articles. Could you comment on that?

P2: I wonder why NNS won't use input data. If they are doing topics in AI data input will come up all the time.

So choice of projects might have an effect. Data input is actually very common in AI projects but in other areas of CS it might be used less. I think the biggest factor will be the choice of projects. Another explanation could be that they use other phrases to express the same meaning of data input, like information input, I guess.

Aah…thinking of Arabic students, they might have equivalent terms in their language so they may use it instead of data input.

2- To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?

P2: No, I do not think so.

3- To what extent do you think that the dissertation or journal article topic might affect either students' or experts' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.

Topics

NS1: Implementation of Game Agents in Unreal Tournament

NS3: The Development of a Negotiation System using Software Agents to attempt to Resolve the Irregularities associated with the Transfer of Professional Football Players (E-commerce Technology)

RC(15AI): Similarity Measure for Anomaly Detection and Comparing Human Behaviours

17AI: Text Summarization Contribution to Semantic Question Answering: New Approaches for Finding Answers on the Web

P2: In the first two, it is likely, but in the last two, definitely. As you can see, these articles are in AI.

4- Can you think of any other reasons that may explain why the native and non-native studentsand the experts use this phrase more or less frequently?

P2: No.

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: From these factors, I think that NNS Arabic students may use other equivalent terms instead of data input.

6- Are there any other comments you would like to add?

P2: No.

Prompt card (7): Section following/section previous

1-You can see from the tables that only native students use *following section/previous section* more than the expert Computer Scientists writing journal articles. Could you comment on that?
P2: Yes. This is specific to Arabic students; for example, Portuguese, Danish where I had taught for many years. These phrases are used everywhere. So, again, I am surprised to see this. Having said that, they might be related to the style of the writing. Most people like to give signs to the readers but for foreigner students they may not use these signs in their writing.

NS could possibly be used to the style of academic writing.

2-To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?
P2: Yes, this could be an explanation. People writing in articles are more likely to mention the name of the chapter or section instead of using the term. They do not have room for that. I agree with you in that eliminations of number of chapters and sections will be in journals but in dissertations students write more sections and chapters; it is more preferable.

3-Can you think of any other reasons that may explain why the native and non-native studentsand the experts use this phrase more or less frequently?
P2: No.

4-Here are three factors that some people have said may explain the differences. What is your own view?
A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: No, none of these explains.

5-Are there any other comments you would like to add?
P2: No.

Prompt card (8)

1- **Why do you think that these three patterns were occur only in students' corpora but not in the reference corpus?**

P2: I think that data access and systems resources are general CS phrases. So I am surprised why it does not show up in the expert writers.

S: Actually, expert writers use this phrase, but in a different way. For example, access data.

P2: Aah…so it is not a complete absence of the term. Variation could occur in using this term. It depends on the style of writer, I guess.

S: What about traffic (n+prp) network?

P2: Looking at number of users, I don't see any significance of this phrase use. Even though it is a specific CS phrase, it is more commonly used as network traffic.

2- **To what extent does the fact the students are writing dissertations rather than journal articles explain why the students use this phrase more often than the experts do?**

P2: It can't be explained by this.

3- **To what extent do you think that the dissertation or journal article topic might affect either students' or experts' use of this academic phrase? I would like to show you some of the students' dissertation topics and the expert writers' journal article topics and ask what you think.**

Topics for Traffic n+prp Network

NNS1: Computer Security

NS20: Dimensioning the Mobile Backhaul

Topics for System Resources

NNS23: Intelligent Web Search using Named Entity Recognition

NNS17: Intelligent Control of an Unmanned Aerial Vehicle

NS3: The Development of a Negotiation System using Software Agents to attempt to Resolve the Irregularities associated with the Transfer of Professional Football Players

NS19: Create a Social Networking Website to rival Facebook

P2: It is not related to these topics.

4- Can you think of any other reasons that may explain why only the native and non-native students use this phrase?

P2: No.

5- Here are three factors that some people have said may explain the differences. What is your own view?

A. Native and non-native writers use language differently;

B. Experienced and inexperienced writers use language differently;

C. Personal style: different writers write in different ways.

P2: I think experience might be an explanation. Aah…I don't think other factors apply. I guess (laughs), a lot of guesses…

6- Are there any other comments you would like to add?

P2: No.

7- Are there any other comments you would like to add about nothing from the interview?

P2: I think that you might need to focus on NS and find the US- and UK-oriented phrases.

Thank you very much for your participation.

Appendix L: List of Codes for CS Experts' Interviews

a- General Questions about Writing Requirement in CS

Codes	Definition
No. of words/No. of pages	This code refers to the number of pages or number of words required in writing MSC dissertations
NS and NNS (students) academic writing style	This code is used whenever a variation between NNS and NS students in their academic writing are expressed by CS scientists
Variation between NNS/NS students' English language proficiency	This codes refers to the variation between NNS and NS language proficiency

b- Collocations from the Categorisation Judgement Task

Codes	Definition
1-DIFF UNCER Difficulty and uncertainty	This code was used when the CS scientist expresses the difficulty of categorising a collocation and uncertainty about the collocation use and meaning
1a- DIFF UNCER MEANING Difficult and uncertain - similar meaning	This code refers when the CS scientist expresses the difficulty of categorising a collocation because of the meaning
1b- UNCER GCS OR SCS No clear cut between GCS and SCS collocations	This code refers when the CS scientist mentions the difficulty of categorising a collocation as GCS or SCS
1c: DIFF UNCER NON ADJA Difficult and uncertain – non-adjacent collocations	This code refers to when the CS scientist mentions the difficulty and uncertainty of categorising non-adjacent collocations
2- CERTAIN	This code was used when the CS scientist was certain

Certainty of categorising the use of the collocations	about his categorisation of a collocation use (GAP, GCSP, SCSP)
3-ADDIT COMM Additional comments	This code refers to when the CS scientist mentions other comments about his categorisation

c- Collocations on Prompt Cards; Collocation Use and Patterns

Codes	**Definitions**
1- COLLOC USE DIF Collocation use difficulty	This code refers to when the CS scientist expresses the difficulty of explaining the collocations use among corpora
2- GEN EFFECT Genre effect	This code refers to when the CS scientist agrees that genre affects the use of collocation among corpora for one of the sub-coded (2a, 2b, 2c) reasons
2a-W DISS vs. W RA Writing in dissertations vs. writing in articles	This code refers to when the CS scientist mentions the various demands of writing in dissertations and in research articles
2b- DISS DEM Dissertation's writing demands	This code refers to when the CS scientist mentions the demands of writing in dissertations only
2c-RA DEM Research articles' writing demands	This code refers to when the CS scientist mentions the demands of writing in articles
2d-GEN N EFFECT Genre - no effect	This code is used when the CS scientist thinks that genre does not affect the use of collocation
3- TOP SPEC COLLO Topic-specific collocations	This code refers to when CS scientist mentions the effect of topic on the use of the collocation
a- AGRE-Agreement	This code is used when the CS scientist agrees that the collocation is topic-specific
b- DISAGRE/UNCERT	This code is used when the CS scientist disagrees and is uncertain about the collocation specificity

Disagreement/Uncertainty	
c- GCOLLOC General collocations - not topic-specific	This code is used when the CS scientist expresses that general collocations are not topic-specific
d- SCOLLOC Specific collocations - topic-specific	This code is used when the CS scientist expresses that specific collocations are topic-specific
4- Other factors given (a, b, c)	
a- NNS vs. NS UL NNS vs. NS in their use of language	This code refers to when the CS scientist mentions any different use of any language aspect between NNS and NS
b- EXPER vs. NOV Experts vs. novice writers	This code refers to when the CS scientist mentions various writing style between expert and novice writers
c- PER ST Personal style	This code refers to when the CS scientist thinks that personal style of the writer may affect the collocation use or patterns
d- ALL N EFFEC ALL (a, b, c) factors - no effect	This code refers to when the CS scientist thinks none of the previous factors (a, b, c) affect the use of the collocation
5-Interviewees' additional reasons	This code refers to when the CS scientist mentions other factors that might affect the use of the collocations
a- Cultural factor	This code refers to when the CS scientist mentions the effect of culture in the use of the collocation
b- Subject-related collocations	This code is used when the CS scientist thinks that the collocations is used in a specific subject
c- Use of equivalent L1 terms	This code refers to when the CS scientist thinks that the use of collocations is affected by the use of equivalent terms from students' L1

d- Codes for Evaluation of Instruments

Codes	Definitions
DIF CJT Difficulty withCJT	This code is used when the CS scientist mentions any comments about the difficulty of the CJT
RES REACT CJT Respondents' reaction	This code refers to when the CS scientist expresses his reaction to the CJT
ADDITON COM CJT Additional comments	This code refers to when the CS scientist mentions any other comments about the CJT as an instrument

Appendix M: Categorisation Judgment Task Results

Phrases with adjacent words	General academic phrases (GAC)	General CS academic phrases (GCS)	Specific CS academic phrases				
			AI	SE	IS	Dictionaries checked	Comments
1-code following/ following code		P1 P2				NF	
2-data input/ input data	P2	P1 P3				GCS	
3-data access	P2 P3	P1				GCS	
4-data user	P1, p2 P3					Gcs	
5-data information/ information data	P1,p2	P3				Gcs	
6-data type	P2	P1 P3				GCS	
7-design system/	P1 P4	P1 P3				GCs	
8-environment development/ development	P3	P1 ,p2 P4				GCS	
9-layer application/ application		P4		P1, p2	P1, p2	GCS	
10-network traffic/ traffic network		P2 P3		P1	P1 P4	GCS	
11-resources available/ available	P1,p2 P3					GAC	
12-resources system/ system resources	P4	P1,p2, P3				GCS	

13-code source / source code		P1,p2 P3				GCS	
14-data layer				P2	P1, p2, P3	GCS	
15-data available/ available data	P1,p2 P3					NF	
16-previous section	P1,p2 P3,					NF	
17-following section	P1,p2, P3					NF	
18-Web site/Site Web	P1,p2, P3					GAC	
19-Open source		P1,p2 P3				GCS	
20-components different/ different	P1, p2	P2				NF	
21-simulation results	P1, p2	P3				GCS	
22-data structure/s	P2	P1 P4		P3		GCS	
23-error rate	P1 P4	P2			P3	GAC	
24-extraction information	P1	P1,p2 P4			P3	GCS	
25-allocation dynamic/ dynamic		P1,p2 P4		P3		GCS	
26-data training		P2	P1 P3			G CS	
27-data test	P1 P4	P1,p2			P3	G AC	
28-computer vision		P2 P3	P1 P4			GCS	

29-process development	P1 P3	P1		P1	P2	GCS	
30-data database	P1 P3	P1,p2				GCS	
31-data raw/ raw data	P1,p2 P3	P1				GAC	
32-design architectural	P1, P4	P2, P3				GCS	
33-design implementation	P1 P3	P1		P1	P2	NF	
34-files vulnerable /		P1,p2 P3			P1	NF	P2 specific to data mining in
35-function ranking	P4		P1, p2	P3		NF	P2
36-document ranking	P1	P3	P1, p2			NF	P2
37-document scope	P1 P3		P2			NF	P2
38-files neutral/ neutral files			P2	P3		NF	P2
39-document cohesion	P1 P3		P1, p2			NF	P2
40- data …time	P1, p2	P3 P4				GAC	
41- number…code	P2	P1, p3	P2			GAC	
42- amount …data	P1, p3	P4				NF	P2
43- other…features	P1, P4			P3		NF	P2
44-other…data	P1 P4	P1		P3		NF	P2

45-data..web	P1	P1 P4			P3	NF	P2
46- method…calss		P2 P4		P1	P3	Nf	P2
47- model…section	P1 P4	P1		P3		NF	P2
48- data …different	P1 P4				P3	NF	P2
49- query …docum ent	P1		P1 P3		P1	NF	P2

Appendix N: Acadmeic Collocations Awareness –raising Activities

1. Source Code

Activity One: Noticing Collocation

The following exercise will help you notice the kinds of words and phrases that are often found with 'code' in Computer Science writing either in the left or right context. Spend some time analysing the concordances of this word and answer the following questions:

A. The word 'code' is used as noun. Look at the words to the right of 'code'. Which words are more frequently used? Can you identify the part of speech of these words?

B. Which words and phrases go to the left and right of 'code'?

4- …knowledge of the source ***code*** or better user…

5- …two million lines of source ***code***, and evaluated the…

6- …We have provided C++ source ***code***, but it is straightforward…

7- …we provide the source ***code*** for Computing the proposed…

8- …and have available source ***code*** and fault repositories…

9- …with source lines of ***code***, alert density from a…

<Extracted from the reference corpus>

Activity Two: Noticing Patterns

1- How many patterns did you find for 'source code'?

2- In the following table, write down the patterns and how many times you found each pattern.

Patterns for 'source code'	Number of occurrences (frequency)

Activity Three: Comparing and Contrasting Patterns between the NNS Students' Corpus and the Reference Corpus

Look at the concordance lines of set (A), which are taken from NNS students' corpus. Spend some time analysing the words and phrases that go together with 'source code'. Then answer the questions below.

Set (A)

1 …compatibility of the source ***code*** written in net framework…

2 …that file from the source ***code***. For example…

3 …simulation using source ***code*** written specially to simulate…

4 …a copy of the source ***code*** is included in appendixes…

5 …represented in the C source ***code*** by the line below…

6 …by running a source ***code*** written in Matlab…

7 …the corresponding source ***code*** was developed to oblige the…

<Extracted from NNS corpus>

A. How many patterns are used by NNS students for 'source code'?

B. Do the NNS students use any of the same words and phrases you found earlier, when you looked at 'source code' in the previous activity?

C. In the following table, write down the patterns and how many times you found each pattern.

Patterns for 'source code'	Number of occurrences (frequency)
Pattern 1	
Pattern 2	
Pattern 3	

D. Compare between the patterns you identified for NNS with those found in the reference corpus in the previous activity.

2. Data User

Activity One: Noticing Collocation

The following exercise will help you notice the kinds of words and phrases that are often found with 'data' in Computer Science writing either in the left or right context. Spend some time analysing the concordances of this word and answer the following questions:

A. The word 'data' is used as noun. Look at the words to the right of 'data'. Which words are more frequently used in this body of data? Can you identify the part of speech of these words?

B. Which words and phrases go to the left and right of 'data'?

1- …Before using the pooled ***data*** of 355 user responses…

2- …of huge amounts of user ***data***; however, in the case…

3- …music is to analyse user ***data***, such as which music…

4- …The user clickthrough ***data*** are collected based on this…

<Extracted from the reference corpus>

Activity Two: Noticing Patterns

1- How many patterns did you find for 'data user'?

2- In the following table, write down the patterns and how many times you found each pattern.

Patterns for 'data user'	Number of occurrences (frequency)

Activity Three: Comparing and Contrasting Patterns Between the NNS Students' Corpus and the Reference Corpus

Look at the concordance lines of set (A), which are taken from NNS students' corpus. Spend some time analysing the words and phrases that go together with 'data user'. Then answer the questions below.

Set (A)

1 …radio channel to a mobile ***data*** user, works by dedicating…

2 …stored data and compare the ***data*** to the user query to…

3 …TextBox controls for accepting ***data*** from the user…

4 …The first stage is the ***data*** collection of user information…

5 …fragmentation of user ***data*** for fitting the physical…

6 …the address of the user ***data***, which is copied into…

7 …If the user ***data*** are matched, then the…

8 …as the amount of user ***data*** carried by the network…

9 …represents the user profile ***data***, and the second one…

10 …to store the user profile ***data*** in a relational database…

11 …the user with clear ***data***. Though…

<Extracted from NNS corpus>

A. How many patterns are used by NNS students for 'data user'?

B. Do the NNS students use any of the same words and phrases you found earlier, when you looked at 'data user' in the previous activity?

C. In the following table, write down the patterns and how many times you found each pattern.

Patterns for 'data user'	Number of occurrences (frequency)
Pattern 1	
Pattern 2	
Pattern 3	

D. Compare between the patterns you identified for NNS with those found in the reference corpus in the previous activity.

3-Data Access

Activity One: Noticing Collocation

The following exercise will help you notice the kinds of words and phrases that are often found with 'data' in Computer Science writing either in the left or right context. Spend some time analysing the concordances of this word and answer the following questions:

1-The words ‘data’ are used as nouns. Look at the words to the right of ‘data’. Which words are more frequently used in this body of data? Can you identify the part of speech of these words?

2-Which words and phrases go to the left and right of ‘data’?

1 …worth of application server access log ***data*** to simulate user…

2 …and track a write access to the protected ***data*** using…

3 …in terms of ***data*** object access. We note here…

<Extracted from the reference corpus>

Activity Two: Noticing Patterns

1- How many patterns did you find for ‘data access’?

2- In the following table, write down the patterns and how many times you found each pattern.

Patterns for ‘data access’	Number of occurrences (frequency)

Activity Three: Comparing and Contrasting Patterns between the NNS Students’ Corpus and the Reference corpus

Look at the concordance lines of set (A), which are taken from NNS students’ corpus. Spend some time analysing the words and phrases that go together with ‘data access’. Then answer the questions below.

Set (A)

1 …being developed, the ***Data*** Access layer should contain the…

2 …Logic Layer, and ***Data*** Access Layer. Any changes…

3 …Business Logic_Layer and ***Data*** Access Layer. The User Interface…

4 …Interface layer and ***Data*** Access layer. The Data…

5 …The ***Data*** Access layer should contain all the…

6…Business Logic layer or ***Data*** Access layer. The website…

<Extracted from NNS corpus>

A. How many patterns are used by NNS students for ‘data access’?

B. Do the NNS students use any of the same words and phrases you found earlier, when you looked at ‘data access’ in the previous activity?

C. In the following table, write down the patterns and how many times you found each pattern.

Patterns for ‘data access’	Number of occurrences (frequency)
Pattern 1	
Pattern 2	
Pattern 3	

D. Compare between the patterns you identified for NNS with those found in the reference corpus in the previous activity.