THE UNIVERSITY
OF BIRMINGHAM

CORPUS USE BY STUDENT WRITERS: ERROR CORRECTION BY THAI LEARNERS OF ENGLISH

by

PATSON JAIHOW

A thesis submitted to
The University of Birmingham
for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language and Applied Linguistics
School of English, Drama and American & Canadian Studies
The University of Birmingham
September 2016

Abstract

Researchers in corpus linguistics and applied linguistics have recommended the use of corpus data by language learners to promote independent learning (Bernardini, 2004; Yoon & Hirvela, 2004; O'Keeffe et al, 2007). However, it is not clear to what extent learners are able to use corpus resources independently, and how they can be trained to use a corpus more effectively. This thesis reports a study of learners using a corpus for error correction. The learners recorded their processes using a think-aloud protocol. The thesis records three main findings. Firstly, the learners found it easiest to spot and correct errors of clause structure, noun class, adjective pattern, and collocation; they found verb pattern the most difficult errors to correct. Secondly, the learners most frequently searched for information about colligation, collocation, acceptability/occurrence of strings in a corpus, and determiner-noun agreement; they searched for information about lexical pattern relatively infrequently. Finally, the learners worked most effectively with the corpus when they entered single words as the search terms and scrutinized the concordance lines for collocates and patterns; they worked least effectively with the corpus when they entered whole strings of words. The thesis also makes recommendations for facilitating corpus use in classrooms and specifies the training that learners need to use corpora effectively.

To my brother

Acknowledgement

The completion of this PhD is possible due to several groups of people. The first is Prof. Susan Hunston, my supervisor who has taught me to be aware and careful of several issues in doing and writing up research. Her critical and constructive feedback was valuable to my intellectual development. She enabled me to undergo through many challenges (and difficulties) I had during my PhD life. I am so grateful to her effort and devotion.

I would also like to express my gratitude to Asst. Prof. Umpairat Sudhinont and Asst. Prof. Sutaree Prasertsan who allowed me to intervene their classes for data collection, to the Faculty of Liberal Arts, Prince of Songkla University (PSU), for facilities for data collection, and to the participants (3rd year English-major students at PSU). I am particularly grateful to Kwang, Pee Kae, Pee Pin, Ying, Trong, and Jade for their help with the pilot study of the error correction test. I also thank Ake, Champ and Jaran for their technical help during the data collection period.

I am thankful for my family and friends in Thailand for their morale support, and friends and housemates in the UK for making my stay here lively.

Lastly, but perhaps most importantly, I would like to thank Prince of Songkla University and the Commission on Higher Education of Thailand for granting me the financial support for my PhD studies.

Table of Contents

List of Tables

Chapter 4

Chapter 5

Chapter 6

List of Figures

Chapter 7

Chapter 1

Introduction

1.1 Data-driven-learning

The advent of computer technology and the development of English language corpora have influenced the ways in which language has been taught. Corpora in particular have led to the emergence of a computer-based language learning approach known as data-driven-learning (DDL), developed by Tim Johns (Johns, 1991). Data-driven-learning is a learning method that demands substantial concentration on the part of the learner; Johns (1991: 2) describes it as an approach where "the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data". By this definition, two things that make DDL a distinctive approach to language learning are the roles of both learners and teachers and the kind of linguistic data used. In the DDL context, the learner is viewed as a "language detective" (Johns, 1997: 101) and as "a research worker" (Johns, 1991: 2) whose job is to "discover" facts about the language being learned by themselves. The teacher, on the other hand, takes the role of a director and coordinator of learners' researching process so that learners can develop discovery strategies that enable them to "learn how to learn" (Johns, 1991: 1). To enable learners to do this, they need to be presented with a set of linguistic data that can give them information about language issues or questions raised either by the teacher, or by the learners themselves, such as "Is it better to say x or y?" or "What is the difference between saying x and saying y?" (Hunston, 2002: 170). To date, the kind of linguistic data that has been used in the DDL classroom is corpus data, usually in the KWIC (keyword in context) concordance lines.

DDL, in more recent work, is referred to as serendipitous learning (Bernardini, 2000; 2002) and discovery learning (Bernardini, 2004). In the serendipitous learning context, a learner is seen by Bernardini (2002: 131; 2004: 22) as a "traveller" as they "are guided to browse large and varied text collection in open-ended, exploratory ways" (Bernardini, 2004: 22), and a teacher's role is described as a facilitator as well as a learning expert who helps learners learn by themselves (Bernardini, 2004).

Based on their experience of implementing Data-driven Learning or Discovery Learning in the language classroom, Johns (1991) and Bernardini (2004) mention many advantages of this approach. For example, learners become more autonomous in their own learning. They are stimulated to inquire about language and develop skills in noticing language patterns and forming generalisations about language. A supportive learning atmosphere is created as both teachers and learners help to find out and share ideas about language. Hunston (2002) adds that this approach is supportive to learning because it motivates learners to remember what they have discovered. Thus, learners' motivation can be maximised by allowing them to find facts about language that serve their urgent linguistic needs. This type of learning, in the constructivist view, promotes learner autonomy and language acquisition in the long run and also fosters learners' development of identifying, hypothesising, and verifying skills (Boulton, 2009; Cobb, 1999; O'Sullivan, 2007; Bloch, 2007; Boulton, 2010; Keck, 2004, cited in Yoon, 2011: 131).

Despite these advantages, DDL has not yet become common practice in the language classroom (Römer, 2011). For some reason, teachers may be doubtful about the learning process and learning outcome. If learning is serendipitous, for example, what happens if the learner is not

curious or has no questions? How is this autonomy reconciled with exam needs? A review of previous work on DDL e.g. Thurnbull and Burston (1998); Kennedy and Miceli (2001); Watson Todd (2001); Gaskell and Cobb (2004); Chamber and O'Sullivan (2004); Lee and Swales (2006); Cresswell (2007); Sun (2007); Yoon (2008); Kennedy and Miceli (2010); Charles (2012) reveals that the number of studies of DDL effects in language classrooms has increased, but is still limited. Moreover, most of these DDL studies have focused on measuring the outcome or product of learning. Little is known about DDL as a learning process, which needs to be explored.

The present study aims to make an important contribution to this field of study by focusing on the learners' processes of investigating corpus data using think-aloud protocols. Insights into the ways learners conduct corpus searches and interpret corpus data can be used as a platform for training language teachers to make the best use of corpora in the classroom.

1.2 What is this study about?

The aim of the study was to find out, through think-aloud protocols, what exactly happens when a learner explores a corpus to find the answer to a question about language. To do this, learners were encouraged to use corpora to correct errors in their writing. It is necessary for them to be responsible for their errors because individual learners' errors can vary and it is impossible for teachers to deal with every single error. Therefore, students need to be trained to correct their own errors or mistakes. One of the most important things in learning a language I believe is to give learners the information they need to make their own language use better or to improve their own language use. In doing that, students have controlled information available to them to do this thing for themselves and this is a way of increasing their autonomy. One of the resources

that are increasingly available is corpora. As noted above, there is surprisingly little work on how learners use corpora and what actually happens when students encounter the corpus and try to use the information in the corpus to correct their own language. The purpose of the thesis is to discover what student corpus-users do and as a consequence to give advice to teachers on how to use a corpus for students.

This set of experimental studies was carried out in Thailand. I wanted to know how students use a corpus to check language rules, pattern and phraseology when they edit their own writing. To demonstrate this, I gave the students some training in how to use a corpus for discovering facts about language use, and in particular, for correcting errors. At the data collection stage, the students were asked to do an error correction test using a corpus, followed by editing tasks where they were asked to identify areas of difficulty in their writing and to use the BNCweb independently to check and improve accuracy. As they did so, they were asked to describe their thoughts and actions aloud and to video-record their working computer screens and their think aloud. These video-recorded think-aloud protocols and the record of the corpus searches they carried out provided the data for the study. These data were used to see how the students went about solving written errors using a corpus and what made them do it well or badly, as well as the problems they encountered.

1.3 A tradition of corpus linguistics

Corpus linguistics is not totally new, but it is not widely known outside the academy either, even among language teachers. A quick and easy way to understand what corpus linguistics is about is to know what a corpus is. According to Hunston (2002: 2), a "corpus" (plural: corpora) is the term used by linguists to refer to "a collection of naturally occurring examples of language,

consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study", and the word "corpus" has recently been used to refer to "collections of texts (or parts of text) that are stored and accessed electronically". This definition provides us with a basis for understanding two aspects of a corpus: its forms and its functions. Regarding its forms, a corpus is "compiled from writing and/or a transcription of recorded speech" (Krieger, 2003: 1). Therefore, the most important characteristic of a corpus is that it represents the form of naturally-occurring or authentic language.

Regarding the functions, corpora have been widely used in language research, as well as language teaching and learning. For language research, linguists can investigate different aspects of a particular language from a corpus. Biber et al. (1998), for example, argue that discovering the pattern of co-occurring features and dimensions of variation that characterise speech and writing would have been very difficult if there were no development of corpus-based methods.

Another important role of corpora in language education is to assist in the production of dictionaries and grammar books (Hunston, 2002). Lexicographers use corpus data for dictionary compilation by looking at how words have been used and write dictionary entries which reflect the central and typical uses of the language. As noted in McEnery and Xiao (2010), the learner dictionaries that claim to be corpus-based include the *Collins COBUILD English Language Dictionary* (Sinclair, 1987), the *Longman Dictionary of Contemporary English* (3rd edition) (Longman, 1995), the *Oxford Advanced Learner's Dictionary* (5th edition, Hornby & Crowther, 1999), and the *Cambridge International Dictionary of English* (Procter, 1995). Grammar books that are based on corpus data are the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad and Finegan, 1999) and *A Comprehensive Grammar of the English*

Language (Quirk, Greenbaum, Leech, and Svartvik (1985), cited in McEnery and Xiao (2010). Another corpus-based grammar book is the *Cambridge Grammar of English* (Carter and McCarthy, 2006). In language teaching and learning, teachers can use corpora as a resource to develop learning materials (Lee, 2005). For example, teachers planning to teach English for engineering students can select useful words or phrases from a corpus specific to the field of engineering to create their materials. A more modern use is to have engineering students collect and examine their own corpora. Teachers can also encourage autonomous learning by having students find a language pattern or rule from concordance materials.

To conclude, a corpus is a collection of spoken or written language stored electronically in computer files. With a corpus and corpus software, linguists can study language as it is used in the real world. This study of language through an analysis of corpus data is called corpus linguistics.

1.4 Revolution in language study

The development of corpora and the method of corpus-based language analysis have revolutionised language study and have partially revolutionised language teaching (Hunston, 2002; Flowerdew, 2009). In terms of language study, corpus linguistics has changed the ways linguists view language. Traditional views on language such as lexis and grammar as separate entities have been questioned and new concepts of language such as lexico-grammar, semantic prosody, and phraseology have been proposed. Studies of language have also shifted onto frequency, lexis and phraseology, and onto register and variation, instead of rules of grammar.

In terms of language teaching, it was found that traditional language descriptions in textbooks do not match the actual use of language found in large corpora (Carter, 1998; Burns, 2001; Burn, Joyce and Gollin, 2001; McCarthy and O'Keeffe, 2004; Thornbury and Slade, 2006, cited in O'Keeffe, McCarthy, and Carter, 2007: 21). This is the point where corpora have influenced language teaching. Insights and linguistic evidence gained from corpora have fed into the development of corpus-based teaching and learning materials such as the production of learners' dictionaries and course books. Moreover, corpora have had some impact on classroom practice.

As discussed in 1.1, the great influence corpora have had over language learning is the emergence of an approach to language teaching known as data-driven learning (DDL). In a DDL classroom, learners are encouraged to take charge of their own language learning by investigating concordances to make sense about language use (see 1.1).

1.5 Benefits of corpora for language teaching

To understand the benefits of corpora for language teaching, there is a need to understand two fundamental characteristics of a corpus, which are its authentic texts and its electronic form (Bowker and Pearson, 2002). Texts in a corpus are authentic because a corpus comprises texts used in the real world. Another characteristic is its electronic form, which can be processed by a computer, making it quick and convenient for corpus users to access the data which consist of millions of words efficiently.

These two characteristics of a corpus are beneficial for language learning. Authenticity allows learners to have access to empirical language data, and not rely on native speakers' intuitions about language, which can be wrong (Aston, 2001; Flowerdew, 1996, cited in Flowerdew, 2009).

Hunston (2002: 20) further explains that intuition is not always a good guide at least to "four aspects of language: collocation, frequency, prosody and phraseology". The electronic form allows learners to play an active role in identifying language patterns and discover facts about language which cannot be fully covered in other printed materials like dictionaries or grammar books (Flowerdew (2009). This point will be discussed again in chapter 2.

1.6 Use of corpora for language teaching

When learners access a corpus, basically they can make use of two outputs from a corpus: a concordance and a list of word frequencies. As Levy (1990: 178) puts it: "A concordance is a collection of all the occurrences of a word, each in its own textual environment". The textual environment of a word can help learners see how a word is used in different contexts, and the frequency information can indicate how many times a word has been used in a corpus and therefore how important it might be for the learner.

These two basic outputs from a corpus have had a practical and outstanding contribution to language teaching. The idea of how different types of corpora contribute to language teaching is illustrated in Gabrielatos (2005). Native-speaker corpora have an influence on corpus-based learning materials development and software design, dictionary and grammar book compilation, syllabus and coursebook design, and language test construction and evaluation (Gabrielatos, 2005). Learner corpora which reflect useful information about learner language and their learning needs have contributed to the understanding of how learners acquire a language and how language should be taught (Gabrielatos, 2005). Learner corpora are also useful for language test construction, language evaluation, and teaching and learning materials development. Corpora

compiled of coursebooks facilitate testing and evaluation of the language that learners have an exposure to, and when compared to learners' L1 and learner corpora, they are useful for teaching and learning materials production.

Römer (2008) and Stubbs (2004: cited in Flowerdew, 2009) classify the use of corpora for language teaching in two ways: direct and indirect applications. A direct application, which is the focus of this study, involves the learners' and teachers' use of a corpus in the language classroom to assist the teaching and learning processes. An example of a direct application for learners is to use a corpus as a reference resource to help themselves while writing in L2, as discussed by Lee and Swales (2006), Kennedy and Miceli (2010), and Sun (2007), and while correcting written errors as discussed in O'Sullivan and Chambers (2006) and Miceli and Kennedy (2002). An indirect application, on the other hand, refers to the use of corpora by researchers and material developers for the compilation of dictionaries, grammar books, or teaching materials. Collins COBUILD English Language Dictionary (1987) is the pioneer of this application.

1.7 Direct applications of corpora for language teaching

As mentioned earlier, in a direct application of corpora for language instruction, a corpus is used by one of the two parties, either by learners or teachers (Taljard, 2012). Learners, on the one hand, can have hands-on experience of accessing a corpus to discover the use or pattern of certain words or phrases autonomously. On the other hand, teachers can access a corpus to extract the relevant concordance lines and use them as an input for teaching or preparing materials. As suggested by Flowerdew (1996), teachers can also use concordances to check language usage

(especially useful for non-native speakers of the target language), to check high frequency words to be taught, to present learners with instances of authentic language usage when teaching a particular language point as well as to develop teaching materials. Used by language teachers as a source of native speaker's advice about the language, corpora are considered "tireless native speaker informants" (Barnbrook: 1996, 140, cited in Römer, 2008).

Moreover, in designing activities that foster learners' engagement in learning, teachers can access corpora, choose and modify the language to suit the learners' proficiency and learning needs. Realising that phraseology is of great importance to language pedagogy and that patterns are useful for learning about lexis, Hunston and Francis (2000, 272) recommend that language teachers "should be encouraged to identify patterns as grammar points for learners to notice" in concordances as an alternative to encouraging learners to notice patterns in texts. Boulton (2008a) highlights concordance applications for data-driven learning, which requires language learners to play an active role in discovering patterns of language use. Hunston (2002) says that this approach is supportive to learning since it motivates learners to remember what they have discovered, and corpus data can draw learners' attention to patterns that have been overlooked by the teacher, or not covered in textbooks. To give learners experience of data-driven learning, the teacher can either have the students work with raw corpus data with the teacher or have the students learn from the concordance lines selected and edited by the teacher. According to Hunston (2002), DDL "is most suitable for very advanced learners who are filling in gaps in their knowledge rather than laying down the foundations (p.171)". However, Boulton (2008a, 2008b, 2008c, 2009a, 2009b, 2010, 2011) demonstrates that lower level learners can also benefit from DDL tasks. Boulton (2008a) argues that most DDL research seems to introduce learners to

hands-on experience of concordancing without adequate preparation and this leads to unsatisfactory outcomes because learners are subjected to three new concepts of concordancing: new kind of input (corpus data), new ways of learning (DDL techniques), and new technical skills (using a concordancing program) all at once, rather than one a time. Boulton (2009a) explains the three new concepts as follows. First, corpus data in KWIC format must be read vertically in order to identify pattern generalizations rather than horizontally in order to identify syntactic meaning. Second, learners need to learn inductively from corpus data, rather than deductively. Third, learners need to learn how to use corpus software. For this reason, Boulton (2008a) advocates using paper-based materials at the initial stage of learner concordancing to develop the leaners' skills in reading concordances under the teacher's control. Once the learners are familiar with the concepts of concordancing, they can be exposed to the use of corpus software.

By using a DDL approach, learners need to be responsible for their own language learning by researching through concordances to find out patterns of the target language. In researching concordances, there are basically two methods (Mishan, 2004: 223). The first is the "bottom-up" approach or induction, where learners will examine concordance lines, try to discover patterns from the occurrences of the target word, and then make conclusions. For instance, they might select a word they want to study, look for its patterns in a concordance, and gain understanding of how the word is used. Conversely, in the "top-down" approach, they have some pre-formed hypothesis about language use or usage, then examine evidence, and try to find patterns to prove the hypothesis from concordances.

It should be noted here that whether to find the patterns or to test the pre-formed hypothesis, students have to think and discover facts through concordances about the language they are learning; as Gavioli and Aston (2001: 241) assert, "a concordance does not make sense in itself: sense has to be attributed to it by the reader, who must induce patterns from a concordance that will account for the data".

In the literature there are quite a lot of recommendations for using corpora in the classroom, but there is not much evidence of what the students are actually doing. Many researchers have recommended corpus use and enumerated the advantages of doing so. Therefore, one striking question posed here is "Why don't all teachers use corpora all the time?" If research keeps suggesting that it is rewarding to use a corpus in the classroom and increasingly corpus data have been integrated to language teaching materials, why do more teachers not make greater use of corpora? Some answers to this might be:

1. Suitable corpora are not available. One of the potential drawbacks of using corpora with learners is that the language presented in the corpus is difficult for learners to understand. Teachers, therefore, do not use corpora with learners because they have no way to find corpora that are suitable for the level of learners.
2. It can be technically difficult for some teachers to use corpus software. Teachers need to have computer skills to operate a concordancing program. If they find it hard to use the software, they will not use it in class. For this reason, corpus software needs to be user-friendly. Learners also need software that is easy to use.
3. Teachers lack awareness of corpus use. Teachers may not know about using a corpus for language teaching because they have not been trained for this. In addition, there are

few materials based on corpus data such as concordance lines that are available for them to use. If they wish to teach students to learn about language from a corpus, they have to write their own materials. In practice, teachers prefer to use existing course books because it is difficult for them to develop their own materials and they do not have time to do so. Teachers typically have heavy teaching loads and substantial amounts of paper work. Some teachers have limited access to the Internet and other facilities such as a computer. All these things can make it difficult for teachers to produce their own materials.

4. Although there is a great deal of research that aims to demonstrate the efficacy of DDL, there is still not enough evidence to convince teachers that classroom concordancing works or is superior to other teaching methods. As reviewed in chapter 2, there is not a huge amount of evidence to show that learners learning with corpora are very much better than learners learning with other approaches. If there was that evidence, then teachers would be much more willing to put effort into gaining expertise in this new approach. However, as classroom concordancing is a kind of new thing to teachers and there is not much evidence that using corpora for language teaching has good effects on learners, teachers may lack confidence in using it because they are not convinced that it works or not.

1.8 Learners using corpora

O'Sullivan and Chambers (2006) suggest that learners need resources that can help them write more accurately and efficiently outside the classroom, and among other resources like a

dictionary or books, Gabrielatos (2005) suggests that corpora can be a potential resource for learning lexis and grammar.

Work on learner use of corpora is very limited. Yoon and Hirvela (2004) argue that, mostly, recent research on corpus use emphasizes the use of corpora from a teacher' perspective on developing teaching materials or activities involving corpus-based orientation. In contrast, very few research studies (e.g. Watson Todd, 2001; Chambers and O'Sullivan, 2004); Gaskell and Cobb, 2004; Yoon and Hirvela, 2004) have been done to investigate learners' actual use of corpora and concordances in L2 writing. These studies largely focus on the learners' correction of errors that have been identified for them by their teachers. This may be claimed to have reduced the learners' chance of editing or revising their writing on their own, which is an essential skill for writing process and writing in the real world when the teachers are not available for them.

Due to the small number of research studies, Yoon (2008) points out that research on classroom-based corpus use and on learners' autonomous use of corpora needs urgent attention. This implies that there is a lot that has not been understood about learners' actual use of corpora both in the classroom and in everyday life. Previous research has often focused on the end product of students' use of corpora for error correction and on their perceptions or attitudes towards such experience, and has tended to neglect the in-between process. Therefore, it is still unknown whether or not learners can actually use corpora autonomously and what it is that they do that makes them fail or succeed.

This study aims to contribute to this line of research by investigating the kinds of lexicogrammatical errors that learners were able to correct using corpus data and the kinds of linguistic information they search for from a corpus when they write in English. More importantly, the study attempts to figure out the process in which learners conduct searches and interpret the results.

1.9 Research questions

My research is about using a corpus for error correction or self-editing. It was conducted as part of a writing class. When students write something in English and then go back over their work to check the accuracy of their language, that is the point at which the corpus can intervene. The ultimate goal of learners using a corpus for error correction in this study is to improve accuracy in phraseology. To achieve that goal, it is important that the learners carry out corpus searches and interpret the results with maximum expertise and efficiency. For this reason, I was curious about learners using a corpus. I wanted to know whether they could identify and correct their errors after consulting a corpus. This general interest gives rise to the following research questions.

1. What kind of lexicogrammatical errors do learners of English find it easiest to solve by using a corpus?
2. When students are writing essays, what language points are they most likely to check in a corpus?
3. What do the students do when they perform a linguistic investigation using a corpus?

1.10 Outline of the thesis

This thesis consists of eight chapters. Following this introduction chapter, chapter 2 reviews previous work on learner using corpora. Chapter 3 describes my research methodology. Chapter

4 reports on the first experiment in the series and identifies which kinds of errors the students found it easy to solve and which kinds of problems they found it difficult to solve using a corpus. Chapter 5 presents the kinds of linguistic features the students looked for from the corpus and the strategies they employed in carrying out searches. Chapter 6 highlights what might be expected to go wrong when students use a corpus independently and what they should be able to do to make their corpus searches more effective. Chapter 7 discusses implications for pedagogy. Chapter 8 concludes the study and discusses both its limitations and its potential for changing teacher behaviour.

Chapter 2

Learners Using Corpora: a review of previous studies

2.1 Introduction

Over the past 20 years, there have been many recommendations for the use of corpora in language classrooms (Tribble and Jones, 1997; Hunston, 2002; Sinclair, 2004; Aston, Bernardini, and Stewart, 2004; O'Keeffe, McCarthy, and Carter, 2007; Bennett, 2010; Reppen, 2010; Richards and Rogers, 2014; Leńko-Szymańska and Boulton, 2015). Although, in reality, corpora have not been widely applied for use with language learners as suggested by the literature, three main driving forces maintain this trend: the availability of more ready-to-use corpora; the ease of building customised corpora; the computer skills people possess (Flowerdew, 2009). Concurrent with this trend is an increasing amount of research on using corpora in the classroom for language reference and error correction in L2 writing at university level. This work has been carried out in varied contexts such as EAP classrooms in the UK, the US, Canada, Australia, and Italy; as well as ESL and EFL contexts in Taiwan, Thailand, and the US. Each of these projects varies in its focus of study and methodology used. The results, however, are sometimes slightly disappointing in terms of how significant they are. Most importantly, most of the research in this area has been restricted to the **output** of the student use of corpora. Very few studies have been carried out to discover the **process** of corpus use by learners. Likewise, most of the literature on pedagogical use of corpora is about corpus-based language studies that have been carried out and information about corpus use in textbooks or course books etc., but not about students doing corpus work. This information is useful for grounding teachers in making use of corpora for

language teaching, but it does not tell us what learners use a corpus for and how they need to be prepared before using a corpus. Therefore, despite the perceived value of learner concordancing, there is no clear understanding of how learners actually use a corpus, how easy or successful it is for them to use a corpus, and if they can learn from a corpus. As my study will show, such an understanding suggests that more needs to be done with students to enable them to use a corpus effectively before simply teaching them the techniques of using a corpus.

In this chapter, the work that has been done in relation to the use of corpora in writing activities and error correction in the language classroom will be reviewed. The chapter begins with an explanation of why learners should be encouraged to use a corpus (Section 2.2). This is followed by a review of the relevant studies, which vary in terms of the corpora used, the kinds of learners involved, research focus, research design, and the questions raised and the findings, and these variations will be used as the topic of the review (Sections 2.3 – 2.7). The chapter ends with a conclusion of how these studies contribute to an understanding of learner use of corpora and how my study can fill the gap in the existing research.

2.2 Errors, focus on form, and noticing

In learning a second or foreign language, it is probable that learners will produce the language that is different in form from the native speaker's language. These incorrect forms of the language produced by language learners are referred to by Corder (1967) as either errors or mistakes, depending on what causes them and whether the learners are able to recognise and correct them or not. If the learners produce the language that is "deviant, ill-formed, faulty, incorrect or whatever" because of their deficiency in language competence and they are unable to correct it themselves, the incorrect form is regarded as an error (Corder, 1967: 165). On the other

hand, if the learners perform the language incorrectly because of the external factors such as fatigue, strong emotion, or carelessness and they are able to correct it themselves using their language knowledge, the incorrect forms are regarded as mistakes or "errors of performance" (Corder, 1967: 166-167). Based on this notion, errors are more vitally important to the process of learning than mistakes because they are caused by a lack of knowledge and learners are unable to correct them. Regarded as a part of the language learning process, errors, in Corder's (1967) view, are significant for language instruction. First, they provide the teacher with information about the learners' progress in learning the target language, so the teacher can predict what remains for the learners to learn in order to reach the goal of learning. Second, they are useful for researchers to understand how learners learn or acquire the language. Third, errors or making errors are regarded as a device or strategy the learners use to test hypotheses about the language being learned. Therefore, in language learning and teaching, errors are not something to ignore, but to deal with. They are a benefit to the process of learning and of teaching because they provide useful input to both.

There are ways for teachers to engage with learners' errors in order to correct them. If learners produce errors because they have an incorrect hypothesis about the language and they never notice that the divergent forms wrong, the teacher can provide them with several opportunities to notice the right things. One way of doing this is to use focus-on-form approach to enable them to notice in a very intense way linguistic form they have not noticed before. Focus on form is a phrase coined by Long (1991) to refer to the learning process where the teacher directs learners' attention to language form after they have been able to get the meaning across for communication. This approach has been advocated by various people such as Schmidt (2001),

Long and Robinson (1998), Swain (1998), and Lyster (1998). This form-focused instruction has also been researched by a number of people, for example, Van Patten and Oikkenon (1996), Doughty and Verela (1998), Williams and Evans (1998), who studied this approach for grammatical rule instruction. Doughty and Williams (1998) report that this approach can also promote lexical acquisition. Focus on form contrasts with focus on forms and focus on meaning. Focus on forms is an approach to language teaching where the focus is merely on discrete grammatical rules and no attention is paid to meaning because it is believed that learning a language is to learn its grammar (Long, 1991; Long and Robinson, 1998). Focus on meaning, on the other hand, limits the teaching process meaning and no attention is directed to discrete grammatical forms of the language (Long and Robinson, 1998). Both focus on forms and focus on meaning have been criticized. Ellis (2005) points out that, through focus on forms instruction, learners may not master some structures unless the structures are practiced repeatedly. Swain (1998) and Lyster (1998) argue that focus on meaning instruction could not equip learners with grammatical competence. Based on this controversial issue, it sounds reasonable to state that focus on form is an approach that could settle the argument about what to focus learners' attention to. To help learners develop communication skills, instruction needs to direct learners' attention to both meaning and form (Ellis, 2005), and that is where DDL, a technique that supports form-focused instruction, and corpus work can intervene.

As discussed in chapter 1, learners engaging in a DDL situation need to be very active. They need to examine lots of language samples from concordance lines in order to find out fact about language use. In processing a lot of language data, learners need to be conscious and carefully notice the underlying rule from the input available to them. Therefore, noticing is one of the

characteristics of DDL. Noticing is proposed by Schmidt (1990) to refer to the process in which learners pay attention to linguistic form, and it is a very important part of DDL because it is a starting point for language acquisition. The concept of noticing is that learning takes place when learners pay conscious attention to language form (Schmidt, 1990; 1994). Although it is not unanimously agreed that noticing must necessarily be conscious, Schmidt (2001) argues that the more learners pay attention to linguistic form, the more they learn about language.

The traditional notion of errors and mistakes discussed above may be useful in describing the process of second language acquisition (SLA), but it assumes too precise a demarcation between what is and is not known. In practice there would seem to be an intermediate situation where something is partially known. In this study, there were many cases where the participants used a corpus to check the accuracy of incomplete knowledge. For example, they know that the noun *study* needs a preposition before it is followed by another noun, but they are not certain if *in* or *of* is the correct one. In this case, they have not yet committed to any of the possibilities, so what they are looking for in a corpus is correction of neither an error nor a mistake, but something in between. The students have some formal knowledge, but not enough, to draw on. Therefore, the word "errors" in this study is used to refer to whatever the students look to correct by looking in a corpus: errors, mistakes, and something in between.

2.3 Why should learners be encouraged to use a corpus?

It is not entirely clear that learners are able to learn a language from corpora and that learning a language from corpora is better than learning from other sources of language. However, there is strong evidence to support why learners should be encouraged to use language corpora. The evidence is … in the influence corpora have on language learning and teaching. Such influence

can be observed in four areas, namely; how language is perceived, what materials are made available to learners, the process of teaching, and the process of learning.

2.3.1 Views on language in corpora and language in textbooks

Because corpora provide empirical evidence of language in real use, the language they represent is seen as authentic, and corpora are highly praised for language authenticity. Krieger (2003) and Bennett (2010) highlight that the core benefit of using corpora is to uncover patterns of language as it is actually used in real-world contexts. This point is supported by O'Keeffe, McCarthy, and Carter (2007) and Mull (2013) who consider that authentic language, rather than invented or contrived examples, is an important source for students to learn from in order to improve their advanced communication skills.

Concurrent with the recognition that authentic language is of vital importance to language learning, the use of intuition-based teaching materials has been extensively criticised. A powerful argument against using intuition for language description is that it is unreliable (Sinclair, 1991; Hunston, 2002; Bernardini, 2004; Flowerdew, 2009; Bennett, 2010). Sinclair (1991) argues that "human intuition about language is highly specific and not at all a good guide to what actually happens when the same people actually use the language" (p. 4). This means that the way people actually use language differs from how they think language should be used, so they are not able to be accurate about their own language use, let alone intuiting the language used by the majority of speakers. Intuition may be helpful for some common aspects of a language shared among its speakers such as sentence structure and grammatical patterns, which do not vary from place to place or person to person. When it comes to making judgments about certain aspects of language such as collocation, frequency, prosody, and phraseology, native speakers' intuition can be less

helpful (Hunston, 2002). Hunston gives the following examples to illustrate her point. For example, native speakers as well as language learners may not be aware of some particular adverb-adjectives collocations like *acutely aware*, *painfully clear*, and *vitally important.* Likewise, it is sometimes easy to intuit that *take* is more frequent than *disseminate*, but it is sometimes difficult to say which word e.g. *fare* or *fantasy* is more frequently used. In terms of semantic prosody, it is beyond the native speaker's intuition to learn that "the phrase *par for the course* is used not only to comment that something frequently happens, but also to evaluate that event negatively" (Hunston, 2002: 21). With respect to phraseology, the pattern *require to be* + V-ed is acceptable, but, according to Owen (1996, cited in Hunston, 2002: 21), it can be hard for native speakers to intuitively explain why such a pattern found in the Bank of English corpus in a sentence such as *"Further experiments require to be done"* sounds incorrect. This hints at Oven's argument that corpus data in the Bank of English can be unhelpful because they suggest that *"required to be done"* would be correct, whereas intuition says it is not. Hunston looks into the Bank of English and explains this mismatch in terms of the kind of verb that is used after *"required to be",* thereby reconciling Owen's intuition with the corpus evidence. She notes that the past participle (V3) that follows *require to be* is usually a verb with specific meaning as found in the example like *"These roses require to be pruned each spring"*, but very few examples with the verb *do*.

2.3.2 Material development

In terms of teaching materials, corpora have been used for creating dictionaries and books. O'Keeffe, McCarthy, and Carter (2007) point out that most of the current English learners' dictionaries are based on large language corpora and that corpus evidence is integrated into the

development of corpus-informed teaching materials. Examples of learner dictionaries that claim to be corpus-based are the *Collins COBUILD English Dictionary*, the *Longman Dictionary of Contemporary English* (third edition), the *Oxford Advanced Learner's Dictionary* (fifth edition), and the *Cambridge International Dictionary of English* (McEnery, Xiao, and Tono, 2006). Examples of corpus-informed coursebooks include the *Touchstones* series and the *English Phrasal Verbs in Use* written by McCarthy and O'Dell, 2004 (Richards and Rodgers, 2014).

These materials are created mainly because of the criticism that the language presented in textbooks does not reflect actual language use and that scripted dialogues, for example, cannot effectively enhance learners' communication skills (Carter, 1998; Burns, 2001; Burn, Joyce and Gollin, 2001; McCarthy and O'Keeffe, 2004; Thornbury and Slade, 2006 , cited in O'Keeffe, McCarthy, and Carter, 2007: 21). O'Keeffe, McCarthy, and Carter provide arguments and reports from many studies to support this criticism. For example, Burns (2001) argues that scripted dialogues reflect patterns and features of natural spoken discourse only to a limited extent. This can hinder students' ability to hold a conversation in real and unanticipated situations outside the classroom. Carter (1998) studies textbook dialogues and finds a lack of main language features of spoken language, i.e. discourse markers, vague language, and hedges, compared to those found in the Cambridge and Nottingham Corpus of Discourse in English (CANCODE). In a similar case, Gilmore (2004) compares spoken discourse features in seven dialogues in textbooks from 1981–1997 with corpus data and reports a significant difference between discourse features, e.g. turn patterns, lexical density, repetitions, pausing, hesitation found in corpora and textbooks. He also examines these features in more recent textbooks such as the Touchstone series by McCarthy, McCarten, and Sandiford (2005a and b; 2006a and b) and

illustrates that the writers are trying to embed discourse features of spoken language shown in corpus evidence in dialogues. An example of the mismatch between the prescription of language rules found in grammar books and the actual use in corpus data can be found in Kettemann (1995), who points out the change of tenses to the past in reported speech that is not actually the case as prescribed in pedagogical grammar.

2.3.3 New views on language

According to Hunston (2002), corpora call into question traditional units of language like phrases and clauses and bring about new concepts or aspects of language such as phraseology, lexico-grammar, register, nuances of language, collocation, frequency, prosody, colligation, and semantic preference (Hunston, 2002; Flowerdew, 2009; Bennett, 2010). Thus, the main reason for encouraging learners to use a corpus is to enable them to observe these language aspects, which, according to Flowerdew (1993, cited in Flowerdew, 2009: 334), are not as salient in other sources of language like a dictionary or grammar book as in a corpus. The concepts and examples of these aspects are directly quoted below.

1) Phraseology: the study of phrases (Bennett, 2010), which includes:
    - Collocation: the statistical tendency of words to co-occur, e.g. *big deal, good deal, great deal*
    - Lexical bundle: a recurring sequence of three or more words (Biber, Johansson, Leech, Conrad, and Finegan, 1999: 990, cited in Bennett, 2010: 9), e.g. *Do you want me to, I don't know what*

- Preferred sequences: preferred sequences of words, e.g. *someone is interested in something, an interesting thing, what is interesting, it is interesting to see* (Hunston, 2002: 9-11, cited in Bennett, 2010:9)

2) Lexico-grammar: Sinclair's (1991) idea that there is no difference between lexis and grammar, or that lexis and grammar are so closely intertwined that they cannot be productively studied separately (cited in Bennett, 2010: 10)
3) Register: situation of use (Bennett, 2010); how patterns may vary across various registers or genres (Flowerdew, 2009)
4) Nuance of language: questions that students might ask that we just don't know the answer to (Bennett, 2010)
5) Colligation: the collocation between a lexical word and a grammatical one, e.g. *head <u>of</u> department, throw <u>one's</u> head back* (Hunston, 2002: 13)
6) Semantic preference: how a word or phrase relates to a group of collocating words that (1) share an element of meaning, (2) are related to particular genres or registers, or (3) belong to lexical sets in terms of synonymy, meronymy, antonymy, etc., e.g. words or phrases relating to measurement, or words or phrases relating to history (Flowerdew, 2009: 332)
7) Frequency: differences in frequency between different words on a particular genre (Hunston, 2002)
8) Semantic prosody: a word that is typically used in a particular environment (Hunston, 2002: 141), e.g. the word *cause* typically collocates with negatively loaded words – e.g. *accident, concern, damage, death* – and thereby takes on a negative semantic prosody (Stubbs, 1996, cited in Flowerdew, 2009: 333)

There is enough evidence from the research being cited that, in general, using a corpus with learners would seem to be a good thing because learners can find some language features that cannot be easily found from reference books. This also provides grounds for hypothesising that using corpora in the classroom is beneficial, but it seems that there is not enough empirical evidence to demonstrate that learners can benefit from corpus data and are able to observe these features from corpora.

2.4 Kinds of corpora

Three kinds of corpora are used in research on corpus use for language learning: ready-made corpora, corpora compiled by the teacher/researcher, and corpora compiled by the students. Ready-made corpora employed by the first group of researchers are the Brown Corpus (Gaskell and Cobb, 2004), the Collins COBUILD Corpus, known as the Bank of English (Yoon and Hirvela, 2004; Yoon, 2008), and the *Independent* corpora (Cresswell, 2007). In addition to these corpora, language in the World Wide Web (WWW) can be classified as one of the ready-made corpora. Watson Todd (2001) introduces his students to making concordances from the Internet via FAST Search http://www.alltheweb.com. In the second group of studies, the researchers create their own corpora of familiar written texts for use with learners. Turnbull and Burston (1998) created corpora of students' own writing (1000-2000 words). Kennedy and Miceli (2001; 2010) compiled an Italian corpus of familiar texts including emails, newspapers, and magazine (called CWIC: Contemporary Written Italian Corpus) to use in their studies. Chambers and O'Sullivan (2004) and O'Sullivan and Chambers (2006) created a French corpus of expert

writing relevant to students' writing assignments in French. Compared with the first group, corpora of this kind are relatively small in size.

In the third group of studies, learners compile corpora of research papers in their own disciplines (specialised corpora) by themselves for their own use. For example, in Sun's (2007) study, the participants use a corpus of academic writing jointly compiled by themselves.

In some studies, a combination of two kinds of corpora is used. For example, Lee and Swales (2006) use Hyland's Research Article Corpus, MICASE (Michigan Corpus of Academic Spoken English), academic texts from the BNC (the British National Corpus), and student-compiled corpora. Charles (2012) uses two corpora of theses compiled by herself to familiarise the students with the idea of concordancing at the initial stage of the study, and the students compile their own corpus of research articles for their own use afterwards. Her study, then, focuses on the students' evaluation of their experience of creating and using their own corpus. In Phoocharoensil (2012), the type of corpora used in the corpus-based grammar instruction is not specified. It can be seen that, among these three types of corpora, corpora created by teachers/researchers have received the most attention in recent research on the use of corpora for assisting foreign language writing and error correction. Additionally, there is an increasing tendency for teachers to train students to build their own corpus as a reference tool for writing. A reasonable explanation for this may be that most of the research in this area has been done with mixed groups of student writers of various disciplines who are trained to be expert writers in their own fields of study. Therefore, it is hard for teachers to respond to individual students' language needs and to provide students with existing genre-specific corpora that meet their immediate needs. What matters here is that research on learners of general English using a general or

available corpus of English for error correction and language acquisition is still rare and more attention is needed.

2.5 Types of students

Research on corpus analysis by language learners has been conducted with different types of learners in terms of levels of study, levels of language proficiency, and ethnic backgrounds. These three elements are taken as a basis for discussion.

With respect to levels of study, two groups of students are identified: postgraduate and undergraduate students. However, most of the studies have been done with postgraduate students, both doctoral (Lee and Swales. 2006), and Master's (Turnbull and Burston, 1998; Watson Todd, 2001; Chambers and O'Sullivan, 2004; Gaskell and Cobb, 2004; Yoon, 2008). For example, Lee and Swales (2006) did a study with 4 PhD students in an experimental EAP course in the US. Turnbull and Burston (1998) studied 2 non-native MA students' concordancing strategies (one Japanese and one Indonesian) in Australia. Watson Todd (2001) did his study with 25 science and engineering Thai postgraduate students (lower intermediate to intermediate) in Thailand. Chambers and O'Sullivan (2004) investigated error correction made by 8 graduate native-English speaking students in Ireland to improve their accuracy in writing in French. Gaskell and Cobb (2004) did a study with 20 Lower-intermediate L2 Chinese learners of English taking a writing course in Canada. Although all the participants' degree of study while taking part in the study is not directly stated, it could be assumed from their educational background (undergraduate degree from China) and from their age range (18–50) that most of them were postgraduate students who had received their first degree in China. To extend the study by Yoon and Hirvela (2004), Yoon (2008) conducted a longitudinal case-study involving 6 ESL

postgraduate students using a corpus in an EAP writing course in an American university. Unlike those who focus on corpus use for correcting language errors in writing, Phoocharoensil (2012) conducted an attitude survey to examine how 17 Thai students of English at Master's level felt about learning grammar through corpus-based instruction.

Compared to those studies conducted with postgraduate participants, quite a small number of studies have researched corpus use by undergraduate language learners. Cresswell (2007) conducted a study with advanced third-year students taking an English Language and Linguistics course required by their degree in Translation or Interpreting in Italy. These students were divided into a DDL group and a non-DDL group. To compare the results from their previous study with the postgraduate students in 2004, O'Sullivan and Chambers (2006) undertook research with 14 undergraduate students learning French in Ireland to discover the similarities and differences in corpus use and perceptions of using the corpus between the two groups of learners. Kennedy and Miceli (2001) compiled the Contemporary Written Italian Corpus (CWIC) for use in an Italian writing course and gradually trained the students to use it for error correction. The training program is called the *apprenticeship approach*. In evaluating how effectively the students used the corpus, they carried out the study with 10 intermediate undergraduate students taking an Italian writing course in Australia. The insights gained from this study inspired them to initiate a new apprenticeship approach to training in using a corpus which is more straightforward as reported in Kennedy and Miceli (2010), a case-study involving three selected undergraduate Italian-major students.

In some studies, the participants are a mixed group of doctoral and Master's students (Sun, 2007; Charles, 2012) and postgraduate students and undergraduate students (Yoon and Hirvela, 2004).

Sun (2007) also researched the effectiveness of an online corpus tool, called the Scholarly Writing Template (SWT), designed to provide his 20 postgraduate students (19 doctoral students and 1 Master's student) in Taiwan with ideas of content development and language input commonly used in academic writing in English. Charles (2012) conducted an attitude survey with 50 advanced postgraduate students in the UK, the majority of whom (63%) were doctoral students. Yoon and Hirvela (2004) carried out a study on the students' perceptions of concordancing with two different proficiency levels at the same time in the US: 8 students (4 undergraduates + 4 postgraduates) from an intermediate academic writing course and 14 postgraduate students from an advanced academic writing course. In total, out of 22 participants, 18 of them were postgraduate students.

In terms of levels of language proficiency, most of the studies are carried out with advanced learners or with learners undertaking advanced language courses (Turnbull and Burston, 1998; Sun, 2007; Yoon, 2008, Cresswell, 2007; Yoon and Hirvela, 2004; Chambers and O'Sullivan, 2004; Kennedy and Miceli, 2010; Lee and Swales, 2006; Charles, 2012; Phoocharoensil, 2012). In fact, many of these advanced learners can be viewed as language users rather than language learners because English is used as the medium of instruction in their postgraduate study. These students might be more motivated to use a corpus as a reference tool for writing in L2 and correcting errors because they have urgent and real needs to express their thoughts through writing in the target language. On the other hand, non-advanced students, especially those who are taking language courses for their undergraduate degree can be well defined as language learners and perhaps they are not as motivated to learn a language as postgraduate students unless they are language students. This group of learners needs resources for acquiring the language,

and it is challenging to provide them with corpus resources and concordancing skills so that they can learn by themselves. However, only a small number of studies have been conducted to involve lower intermediate – intermediate learners in concordancing (Watson Todd, 2001; Kennedy and Miceli, 2001; Gaskell and Cobb, 2004; O'Sullivan and Chambers, 2006). There are even fewer studies into the use of corpora by undergraduate language students, who are more motivated to learn about language than non-language students.

With respect to ethnic background, it is somewhat surprising to learn that most of the students taking part in the studies are Asian nationals (Thai, Korean, Taiwanese, Chinese, Japanese, and Indonesian) and a few of them are Europeans (Italian, Irish, and Romanian), despite the fact that the research has been conducted in different geographical areas (Taiwan, Australia, Thailand, Canada, Ireland, Italy, Australia, the US, and the UK). Most of these students are learning English, but some of them are learning other languages such as Italian (Kennedy and Miceli, 2001; 2010) and French (Chambers and O'Sullivan, 2004; O'Sullivan and Chambers, 2006).

In sum, the participants in these studies vary not only in the level of degree they are pursuing while taking part in the studies (undergraduate – postgraduate) and levels of language proficiency (lower intermediate – advanced), but also in ethnic origin. Another important issue raised by the information about the participants of these studies is that most of these studies involve a mixed group of participants of different backgrounds in terms of age and courses of study while participating in each of the studies. Very few studies have been carried out with homogeneous groups of learners of English. Given that corpora are particularly more useful for linguistic students, who are observant language learners, it is assumed that these students benefit more from corpora and use them more effectively than non-linguistic students. Still, research that aims to

provide an understanding of how homogeneous groups of English-major students use a corpus is relatively rare as most of the studies have involved heterogeneous group of learners of English from various disciplines. The study by Cresswell (2007) is the only study found to have involved a homogeneous group of learners of English, but its focus is on exploring the use of some connectors, and not on attempting to understand how learners use a corpus in general.

2.6 Focus of study

Research on corpus use for error correction in writing focuses on four main areas. The first group of projects focuses on learners' self-correction of errors (e.g. Watson Todd, 2001; Gaskell and Cobb, 2004; Chambers and O'Sullivan, 2004, O'Sullivan and Chambers, 2006). The focus of the second kind of research is on a corpus as model (Sun, 2007; Yoon, 2008; Charles, 2012). The third focus is on students' attitudes towards and evaluation of corpus use (e.g. Yoon and Hirvela, 2004; Kennedy and Miceli, 2001; 2010; Sun, 2007; Charles, 2012, Phoocharoensil, 2012). The last focuses on the effects of corpus consultation on language learning (Thurnbull and Burston, 1998; Lee and Swales, 2006; Cresswell, 2007).

2.7 Research design

Different research methods have been employed in the related studies. Case-studies, among these methods, are most frequently used, for example, by Turnbull and Burston (1998), Lee and Swales (2006), Yoon (2008), and Kennedy and Miceli (2010) to discover the individual students' ways of using corpora. A survey methodology, perhaps because of its ease of application, is another kind of repeated research design adopted by Yoon and Hirvela (2004), Sun (2007), Charles (2012), and Phoocharoensil (2012). The general purpose of these surveys is to elicit the students' reactions to or attitudes towards their experience of using available and specially designed

corpora for language learning and error correction in their own writing. A quantitative empirical method is adopted in the studies where the main focus is on the students' ability to self-correct their own errors (Watson Todd, 2001; Chambers and O'Sullivan, 2004; O'Sullivan and Chambers, 2006). A comparative method of pre- and post-test design (Gaskell and Cobb, 2004) and DDL and non-DDL groups (Cresswell, 2007) are adopted in the studies aimed at evaluating the effects of corpus consultation on language improvement and development. A retrospective study, the least frequent approach taken by Kennedy and Miceli (2001), is used to discover how learners carry out corpus investigation for error correction.

As noted below, the empirical studies by Watson Todd (2001), Chambers and O'Sullivan (2004), and O'Sullivan and Chambers (2006) provide interesting statistical evidence of the students' success in error correction with the assistance of corpus investigation. The comparative studies by Gaskell and Cobb (2004) and Cresswell (2007) also provoke debate about whether it is better to encourage learners to learn a language through corpus-based tasks than through traditional methods. The case-studies by Turnbull and Burston (1998), Lee and Swales (2006), Yoon (2008), and Kennedy and Miceli (2010) are advantageous in that they allow for in-depth data to be gathered from the participants. One main drawback of this kind of study is that it employs a very small number of participants and the results obtained cannot be generalised. As opposed to this, surveys enable the researchers to collect data from large groups of participants in a more manageable way. The data obtained, however, may provide only a superficial understanding unless they are complimented by other sources of data such as interviews.

Of all these research methods, the retrospective study by Kennedy and Miceli (2001) contributes a great deal to my study that seeks to find out what learners actually do when they are

investigating a corpus to self-correct errors in their writing. The difference between their research and my research is the data collection procedure. The data collection in the study by Kennedy and Miceli is more learner-oriented. They attempt to understand how learners use corpora from the students' point of view through their written accounts of corpus work and the follow-up interviews, supported by the evidence from the video recordings of the students consulting a corpus while revising their writing. The only criticism of this data collection method is that the students might not remember all the major steps taken, so their accounts could be distorted. To keep a balance, I have made my study more teacher-oriented by employing a more innovative way of looking at the students' process of investigating corpus data from the teacher's point of view through the students' think aloud protocol and recordings of on-going computer screens.

2.8 Research questions and findings

The existing research has sought to find out the answers to five main questions: 1) the feasibility of learner use of corpora, 2) the effectiveness of students' use of corpora, 3) how students use a corpus, 4) students' evaluation of and attitudes towards corpus use, and 5) factors that mediate corpus use. Each group of research questions and the results are explained below.

2.8.1 The feasibility of learner use of corpora

There is a group of studies where the focus is on error correction and the purpose of the research is to test the possibility of encouraging learners to use a corpus for error correction. The research questions raised in these studies vary according to the researchers' views on corpus use. Gaskell and Cobb (2004) pose a very basic question about learner use of concordances for error correction. They ask whether learners can use a corpus to correct their own writing. It is quite

clear that the question itself does not raise any other issues in particular, but the students' ability to correct the errors in general. Turnbull and Burston (1998) link corpus use to the concept of learner autonomy by asking *"*How far can students in a particular educational setting take charge of their own learning?*"* (p. 10). They also go further by raising the question of to what extent this method can be integrated with existing teaching methods. Thus, the importance of this research is that it raises the possibility of bridging the gap between the corpus-based language teaching method and other teaching methods currently adopted. To obtain more convincing answers, Watson Todd (2001) and Chambers and O'Sullivan (2004) conducted quantitative research to discover if it is likely that learners are able to learn from concordances and to correct their own errors based on the findings from the corpus output. Watson Todd (2001) makes a strong link between inductive learning and concordancing. His research aims to find out if the language patterns the students induce from the concordances selected by themselves are valid and useful. He also increases the value of the study by attempting to discover the correlation between the students' ability to induce valid patterns, the ability to apply the induced patterns, and the ability to self-correct their errors. Chambers and O'Sullivan (2004) who introduce a corpus to their learners of French investigated whether learners could improve language accuracy in writing through corpus consultation and studied in detail the types of successful changes the students made as a result of consulting a corpus of French compiled by the researchers.

These questions, in broad terms, reveal if learners are likely to use corpus resources to self-correct their errors. In more specific terms, they raise practical issues of what types of errors learners try to correct using a corpus and what lexicogrammatical errors learners are likely to tackle with ease with the assistance of corpus data.

Findings

Despite some of the unsatisfactory results and the difficulties learners encounter, the researchers hold positive views about the feasibility of learner concordancing. The findings by Gaskell and Cobb (2004) at the initial stage are more encouraging than those of the final stage. They report that the students, during the first four weeks, are able to correct over 80%–100% of the errors when they are provided with the concordance lines relevant to specific errors. When they use the corpus independently to correct the errors in their writing in the following weeks, the results show that 60%–70% of the error correction submitted for analysis are made correctly during weeks 6–8. However, during weeks 9 and 10, less than 50 % of the errors are successfully corrected. The researchers attribute the decrease in the rate of successful error correction to the students' worry about the upcoming exams. However, the results imply that learners may have difficulty identifying concordance lines relevant to the errors when using a corpus independently.

A review of case studies of learner autonomy of two students using a corpus yields differing results. Turnbull and Burston (1998) found a completely different degree of success in using a corpus independently to correct the errors chosen by the students themselves. The Japanese student makes good and rapid progress in searching the corpus and investigating the corpus data. At first, she is not quite sure how concordancing is useful. Once she recognises in what ways it is useful for language discovery and gets used to inductive learning, she is able to carry out many productive searches systematically and comfortably and make several useful observations. As opposed to this, the Indonesian student perceives that his corpus use is ineffective and he spends less than half the time spent by the Japanese student on concordancing. He finds the corpus output too overwhelming to stimulate him to learn and reports that he cannot make any successful investigation. In total, he incidentally makes only two useful observations when using the corpus

independently. Turnbull and Burston conclude that it does not necessarily follow that investigating concordance lines will stimulate inductive learning and that individual differences between learners play a key role. Whether or not learners will be motivated by and successful in concordance investigation largely depends on their abilities to deal with inductive learning and familiarity with such an approach. The results also convince them that concordancing activities can be integrated with the traditional approach to language learning as the participant who is more motivated and familiar with inductive learning benefits from corpus use. To help learners with different learning styles and preferences use a corpus effectively, Turnbull and Burston suggest providing them with gradual training to guide them through the process of inductive corpus investigation.

As opposed to the conflicting results in the above studies, Watson Todd (2001) discovers that, overall, the results of the students' inducing language rules from their self-selected concordance lines and applying those rules to error correction in their own writing are positive. The students are able to observe valid patterns from the examples and correct their errors effectively using the induced patterns. There is also a high positive correlation between the students' ability to induce sensible language patterns and the ability to self-correct their errors, meaning that the students will be able to correct their own errors if they can induce language patterns. Nevertheless, these two abilities can be affected by the part of speech of the target words whereas the number of patterns of usage and the number of meanings of the target lexical item can influence the students' ability to apply the induced pattern in self-correction of their errors. The results of this study are statistically dependable and interesting but, to some extent, are not convincing to Yoon (2011), who argues that the positive results may be affected by the fact that the participants select

only ten concordance examples comprehensible to them and ignore the lines that are difficult for them to understand. This argument is not wholly convincing. In reality, in dealing with hundreds of concordance lines, learners need to be selective in choosing the lines that are relevant and make sense to them. It is reasonable for learners to ignore the lines that are beyond their ability to understand and the real purpose of learner concordancing is to discover facts about language that they want to find from the corpus output. If they are able to correct patterns that fit the contexts of use, they accomplish the goal of concordancing.

The results of a more specific study on the types of changes made based on the corpus data also suggest that concordancing is useful for self-correction. Chambers and O'Sullivan (2004) discovered that corpus consultation helped their students make correct changes to their written errors in French. Overall, 75% of changes the students made were correct. The most frequent changes that were made correctly by using a corpus are grammatical errors (gender and agreement, prepositions, verb forms/mood, use of the negative, syntax), misspelling, and lexico-grammatical patterning.

From these studies, it seems reasonable to assume that the possibility of learner concordancing for error correction depends heavily on the extent to which learners have control over their concordancing activities and on learners' inductive learning skills, especially in reading and interpreting concordance lines. Learners tend to become more successful in error correction when the concordances are selected and provided by the teachers. When they have to make their own concordances and have more or full control over their concordancing, the chances of making successful correction can be lessened. In a worse case, when they are given freedom to use a corpus, they can be demotivated and less likely to use it to correct their own errors. Despite the

difficulties of concordancing for some learners, none of the studies express a pessimistic view on learner concordancing for error correction. Instead, the researchers take the view that appropriate training can improve learners' use of corpora.

2.8.2 Effectiveness of students' use of corpora

With the intention of enabling learners to use corpora for checking language patterns and usages while composing and correcting errors, researchers probe the more general question of how effective it is for students to use corpora. Nonetheless, what is meant by effective varies according to the focus of the study. The most common question is raised by O'Sullivan and Chambers (2006): how to evaluate effectively students' use of a teacher-compiled corpus to assist their writing in French. To answer this question, the end product of successful and unsuccessful changes the students made as a result of consulting the corpus was taken into account. Another study that emphasises the end product of the student use of a corpus is by Cresswell (2007). This study is different from O'Sullivan and Chambers (2006) in that it focuses on the corpus-based descriptions of the target items (connectors) the students made rather than on the actual use of the language items in writing. To put it simply, the study focuses on whether the descriptions of connector usages and uses are accurate or good enough to make data-driven learning possible. However, Cresswell extends the analysis to discover whether the students actually use the target connectors in their writing after the participation in data-driven-learning activities as he asks "After 'communicative DDL' is there more 'genuine use' of the connectors studied?"(p. 272). To Gaskell and Cobb (2004), an effective use of corpora by students, in a sense, means the students produce fewer errors after having recourse to corpus investigation, so one of the research

questions they ask is whether or not correcting errors using a corpus will reduce the number of errors in free writing.

While the above-mentioned studies take a quantitative approach to measure the effectiveness of the students' use of language corpora, other researchers take a qualitative approach to evaluate how effectively the students use corpora and shift their attention to the process of concordancing and learning. Yoon (2008) focuses her attention on the influence of concordancing on learning and raises the question of how concordancing affects the students' language learning and L2 writing approaches. In other words, she wanted to know how the students understand language learning (the concept of learning a language, e.g. vocabulary and grammar) and if the students' writing process changed as a result of corpus use. The research by Kennedy and Miceli (2001; 2010) provides real insights into the effectiveness of the ways students process concordance lines. Kennedy and Miceli (2001) examine in detail the process of concordancing trials the students underwent while using the Contemporary Written Italian Corpus (CWIC), a specially-designed Italian corpus, to correct errors in Italian. They then identify what goes wrong at each step and propose what should be done to avoid the pitfalls of concordancing process. Because the study in 2001 appeared to be too intellectually demanding for the students, Kennedy and Miceli (2010) conducted a more straightforward study to scrutinise the two functions of concordancing for language learning. The three students were also provided with monolingual and bilingual dictionaries and grammar books as a supplement to corpus data. The questions they ask are what functions the students use the concordances for – for pattern-hunting (exploring language patterns unknown or misused) or for pattern-defining (finding examples or testing the

hypothesis about the language when the students have the target patterns in mind) – and how effectively the students can do this.

In sum, O'Sullivan and Chambers (2006), Gaskell and Cobb (2004) and Cresswell (2007) measure the effectiveness of learners' concordancing by looking at the students' end product of using a corpus. On the other hand, Kennedy and Miceli (2001; 2010) and Yoon (2008) employ a more qualitative approach to measure effectiveness. Therefore, insights gained from this work could be expected to be well-balanced.

Findings

O'Sullivan and Chambers (2006) make the correct assumption that corpus consultation seems to be more effective for certain types of errors and for specific group of learners. Their study reports that about 73% of the errors are accurately corrected. It appears that the common errors undergraduate students make and are able to correct most effectively using the corpus are grammatical errors (prepositions, articles, gender and agreement and verb forms) and lexical errors (word choice and informal usage) respectively. A comparison with the results obtained from the postgraduate students in Chambers and O'Sullivan (2004) shows that corpus consultation was useful for the discovery of lexical and grammatical patterns for both groups of learners. For the undergraduate students who seem to encounter more difficulties and become more overwhelmed than the postgraduate students, especially in coping with the numbers of concordance lines, it was found to be particularly useful for correcting preposition errors.

In contrast to the earlier findings by O'Sullivan and Chambers (2006), however, the data-driven-learning activities in Cresswell (2007) were found to be moderately effective. Only 53% (8 out of

15 groups) of learners in the DDL group gave accurate corpus-based descriptions of how the specified connectors are used. The most striking results emerge from the comparison of the genuine use of the connectors focused on in the study between the DDL group after having completed the DDL tasks and the non-DDL group. There is no noticeable effect of DDL on the use of connectors in terms of variety of position in a sentence and quantity of use. These findings correspond to the study by Sripicharn (2003) who compares the effects of learning language items through corpus-based instruction and through traditional method by Thai learners of English, and discovers no marked effect of classroom concordancing on language learning. To some extent, these results are not very encouraging.

Similar to Cresswell's (2007) findings, no decrease in the number of errors as a whole was detected in the comparison of the pre- and post-test writing on the same topic under the same circumstances by Gaskell and Cobb (2004). However, by categories of ten error types, seven types of errors reduced: gerunds/infinitives, noun plurals, prepositions, modals, capital and punctuations, word order, and pronouns. Of these, the last three types decreased significantly. Two types of errors substantially increased (articles and subject-verb agreement) and one increased significantly (noun pluralisation). Contrary to the researchers' expectation, the overall increase in the amount of errors in the post-writing does not indicate that it is fruitless for teachers to train students to use a corpus as eight out of twenty individual students produced fewer errors. As pointed out by the researchers, the rise in the number of errors is due to the longer texts the students produce. A possible and sound explanation for these results may be that retention of a learned word may depend on the extent to which the word has been used by the learner. In this case, corpus searches might not guarantee the students' retention of the search

items as the purpose of their searches is to serve their particular needs at that time. Therefore, there is no point comparing the number of errors in the pre- and post- writing unless there is proof that, in the post test, the students actually use the same vocabulary items they have previously searched for and the descriptions made are valid, but they still make the same errors. A very simple way to prove the effect of corpus use is to examine whether the assigned errors are accurately corrected as mentioned previously in the report of their study in 2.7.1 and as found in other studies, e.g. Watson Todd (2001).

Despite the unconvincing quantitative results in the study above, it would appear that corpus consultation has the desired effect on language learning. In addition to helping the students solve lexicogrammatical errors, corpus consultation, according to Yoon's (2008) study, helps raise the students' language awareness and leads them to learning about the language. Obviously, the students were found to become more aware of word collocation and pay more attention to it. More interestingly, corpus consultation seems to change the students' views of language. One student who initially held a strong belief that learning language equates to learning grammar and that words and grammar are not related later grasped a mixed concept of "lexico-grammar" where words and grammar need to be viewed as closely related. Two (out of three) of the students felt that they gained more confidence in their writing. Regarding the effect of corpus use on the students' approach to writing, Yoon found that corpus use does not change the students' writing process (outlining or drafting, writing, and editing). However, the students reported that, during the writing and editing stages, they paid more attention to word usage and collocation when writing without a corpus. Therefore, the main change is that corpus use helps

individual students develop editing skills and habits and increases their confidence in writing quality.

Strong evidence of how the students use a corpus from the beginning to end and how effectively they use it is found in Kennedy and Miceli (2001). The researchers gave the students training in using the specially-designed Contemporary Written Italian Corpus. The way the students were trained to use this corpus is called *the apprenticeship approach*. After training, the students used the corpus to correct errors in two texts, and the researchers examined in detail how the students went about using a corpus while correcting errors underlined by the teacher in the first text and errors identified by the students themselves in the second text. The findings show that the students often followed four steps during this process: "(a) formulating the question; (b) devising a search strategy; (c) observing the examples found and selecting relevant ones; and (d) drawing conclusions" (p. 81). Problems and difficulties the students encountered in each of the steps were identified. The researchers observed that the students were able to make many useful observations from the concordance lines, but they were not aware of what had gone or could go wrong with their investigation and what should be done to avoid the problems or to achieve a plausible outcome. In response to the findings, Kennedy and Miceli give advice on how to avoid the pitfalls of concordance investigation at each of the steps. For example, in formulating the question in the first step, students need to state the question clearly and make sure that it is specific enough for the situation and so forth. The results of this study lead to the conclusion that it is difficult for learners to use a corpus because it requires substantial researching skills, and the students do not know how to avoid problems and difficulties they have. The researchers remark

that the students need more training in how to observe concordances and make a logical conclusion to develop concordancing skills.

Inspired by their informed judgment about learners' low efficiency and difficulties in interpreting concordance output from the previous study, Kenedy and Miceli (2010) further studied three students using the Contemporary Written Italian Corpus and a bilingual dictionary. Kennedy and Miceli identify two basic functions of using a corpus and a dictionary, which do not require learners' high language proficiency level: pattern-hunting and pattern defining (see 2.7.2) and introduce these functions to the students. Pattern-hunting is used for enriching content and language for writing and pattern-defining is used for improving language accuracy. "Both pattern-hunting and pattern-defining functions entail exploring the corpus in search of models for word patterns to employ in one's own text (adapted as necessary), but their departure points differ" (Kennedy and Miceli, 2010: 31–32). To compare which functions the students used the corpus and dictionary for and how effective their use was, the researchers asked the students to modify and add more information to drafts of their autobiography in Italian using the corpus. The findings from the interviews, computer-screen recordings, and discussion suggest that all three students used the reference resources in different ways. Only one student (S1) used both functions of concordancing i.e. pattern-hunting and pattern-defining. This student carried out eleven pattern-hunting searches and two pattern-defining searches. The other two students (S2 and S3) used the corpus for pattern-defining operations only. S2 rarely used the corpus and preferred the dictionary. She used the corpus for pattern-defining purposes only once and was not successful. S3 successfully conducted two pattern-defining searches using the corpus. While these two functions were not extensively used by individual students, Kennedy and Miceli

identify a new way of using a dictionary and corpus for checking an Italian equivalent for a given English pattern, called the "find-an-Italian-equivalent operation" (Kennedy and Miceli, 2010: 38). This new function, which Kennedy and Miceli had not anticipated, was significantly used by all three students. S1 performed three dictionary searches for Italian equivalents, one of which was successful. S2 was successful in using the dictionary for finding two equivalents in Italian. S3 was an extensive user of this technique. He conducted eleven searches: five with the dictionary only and six with a combination of both the corpus and the dictionary. Overall, of these students, two (S1 and S3) achieved a relatively high degree of success, and they felt it was rewarding to use the corpus. S2, on the other hand, was moderately impressed with corpus use, and she found it grueling to deal with corpus data compared to a dictionary, which she felt more comfortable with and made most progress from. According to this study, whether or not students are successful in using a corpus, they seem to prefer a dictionary. Hence, Kennedy and Miceli suggest that learners need to be taught to appreciate the nature of corpus use in that it does not always provide a satisfactory outcome as well as to learn to conduct searches and interpret the results effectively.

Taken together, the quantitative findings in O'Sullivan and Chambers (2006), Cresswell (2007), and Gaskell and Cobb (2004) imply that learners' independent use of language corpora is effective only to a limited extent. For example, there is no difference in the students' use of connectors between the DDL and non-DDL groups in Cresswell (2007), and there is no decrease in the number of errors in writing in pre- and post- concordancing tasks in Gaskell and Cobb (2004). One might conclude that it is pointless to promote learners' concordancing if it makes no difference to their language development. On the other hand, the qualitative findings in Yoon

(2008) and Kennedy and Miceli (2001; 2010) suggest that learner use of corpora should not be dismissed. It has positive effects on learning and cultivates a more learner-centred approach to language learning. As described, learners are positive towards the use of corpora and become more confident in their language use. More importantly, they take charge of their own learning and become more aware of the language. These findings, therefore, enhance our understanding of the extent to which learners can make successful use of language corpora and what skills or knowledge they need to enable them to make optimal use of corpora. Indeed, the interesting findings by Kennedy and Miceli, especially in the study from 2001 showing the steps in a corpus investigation, not only reveal how effectively the students use the corpus, but also make the most substantial contribution to an understanding of how learners use a corpus that other studies in the section below seek to identify.

2.8.3 How students use a corpus

Another set of significant questions which is posed in this research area and needs to be answered through extensive research concerns the practical issue of how students use a corpus. The research questions raised in this group of studies are more or less the same as those concerning the effectiveness of learners' concordancing, but the significance of questioning how the students use a corpus is to gain a deeper understanding of what hinders successful corpus use and what to do to promote more effective use of language corpora. Among others, Yoon (2008) asks the most general question of how ESL learners use a corpus in academic writing in L2. Sun (2007) probes a similar issue, but her study is more comparative, aiming to investigate how learners with different levels of writing proficiency and different publication experience use a corpus (the SWT). A more specific question focusing on difficulties learners encounter while working with a

corpus is set by Yoon and Hirvela (2004). Although this question is not as straightforward as the former, to a great extent, it offers hints about how students use a corpus. All these questions lead to the core of understanding how students use a corpus. That is knowing how much help learners need in order to develop appropriate corpus-based learning skills (Turnbull and Burston, 1998) and how to "improve the apprenticeship to cultivate skills, knowledge, and attitudes" (Kennedy and Miceli, 2010: p. 35).

Findings

The discipline that students are studying is found to have an influence on the ways students use a corpus. The extent to which learners use a corpus depends on the requirement of writing in their courses. The results in Yoon (2008) reveal that the students favour corpus use for checking usages of words. Out of 6 students, three used the corpus more frequently than others. They used it for writing class assignments, writing for their own purposes, and writing in other courses. The other three students, on the other hand, used the corpus only for the writing class. The factor affecting the corpus use is the difference in the field of study (see findings 2.7.4 for more details). Those who frequently used the corpus are in the fields of history, education, and science education, which require the students to write more than the other science-based students. These students most frequently conducted searches on preposition usage. The second most frequent searches conducted were on word usages (the kind of complements the verb takes, voice, and verb collocation) and word contexts. Verb forms were found to be the most problematic word class as the students most frequently searched for them. To a great extent, the results are interesting, but, based on the question raised, they are fairly disappointing as the findings mainly report what the students used the corpus for instead of describing the process of how the students

actually used the corpus. It would have been more useful if the steps of using the corpus taken by the students had been described more precisely.

Sun (2007) provides more straightforward and innovative results, based on her research on designing a concordancing program (the SWT). She finds that students with different research-related background and different language proficiency exploit the SWT differently for their scholarly writing. The student with more research-related experience less frequently referred to the SWT to study the moves in the research papers and modified the specific moves in the information template to suit his own writing. The students with low research-related experience relied heavily on the SWT and tended to accept what they found in the information template without any modification. The student with lower language proficiency also referred to the language template more often than those with higher proficiency. They also used it for different purposes. The low language proficiency student used it for finding model sentences to fit his writing. In contrast, the students with a good command of English used the language template to check the correctness of their English, e.g. tense usages and collocation. It is not surprising to learn that students with less research experience rely more on the reference tool, but what makes this study more interesting is the descriptive analysis of how the three selected students with different background used the SWT to assist their scholarly writing. These descriptions – though based on a very small number of students – lead to a comparative summary of how students varying in terms of their publication experience and language proficiency use the corpus (see Sun, 2007: 335). Thus, the study contributes to an understanding of how language learners actually use a corpus and what help they need to use it more effectively.

Generally, teachers wishing to introduce learners to language corpora may think that it is difficult for students to use a corpus independently. The results of the study by Yoon and Hirvela (2004) who look into learners' difficulties using a corpus indicate that it is neither very easy nor very difficult for students to use a corpus. Generally, the intermediate students encounter more problems than the advanced students. The most serious problem raised by most of the students, especially the intermediate students, is the amount of time spent on analysing corpus data. The other issues that at least half of the students from the intermediate and intermediate classes find it difficult to deal with are truncated concordance lines, low Internet connection speed, and corpus output interpretation. In reading concordance lines, unfamiliar vocabulary items and text authenticity do not cause much difficulty. Remarkably, there are a few points that the advanced students have slightly more difficulty with – search techniques, concordance line analysis, and the large number of concordance lines. These findings are important in that they allow teachers to foresee what difficulties learners will have and what should be done to avoid or minimise the risks of classroom concordancing. However, the results are based on the rating-scale attitude questionnaire where the researchers' predicted statements of problems are mixed and where the names of the respondents are present for the purposive follow-up interviews. This may have compromised the data as the students might not want to express their genuine negative attitudes or feelings or might not want to take part in the follow-up interviews. These results, then, should be treated with caution. If the open-ended question on the students' concordancing difficulties had been added, more genuine and unanticipated problems from the students' own standpoint might have been provided.

Apart from the difficulties caused by the corpus itself and its nature, learners' preferences for their own learning in general can affect their corpus use. Turnbull and Burston (1998) report that learners with different learning styles and inductive learning skills achieve varying degree of success in concordancing activites. The main cause of the students' delayed development of concordancing skills is the limited training provided. With this minimal training, the student who gets used to inductive learning benefits more from a corpus whereas the student who is not familiar with an inductive approach experiences greater difficulties in performing concordancing tasks. This implies that, in introducing learners to use corpora for language learning, learners' differences in learning style and familiarity with inductive approaches to learning should be taken into account. Turnbull and Burston also find that the limited data and contexts available determine the types of corpus investigation and the number of successful investigations. The types of investigation and observation made are mostly limited to local searches based on the KWIC format. The searches for usages of "prepositions, adverbs, articles, and morphemes such as the –ing participle and –ed participle" are the most successful. Content words with low frequency, function words strongly associated with other words beyond the KWIC format, and some expressions showing the subtlety of semantic complexities are not easily searched for in this study. To help students develop appropriate concordancing skills, Turnbull and Burston propose that students should be provided with comprehensive and guided training. They remark that demonstrations on performing a variety of searches and guidance in observing the underlying language patterns would help all kinds of learners to benefit from independent use of corpora for language learning.

With a similar focus, Kennedy and Miceli (2010) explain that all three students used the Contemporary Written Italian Corpus in different ways and became confident users. In addition to finding word patterns and testing their hypothesis about language as expected, all of them developed a new function of concordancing, which was not previously anticipated, by using the corpus to find Italian equivalents and felt that it was rewarding to use the corpus independently. For the other student, the apprenticeship training in using the Italian corpus was not inspiring enough to provoke her to shift her attention from the grammar book and dictionary. It was hard for her to differentiate between the use of a corpus and other traditional reference tools and she considered the corpus unnecessary. In training students to use a corpus for apprenticeship learning, Kennedy and Miceli recommend putting more emphasis on the nature of corpus use.

The results of research that has sought to uncover the ways learners use a corpus are relatively consistent. They suggest that applying corpus use to language learning is more straightforward for advanced learners as they can cope with complex concordance investigation better than lower proficiency learners. The results in Sun (2007), Turnbull and Burston (1998), and Kennedy and Miceli (2010) are provocative because these studies provide more insightful information about how the more and less successful students use the corpus and how they go about it. This provides grounds for learner preparation for effective concordance tasks. The main critique of this work is that the results are based on a very small number of participants who vary a great deal from one another in terms of learning background, age, and language proficiency, etc. and may not reflect a larger group of corpus users in general.

2.8.4 Students' evaluation of and attitudes towards corpus use

Other than an interest in testing the feasibility of implementing concordancing in the language classroom and for error self-correction, one of the issues found to have received greater attention from researchers and teachers who have introduced learners to the use of corpora for serving their language needs is the students' reactions to or attitudes towards this experience. The attitude questions are of minor importance in many studies, e.g. Gaskell and Cobb, 2004; Chambers and O'Sullivan, 2004; O'Sullivan and Chambers, 2006; Lee and Swales, 2006. These studies evaluate the students' attitudes towards using a corpus in general. They simply asked if the students find concordancing useful or helpful. Gaskell and Cobb also extended the question to find out if learners would use a corpus independently subsequent to training.

In other studies by Yoon and Hirvela (2004), Phoocharoensil (2012), Sun (2007), and Charles (2012), the students' attitudes towards their use of corpora have become the main focus of the study. Some of these studies assess the students' reactions to using general corpora. For example, Phoocharoensil's only research question is about the attitudes of Thai learners of English towards grammar teaching through concordances. Yoon and Hirvela attempt to gain wider understanding of aspects of students' attitudes towards concordancing. They raise four questions, three of which are relevant to the students' attitudes. They began by eliciting the students' overall evaluations of corpus use for academic writing in L2 and narrowed down the question to focus on what ways concordancing benefits their academic writing. The rest of these studies aim to evaluate the students' perceptions of their experience of using a teacher-compiled corpus as a reference tool (Sun, 2007) and building their own corpus (Charles, 2012). Sun's

study also differs from others' in that it emphasises the differences in reactions between learners with different writing and publication background.

Research on students' attitudes towards concordance use for writing in L2 covers a wide variety of attitudes: students' attitudes in general, their attitudes towards a specially-teacher-designed corpus, and their attitudes towards creating their own corpus.

Findings

In general, the studies on learners' attitudes towards using corpora and towards the effectiveness of the teacher- and student-compiled corpora yield similar results, which are not contrary to expectation. The majority of the students are optimistic about corpus use. The most basic attitude survey of corpus-based grammar instruction through questionnaires and interviews by Phoocharoensil (2012) shows that the students are very positive towards corpus-based grammar learning. Most of them feel that corpus-based grammar instruction is better than other methods and they are proud of the outcome of their own learning. This study allows only a superficial understanding of learners' attitudes towards their experience with learning grammar points from the concordances. Its significance is that corpus-based grammar instruction leads to self-esteem, which is highly desirable for learning.

The previous study does not treat learners' attitudes in detail. The comprehensive studies by Chambers and O'Sullivan (2004) and O'Sullivan and Chambers (2006) suggest that both their postgraduate and undergraduate learners of French have positive attitudes towards using the corpus of French to improve their writing. Based on the study in 2006, 71.43% (10 students) of the undergraduate students find corpus consultation helpful or very helpful while 28.57% (4

students) of them who miss half and more of the training comment that corpus consultation is slightly helpful or unhelpful for them. With respect to future use of corpora, 42.86% (6 students) of the students report that they will use a corpus to improve their writing while 28.57% (4 students) prefer to consider using it. Those who miss quite a lot of the training insist that they will not use a corpus in the future. The students think that corpus consultation is useful for checking word contexts, sentence structure and idiomatic expressions, and identifying differences in meanings and usages of words.

In comparison, the postgraduate students' (Chambers and O'Sullivan, 2004) attitudes are slightly more positive than the undergraduates. On average, all of the students rate concordancing helpful and the majority rate it very helpful. These students also have more interest in corpus exploitation in the future. Yoon (2011) comments that these findings correspond to the assumption by Johns (1988); Turnbull and Burston (1998); and Granath (2009) that concordances have a more beneficial effect on advanced learners' language learning.

While O'Sullivan and Chambers (2006) find that the Master's degree students perceive the higher value of concordancing in writing, extreme results relating to motivation are found in Lee and Swales (2006). They report that the doctoral students are highly motivated to use a corpus to improve their writing and have very positive attitudes towards corpus use. These students also find their self-compiled corpus useful.

Among the studies that focus on students' evaluation of and attitudes towards using a corpus in general, the results in Gaskell and Cobb (2004) are pleasing but less promising. Although the survey results indicate that all twenty students learn a lot from the corpus and that the corpus

helps improve their writing, only eight of them express their intention of using a corpus in the future. The further analysis of their actual use of the corpus to correct the errors reveals that the rate of using concordances to correct the errors is a lot higher when they are given the precast links than when the links to the concordance lines are not given. This suggests that most of the students tended not to use the concordances for error correction independently unless the concordance lines were made ready for them. The results also show that the rate of successful correction reduces after the students make their own concordances. Most of the errors (80%–100%) are accurately corrected when the relevant concordance lines are provided, but about 60%–70% of the errors are accurately corrected when the students make concordance lines independently in the consecutive weeks. At the end of the experiment, the rate of successful error correction with the learner-made concordance lines fall dramatically to less than 50%. The researcher attributes this to the students' concern over the upcoming final exams.

The research by Yoon and Hirvela (2004) provides more new insights into the students' evaluation of their corpus use and the results are, to some extent, slightly surprising. Generally, the students are positive towards corpus use for writing in L2. Surprisingly, the intermediate students find it easy and helpful to use the corpus and are more positive than the advanced students, who encounter more technical problems. Despite the overall evaluation that the intermediate students are more satisfied with concordancing than the advanced students, both groups, especially the advanced students, demonstrate their interest in recommending corpus use to others, particularly to the fellow students in their home countries. When asked to compare dictionary use and corpus use, the students agree that a dictionary is useful for studying word meanings while a corpus is helpful for exploring lexical patterns. More importantly, they value

corpus use for improving their writing skills and increasing their confidence in writing. In terms of the benefit of corpus exploitation to writing, both the intermediate and advanced students agree that concordancing is the most useful for acquiring vocabulary and phrase usages and for studying word meanings respectively. Again, the intermediate students perceive the usefulness of concordancing for these two aspects of language more than the advanced students. This challenges O'Sullivan and Chambers' (2006) findings that advanced learners find corpus use more encouraging than intermediate students and they score better, and it challenges the assumption that advanced students benefit more from corpus consultation (see above). Yoon and Hirvela attribute this difference to the greater direct training the intermediate students receive while the advanced students explore the corpus more on their own.

While the above-mentioned studies have dealt with students' attitudes towards using corpora, Sun (2007) finds that the students had very positive attitudes towards the SWT. They thought that it is of great help particularly with sentence structures, idea development, organisation, and section prompts. Meanwhile, the usefulness for the other four areas — word choice, paraphrasing, grammatical patterns, and punctuations —was rated slightly lower. Even though they comment that the corpus is not large enough to provide sufficient examples, the SWT is useful for scholarly writing in their own disciplines and they will keep on using it. With respect to the students' writing status, it is found that the students who are currently writing research articles evaluate the usefulness of the SWT significantly higher than those who are not currently writing research articles. However, there is no significant difference in the evaluation among the students with and without publication experience.

In the same way, Charles (2012) reports that students have extremely positive attitudes towards their experience of building and using their do-it-yourself disciplinary corpus of research articles/theses. Over 90% of them find it easy to compile a corpus. Most of them are successful in building a corpus of 10–15 articles and enthusiastic to use their own corpus. About 90% of them say that the corpus helps them improve their writing and they have the intention of using their corpus in the future. However, Charles points out that the results should not be taken too uncritically because the students' positive responses might have been given to show politeness and respect to the teacher. She proposes that more accurate responses from the students could be elicited with the help of the third person.

As explained above, the research on learners' evaluation of and attitudes towards concordance use for error correction and language learning covers three main respects. The larger proportion focuses on learners' perceptions of their experience as a corpus user in general. The much smaller proportion focuses on learners' evaluation as a corpus builder and as a user of a teacher-compiled corpus. The results consistently show that the majority of students are positive about using and creating a corpus. These positive responses indicate it is feasible to encourage language learners to use a corpus for independent language learning.

2.8.5 Factors that mediate the corpus use

Efficient use of corpus technology for language discovery may largely depend on various factors. Basically, students need to have computer skills to operate the concordancing software. If they are to use an online corpus, they need an Internet system that is stable and fast enough to download the information. To some extent, students need to have some linguistic knowledge and researching skills required for forming language hypotheses and drawing conclusions about

language. They also need motivation and perseverance to observe the concordance output. All these factors can affect students' use and experience of concordancing. In response to this, a few studies have sought to uncover the factors affecting and influencing the degree of success in using a corpus independently as a reference tool for writing in L2. Yoon (2008) focuses on both learners' factors and external factors and poses the broader question of "What are individual experiences and contextual factors that mediate the influence of corpus technology on students' L2 writing" (p. 33). As opposed to this, Kennedy and Miceli (2010) place a strong emphasis on internal factors as they set a question specific to the learners rather than to external factors – "What factors, especially skills, knowledge and attitudes, affected their propensity and ability to use these reference-resource functions?" (p. 35).

Findings

Many factors have an influence on the students' use of corpora. Yoon (2008) finds that both individual experiences and contextual factors determine the frequency of corpus use, choice of language items searched, insight of interpretation, and degree of success in using a corpus. The individual experiences and external factors include experience of writing in L1 and L2, motivation for improving writing, nature of the field of study, resource needs, familiarity with concordancing, convenient time, language proficiency, and writing proficiency. According to Kennedy and Miceli (2010), learners' internal factors determine the ways they use a corpus. They discover that the students' ability to use a corpus and the extent of using it are influenced by their attitudes towards corpus exploitation, their understanding of corpus use, and their computer skills. Similarly, Lee and Swales (2006) demonstrate, with their doctoral students, that motivation plays a significant role in determining the use of corpora for inductive learning.

From these results, it can be inferred that internal factors are key to the degree of success in using a corpus. If students have motivation, good attitudes, adequate computer and language skills, they tend to put effort into investigating concordance lines.

2.9 Conclusion

According to this review, trends in research on corpus use by learners during the past twenty years have not changed much in terms of research focus, which concentrates on the effects of classroom concordancing, learners' attitudes towards designing and using a corpus, the role of corpus as model, and self-correction of errors. However, within this trend, one thing that becomes obvious is that the amount of work that employs teacher-compiled corpora and learner-compiled corpora has increased significantly, compared to the work that employs ready-made corpora, which is much smaller in the number of studies.

As shown in this chapter, I have divided previous work on learner corpora into five groups according to the research focus. These are 1) the feasibility of learner use of corpora, 2) the effectiveness of learner use of corpora, 3) how students use corpora, 4) students' evaluation of and attitudes towards corpus use, and 5) factors mediating corpus use. Generally, the findings from most of these studies are positive in that advanced learners tend to find corpus work easy. In cases where the findings are not entirely satisfactory, none of the researchers express a pessimistic view of corpus use for language learning. Instead, they offer ideas of how to help students to find corpus consultation more useful. These findings are significant in that they have attempted to explore learner use of corpora from different perspectives. Insights gained from these studies indicate the current potential of learner use of corpora. This, in turn, helps shape further applications of corpus resources in language pedagogy. The first group of findings on the

feasibility of learner concordancing shows that it is possible for learners to investigate corpus data and observe language usage from concordances. Students are able to induce valid rules about the language and apply the rules to self-correction of errors. However, the second set of findings on the effectiveness of learner use of corpora reveals that students do not make highly effective use of corpora. The most interesting result found in this group of studies is that no obvious effect of corpus consultation on language learning is found, compared to non-corpus-based language instruction. The third group of studies, to some extent, provides an understanding of how learners use a corpus and, to a greater extent, reports what difficulties learners face in investigating concordance lines. This information is particularly useful for learner preparation for corpus use. Despite the difficulties students encounter in the third group of studies, the studies in the fourth group show similar results that students enjoy using the corpus and are highly positive towards corpus use. The last group of findings convinces that the ways learners use corpora are greatly influenced by their attitudes towards corpus use and motivation to use it, as well as their individual experience and other external factors such as writing experience, time convenience, and nature of the course taken.

The results of these studies suggest that encouraging students to use a corpus is a viable and rewarding thing to do. However, it is not entirely understood what learners do when they use a corpus and how they can be prepared to use a corpus more effectively because most of these studies have focused on the output of learner use of corpora and on their attitudes towards their experience of concordancing. Studies aimed at examining the process of investigating concordances by learners are relatively rare. In addition, the studies under review have left several gaps. First, most of them have been conducted with advanced students at the

postgraduate level. Second, most of the students in each of the studies are enrolled in multi-disciplinary courses of study. These students vary a great deal in terms of age, learning experiences, etc. Third, most of these students were not language students. They were enrolled in language courses as part of the requirement for their course degree. Last, the students in these studies did not have full control over their choice of investigation. In most cases, they were asked to search a corpus for the language items specified or identified by the teachers.

From these findings, the research by Kennedy and Miceli (2001) is most significant to my study because it seems to be the only study to date that has attempted to provide thorough understanding of how learners use a corpus. What makes a sharp distinction between my work and Kennedy and Miceli's work is a wider focus of the study and the larger number of participants involved. Where Kennedy and Miceli provide a detailed description of what the students do that goes wrong at each step of investigating concordances based mostly on video recorded pair-work data and follow-up interviews and partly on the students' accounts of using the corpus, I try to make a more general conclusion about a larger group of students' use of a large corpus. Therefore, my study focuses more on what the students find that does not work, as well as what they do that works. Another aspect of their study that is particularly interesting is that they are able to study the outcome of the process or the product and can identify the process of investigating concordances from beginning to end. This kind of process study that goes into a great deal of detail can only be done with a small number of participants. What my study attempts is a balance between a process-oriented study and a broader product-oriented study by focusing on the kinds of searches the students make and what they are thinking and doing while dealing with these searches.

As mentioned, the main drawback of Kennedy and Miceli's work is that it involved a fairly small sample of students all of whom were learners of Italian using a small corpus of Italian, not learners of English. Most of the work on use of corpora by learners of English, on the other hand, has involved postgraduate students with upper-intermediate to advanced proficiency levels, who tend to find corpus work easy and useful. Still, less advanced students' use of corpora needs to be re-explored. Therefore, it is interesting to extend this question to a larger homogenous group of English-major students at a lower level to see whether the results are the same.

Chapter 3

Participants and data collection procedure

3.1 Introduction

The purpose of this study is to examine how learners use corpora for linguistic investigation and what language points they tend to look for from a corpus while writing essays in English and correcting errors. Therefore, the ultimate goal of the study is to find out to what extent Thai learners of English can make optimal use of corpus resources and how they can be prepared and encouraged to use a corpus more effectively. This goal gives rise to the following research questions.

1. What kind of lexicogrammatical errors do Thai learners of English find it easiest to solve by using a corpus?
2. When students are writing essays, what language points are they most likely to check in a corpus?
3. What do the students do when they perform a linguistic investigation using a corpus?

To achieve this goal, the types of language problems that students could easily solve using a corpus, the searches they carried out, the linguistic information they wanted to check, and the ways they interpreted the search results have been investigated. This chapter describes the methodology adopted in this study. It also discusses problems and solutions during the data collection process.

3.2 Population

The population of this study was third-year English major students at the Faculty of Liberal Arts, Prince of Songkla University, Thailand. There were 55 students, split into two sections, and they were selected because they were enrolled on 892-313 Academic English Writing, a compulsory course offered by the Department of Languages and Linguistics in the second semester of the academic year 2012 to cater for the third-year English major students. Therefore, it was convenient and practical to offer them the opportunity to complete class writing tasks in an environment that would directly benefit them and would also contribute to my study.

3.3 Participant recruitment and ethical approval

The students' participation in this study was voluntary, and ethical approval for the study was obtained from the University of Birmingham ethics committee. To recruit the participants, I made contact with authorised people involved at Prince of Songkla University. First, I asked for permission from the Dean of the Faculty of Liberal Arts to collect data and use the facilities and resources at the faculty. Then, I talked to the two teachers responsible for course 892-313 Academic English Writing about my research and data collection plan. After that, I approached the students by intervening in writing classes and talking to the students as a group about my project, the confidentiality of participants, the security of data, and their right to withdraw from the project. To ensure that the students were fully informed about the research project and to obtain valid consent, at the end of the talk, they were given an information letter (see Appendix 1) and a consent form (see Appendix 2) to consider. They were also given a questionnaire (see Appendix 3) and time to think about whether or not they would like to take part in the study. Those who wished to participate in the study signed the consent form, completed the

questionnaire, and returned both of them to me. The students also had a copy of the signed consent agreement given to them. Withdrawal from the study could be done at any time they wished during the data collection stage without consequence or penalty by talking to me in person or via email. The data of those who withdrew would be taken out and destroyed confidentially afterwards. Initially, 45 students volunteered to take part in the study. Later, 7 of them withdrew during the data collection preparation and at the commencement of data collection. Detailed information about the participants will be given in the next chapter.

Owing to the fact that the students from both sections had different free time and it was difficult to meet at the same time to arrange training and data collection, most of the arrangements for training in using the BNCweb and for data collection were made via a Facebook discussion. In cases where further discussions with individual students about issues arising from training and data collection were needed, but the students were not available to have face-to-face discussions, the discussions were also held on Facebook. To be precise, in this study, a Facebook discussion was used only as a means of communication with the students, not as a means of data collection.

3.4 Instruments

The instruments used for collecting the data during my research trip to Thailand comprised a questionnaire, an error correction test, and video recordings of the students' corpus use and think-aloud protocol. In order for the participants to generate data for the study by using a corpus to check and correct their problematic language use in writing, they were asked to produce texts or writing samples.

3.4.1 Questionnaire

A questionnaire was used in the initial stage for collecting personal information about the subjects (e.g. age and gender), internet and corpus use, and their experience of learning English (number of years of learning English, Grade Point Average, etc.). The questionnaire was not used to collect data for the research questions, but the information obtained was used mainly for describing the background of the participants of the study, and also for predicting the subjects' general English language ability and their ability to perform the tasks during the data collection stage.

3.4.2 Error correction test

An error correction test (see Appendix 4) was used to examine what lexicogrammatical errors are most successfully solved by using a corpus. The test consisted of 20 items with different types of linguistic features to be tested namely collocation, pattern, grammar, word order, and word choice. The sample sentences were taken from corpora which would not be used by the participants in this study. The error in each sentence was constructed to exemplify errors learners tend to make in each type of the linguistic features under investigation. There are two identical versions of the test (Version 1 and Version 2). The construction of the test will be explained in detail in the next chapter.

3.4.3 Students' writing samples

I use the term 'Students' writing samples' to refer to drafts of written work the participants produced for class; students were asked to correct some language errors using a corpus to check the words or phrases they had problems with while writing. The two writing samples provided by each student were composed of an academic paper and an essay. These writing samples were

not used for the main analysis but they gave information about which errors the students had identified and what they were likely to check in the corpus in order to correct those errors. Instead, the data obtained from the video recordings while the students were searching a corpus and interpreting the results (see 3.4.4) in order to correct the errors in their writing were analysed for the study.

3.4.4 Video recordings of the students' use of corpora and think-aloud protocols

While using a corpus to complete the error correction test and to search for the words or phrases they had problems with while completing a writing task and editing their drafts, the students were asked to think aloud. While doing so, they were asked to capture the computer screen and record their think-aloud protocol using the Camtasia Studio 6 software. Therefore, recordings of students' corpus use and think-aloud protocols comprise electronic files containing detailed information about the process each subject went through while using a corpus and interpreting the corpus data.

Think-aloud is defined as "a research method in which participants speak aloud any words in their mind as they complete a task" (Charters, 2003: 68). The think-aloud method has its roots in cognitive psychology and it has been used in SLA research during the past few decades as a means of understanding the leaners' cognitive processes of acquiring the language (Yoshida, 2008). To obtain data for research, the participants are asked to tell the researchers their thoughts while completing learning tasks by thinking aloud or talking to themselves. In the traditional method, the think-aloud protocols are tape- or video-recorded and transcribed for analysis. Think-aloud protocols can be classified as retrospective or concurrent (Yoshida, 2008). Retrospective think-aloud protocols are provided by the participants in a delayed manner after

they have completed the task by recalling what they were thinking while doing the task. Concurrent think-aloud protocols are provided by the participants in real time while they are engaging themselves in the given task.

The two categories of think-aloud protocols mentioned above suggest that concurrent think-aloud protocols provides more accurate data because it reflects what the participants are actually thinking and doing in real time. Yoshida (2008) regards this as the major benefit of think-aloud protocols. In contrast, the data obtained through introspective think-aloud protocols may be distorted in some way as the participants may forget what they were thinking while doing the task. Olson et al (1984) state that think-aloud protocols can be used as one of the most effective ways of eliciting a higher level of thinking processes.

In exploiting think-aloud protocols to gather data for research, think-aloud method can have practical disadvantages. One of the disadvantages pointed out by Charters (2003) is that this method is expensive and time consuming, so it can be used only with a small number of participants. The other disadvantage of this method is that it is difficult to analyse transcribed data. In reading the transcriptions of learners' doing something, it is possible that the researchers reports only the processes or strategies they attend to and ignore the processes they unconsciously do not attend to, so the report is poorly incomplete (Yoshida, 2008). For this reason, Charters (2003) states that using think-aloud protocols to gain data needs to be done with care.

Despite the fact that the traditional think-aloud method is time consuming and that it can practically involve a relatively small number of participants, this method is highly praised for its insight into the participants' cognitive processes. This study modified the typical methodology in

order to have a larger group of participants think aloud simultaneously at the data collection stage by using headphones and screen-capturing software to record what the participants were thinking along with the screen capture of what they were doing as part of their DDL process at that point in time. This innovative alternative to the traditional approach to think-aloud protocol analysis led to the data being obtained from a large group of participant in a short time. The data from the screen capture could be used to supplement the think-aloud data, making the understanding of learner concordancing more precise. Chapter 6 explains how these data were analysed.

3.5 Procedure and problems

The data collection procedure for this study falls into two stages: training students to use the BNCweb and the Camtasia Studio 6 software, and data collection.

3.5.1 Stage 1: Training students to use a corpus and the Camtasia Studio 6 software

The corpus used in this study is the British National Corpus (BNC), which is accessed by the BNCweb, the concordance software. Before collecting the data, the participants received training in how to use the BNCweb to search for a target word and how to observe how the target word is used or occurs in the concordance lines. In order not to disrupt the students' class time, training was conducted outside regular classes. At first, I planned to give the same training to all the participants promptly at the same time each week. However, because the students took different courses, they could not attend the training at the same time. To solve this problem, I repeated each of the training sessions twice either on the following day or on the same day, the first for the majority of the participants and the second for the rest who missed the previous training session. In total, the participants were provided with three BNCweb-training sessions in three consecutive weeks. The first training session lasted for two hours and the second and the third lasted one

hour and a half each, so in total, the students received five-hour training. Prior to the training, the students reported that they had no concordancing skills as they had never used a corpus before, but they had fairly good computer skills. During the first two hours of training, the students were introduced to the concept of concordancing and the BNCweb. They learned how to access and navigate the BNCweb, how to make queries and sort concordance lines. In the second session, the students practiced conducting simple searches and interpreting the results. In the last session, after getting familiar with using the BNCweb, the students received training in conducting advanced searches using grammatical category labels (e.g. NP0, AV0, and AJ0), wildcards (e.g. ?, *, and +), and metacharacters (e.g. /, ( ), and { }) and interpreting concordance lines. At the end of the training, the students were expected to be able to use the BNCweb to make their own queries and could interpret concordance lines correctly. However, before the experiment, the students were not tested to measure how much they had achieved from this corpus training. The handout explaining how to use the BNCweb (see Appendix 5) was partially adapted from Dr. Neil Millar's handout used in the course Research Methods in Corpus Linguistics offered to the MPhil Corpus Linguistics and MA Applied Corpus Linguistics students at the University of Birmingham in the autumn term 2011.

After the corpus training, the participants received training in how to use the Camtasia Studio 6 software to record the process they would go through while searching a corpus and working with the corpus output at the data collection stage. Due to the problem noted above, this training, lasting about an hour, was repeated three times, the first two on the same day and the last on the day when the data collection began.

3.5.2 Stage 2: Data collection

Originally, at the data collection stage, the participants were asked to undertake two tasks in the computer room: an error correction test and a writing task. This stage also took place outside ordinary class time.

3.5.2.1 Task 1: Error correction test

The purpose of the error correction test was to examine what lexicogrammatical errors are most successfully solved using a corpus (RQ 1). In doing task 1, the participants were asked to correct the error underlined in each item in part 1 (items 1-15) and choose the most suitable word to complete each sentence in part 2 (items 16-20) using a corpus. To distinguish the items that the participants could correct using their existing knowledge from those they could not correct themselves and needed to use a corpus to help, students were asked to complete the test twice. First, I gave them Version 1 of the test and asked them to look for the items they could correct by themselves and to correct them. After the participants finished correcting the items they could correct by themselves in Version 1, I collected the test paper and gave them Version 2 to complete.

In Version 2, which was identical to Version 1, the participants were asked to use a corpus to help correct the errors they could not correct the first time in Version 1. At this stage, they could also use the corpus to check the answers for the items they had corrected previously in Version 1 if they wished. In this Version, the participants could spend as much time as they wanted on the task.

3.5.2.2 Task 2: Editing task

My plan for this task was that the participants would be required to produce a piece of writing for an Academic English Writing class and spend an hour doing so. The topic or content of the writing was supposed to be assigned by the course instructors. Within an hour, it was not necessary that they had to finish a complete draft, but their writing should be no less than 300 words in length. After that, in their drafts, I would ask them to highlight as many things as they wanted that they were uncertain about and needed to use a corpus to check. Once the participants finished their first drafts, they would have to edit the drafts by searching a corpus to check if the word or phrase or pattern they underlined was used correctly. If not, they needed to correct it. Any errors changed or corrected based on the corpus data would be underlined, and I would collect these corrected drafts for analysis.

However, as the data collection for task 2 was intended to be done in a natural classroom context, I could not conduct the data collection process as planned and there was a slight change to it. In the Academic English Writing course from which the data were being collected, the participants were assigned two pieces of academic writing: an essay and a research paper. The problem was that by the time the data for task 2 was about to be collected, the participants had already submitted the essay to their teachers. What they were doing at that time was writing a first draft of the research paper based on their essays. Before submitting the first drafts to the teachers, they were asked to do a peer review focusing on every aspect of writing a good research paper learned in class including language accuracy. I decided to use this writing as the main source of data for task 2 in my study, so I talked to the teachers about my plan and asked them to tell the

participants to do a peer review on other aspects of writing a good research paper other than the language which would be left for themselves to check from a corpus.

After the students had carried out a peer review of their writing, I asked the participants to bring their first drafts to the computer room to self-correct the language errors. Before using a corpus to check language use, I asked them to identify and underline at least 10 errors in their own drafts. After that, they were asked to access a corpus to check and correct the errors underlined. While the participants were working on this task, I circulated and observed that some of them identified less than 10 errors. I encouraged them to look for more errors in the drafts, but they said that they did not recognise the errors and could not identify more errors in their writing. After they finished the task, I collected the paper.

To deal with the problem that the participants could not identify their own errors and had not identified enough language problems for my study, I decided to go through the drafts myself and look for the errors that I thought it would be useful for the participants to look up in a corpus with the advantage that I would have rather more control over the searches the participants would conduct. However, I was not confident that the students would be happy and willing to come and do the same editing task again because, according to my data collection plan, they had completed all the necessary tasks. To gauge their interest in doing this, I asked if they were willing for me to locate the errors for them to look up in a corpus. I received a better response than expected because they needed someone to read their work and point out the errors they had made so that they could be able to correct their own errors and achieve better results for their course work. I read through each piece of writing, looked for the errors that I thought would work with a corpus and underlined them. While reading and identifying errors in some of the participants' drafts, it

seemed to me that what I was reading was not an accurate reflection of their own language use. I suspected that they had cut and pasted text from the sources they cited and as a result I could find only a few grammatical errors for them to look for in a corpus. However, I was still able to have the students search a corpus for the errors I had chosen because I could find plenty of errors in other pieces of writing. In the meantime, I looked for an opportunity for the students to perform an additional writing task that could reflect more accurately their genuine writing ability.

To overcome this problem, I tried to find a way for the students to use corpus resources to enable them to write properly, using their own words. However, as I had no clear evidence whether or not the students had copied and pasted others' work into their own or whether they had fairly high language proficiency and could produce written work of high quality, it seemed unreasonable to ask them to rewrite. On the other hand, if I could do so, this would probably have required a lot of input on my part and it would have taken a lot of time. In addition, it would probably have affected the teachers' plan. I, therefore, talked to the teachers who ran the course about the problem and this gave rise to task 3 for my data collection.

3.5.2.3 Task 3: Writing task

In addition to writing a research paper, during the last week of the course, the teachers planned to have the students write in class about what they had learned from the course. The topic for section 1 was *"What I have learned from the course Academic Writing in English"* and that of section 2 was *"What I have gained from the course Academic Writing in English".* I decided to include this writing task in my study. To complete the task in an environment that would contribute to my study, the teachers agreed to have the students undertake the writing task in the computer room under my control. The students from each section were allowed to write and edit

their writing during their class time for 90 minutes. While writing, the students who participated in my study were asked to use a corpus to check and correct the language problems they encountered.

While searching the corpus for correcting the errors in the test in task 1 and in their own writing in tasks 2 and 3, the participants were asked to think aloud, capture the computer screen in progress, and record their think-aloud protocol using the Camtasia Studio 6 software. This method of data collection was used to allow a deeper insight into what the participants actually did during performing the given task as the software could record everything the students did while using the corpus and interpreting the results. To ensure the anonymity of the participants and the confidentiality of data during the conduct of the research, each participant was identified with and referred to by a code number. The participants, at the end of each task, were asked to save the files using their code number in the file name. Also, they wrote their code number on their paper work. Therefore, both the paper work and electronic files had their code number and they could be easily matched up. To ensure security of the data, the completed spreadsheet regarding the personal details of the participants and their written work were stored in a locked drawer in my office where only I had access to the data. The recorded files of each participant were stored electronically in an external hard disc which was kept in a locked drawer and in my personal laptop to which no one else could have password-protected access, After the end of the project, the data were disposed of by shredding the paper and deleting the electronic files.

3.6 Other problems arising during the data collection stage

As seen above, the data collection procedure did not go as well as I had hoped. Apart from those problems mentioned, arranging the time for data collection and maintaining the number of

research participants was challenging. Due to the fact that the data collection for this study was done outside ordinary class time and the students were tied up in evening classes and other activities, it was difficult to arrange for all the participants to come for training in using the BNCweb and completing each task together at the same time. To make it convenient for the students to come for each training session, I had to provide them with options and I had to repeat the same training with different group of students. This, to some extent, slowed down the process of data collection, but it also helped maintain the number of participants taking part in the study.

As the students who volunteered to participate in this study could withdraw at any time during the data collection, maintaining the number of participants in order to have a large number of subjects was very important. Out of 55, 45 students volunteered to take part in the study. However, during training, 4 students informed me of their withdrawal from the study, and another 3 did not attend on the day the data for task 1 were collected. Therefore, eventually, only 38 students took part in my study. I considered finding more participants before proceeding with the data collection for task 2. However, I consulted a statistician and found that it might be unsound to recruit more participants as those 38 students had already been trained and had already completed task 1. More importantly, the rest of the student population were not willing to volunteer and 38 participants out of 55 is statistically acceptable. To encourage the participants to take part in the following tasks, I asked them to come at a time convenient to them to do task 2 and kept reminding them of the writing task 3 in the last week. As a result, all 38 students completed all tasks.

In sum, although there was a slight change in the data collection stage, it gave satisfactory results. Instead of undertaking two main tasks, the participants, as a result of the data collection

procedure adopted, undertook three tasks: one error correction test, one editing task, and one writing task. I, therefore, ended up proceeding with better data collection than I had planned because I obtained data from three sources for my study. The analysis of these data will be explained in the following relevant chapters where I discuss the results of my research.

Chapter 4

Types of Lexical and Grammatical Errors Most Successfully Solved Using a Corpus

4.1 Introduction

This chapter gives an account of experiment 1 conducted in Thailand in November 2012–February 2013. The aim of the experiment was to have students use a corpus to correct a number of incorrect items in order to find the answers to research question 1: What kinds of lexicogrammatical errors do Thai learners of English find it easiest to solve by using a corpus? The sections that follow detail the material analyzed to obtain the data, the methods of data collection, and data analysis and results. A discussion of the results is also given.

4.2 Subjects

As mentioned in the methodology chapter, initially 35 third-year English major students, enrolled on the course 892-313 Academic English Writing at the Faculty of Liberal Arts, Prince of Songkla University, Thailand, in the academic year 2012, volunteered to participate in the study. However, at the beginning of the data collection stage, seven of them decided to withdraw from the study, therefore in total, 38 students took part. Full details about the participants are given in the section below.

4.3 General information about the participants

The students who initially volunteered to take part in the study were given the questionnaire to complete regarding their personal information, English learning experience, and Internet use. At

the data collection stage, 38 of these students were participating in the study. The information about these participants obtained from the questionnaire is outlined in Table 4.1 below.

Table 4.1: Summary of the participants' information

Questions	Answers	Number of students
1 Gender	• Male • Female	6 32
2 Age	20–22 years old	20 years old = 13 21 years old = 22 22 years old = 3
3 Years of age when starting to learn English	2–9 years	2 years = 1 3 years = 2 4 years = 9 5 years = 8 6 years = 3 7 years = 10 8 years = 2 9 years = 3
4 Years of learning English	12–19, 15.53 in average	12 years = 4 13 years = 1 14 years = 1 15 years = 16 16 years = 2 17 years = 10 18 years = 3 19 years = 1
5 Number of students who had been to English-speaking countries for up to 3 months or longer	7	The USA = 6 New Zealand = 1

Questions	Answers	Number of students
6 Pre-requisite courses taken and grades obtained	890-210 English Grammar in Use	A = 12 B+ = 8 B = 4 C+ = 4 C = 3 D+ = 6 D = 1
	892-310 Paragraph Writing in English	A = 9 B+ = 6 B = 12 C+ = 1 C = 5 D+ = 4 D = 1
7 Grade Point Average Cumulative	3.36 in average: Highest 3.81 Lowest 2.6	
8 Views on writing correct English	• Very difficult and worrying • Difficult but not worrying • Not difficult	16 20 2
9 Comparison of their English level to the class	• Below average of the class • Average to the class • Upper average of the class	9 28 1
10 Frequency of using the Internet	• Everyday • Four times a week	36 2
11 Internet use (more than one answer applies)	• Email • Chat • Playing games • Entertainment • Searching information • Social networking • Online dictionary • Checking English	22 28 12 33 34 34 27 17
12 Number of students who had used a corpus before	2	
13 Frequency of checking the use of uncertain word or phrase when writing in English	• Every time • Often • Sometimes	8 23 7

Table 4.1 indicates that these students were reasonably homogeneous in terms of gender, age, and their experience of learning English at university. Given that they had been learning English for over ten years and that they were from the same major and had gone through the same learning process at university, it was assumed that all of them possessed a reasonable amount of English skills to perform in the study at their optimal level. Their Grade Point Average Cumulative, which is relatively high, also affirmed my belief that these students had sufficient knowledge of English to take part in the study. It is important that the questionnaire data revealed that these students used a computer quite often in their daily lives. Thus their computer skills were good enough to be trained to use a corpus and there were no problems anticipated in doing the experiment.

4.4 Material

The material used in the experiment to find out what kind of lexicogrammatical errors are most successfully solved using a corpus was an error correction test (see Appendix 4). The test consists of 20 items with errors based on different types of language features namely collocation, pattern, grammar, word order, and word choice. These items were selected based on assumptions about the types of errors that EFL students tend to make when writing in English and about the types of errors that a corpus could be useful for correcting. The basis of my assumptions about error types will be explained later in this chapter. The correct or actual sample sentences in the test were taken from corpora which were not intended for use by the participants in this study. Where appropriate, something in each of these sentences was changed in order to construct an error that exemplifies a particular type of error learners tend to make in each type of linguistic features under investigation. The design of the test used in the study is described below.

4.5 Designing and piloting the test

As mentioned earlier, the purpose of the test is to find out what language problems are easily solved using a corpus. Therefore, the aim of designing the test was to include a range of types of language features to be tested with the help of a corpus. The process of designing the test fell into four steps.

Step 1: Preliminary considerations

Before the actual design of the test, some practical considerations were taken into account. My preliminary consideration was that I wanted a range of language features and a principled set of error types. My decision was based on choosing language features in writing that I thought the learners would be able to find using a corpus. Then I considered how written language features are described. My assumption was that there is grammar, lexis, and discourse. I considered each of these in turn, and for grammar, I used Willis' (2003) classification: grammar of structure, grammar of orientation, and grammar of class. The grammar of structure is "the way items – words and phrases – are sequenced to make up larger units" (Willis, 2003: 29). For example, the structure of an English clause is **subject + verb + object**, and the basic noun phrase structure is **(determiner) + (adjective (s)) + noun**. The grammar of orientation refers to the systems of tenses and determiners to "show how the things we are speaking or writing about are related to the real world and to other elements in the text" (Willis, 2003: 34). In the clause *"my wife works in the garden most weekends"*, Willis (2003: 34) demonstrates that we can identify who the message is about, and whether it refers to the time in the past, present, or future, but we cannot find an 'orientation' in the clause *"wife – work – garden – weekend"* as there is no article or determiner to show whose wife it is, and the verb does not express the time reference because it is

not conjugated. The grammar of class refers to "words which relate to the same pattern as belonging to the same group or class" (Willis, 2003: 41). For example, there is a class of double object verbs such as *give* and *bring*. There is also a class of evaluative adjectives such as *good* and *interesting* belonging to the pattern *It + Be + adjective + to + verb*.

In total, five types of written linguistic features were classified: grammar of structure, grammar of orientation, grammar of class, lexis, and discourse. The types of linguistic features to be tested with a corpus were then identified based on these classifications. By definition and with the help of a corpus which presents examples of language use in the form of concordance lines which are sometimes in incomplete sentences, and from my perspective and my initial attempt to use a corpus to find information about language, I realized that there are limitations to what learners might be expected to use a corpus for. It is practically unsound to test discourse through corpus analysis because the discourse features, such as syntactic complexity and conjunction use or generic structure, are the features that would be difficult for students to find as they are context-dependent. Identifying an error in each of these three cases would be dependent on a large amount of co-text. For example, the selection of a correct conjunction may depend on reading a very long structured text to determine whether the conjunction is appropriate, and it certainly cannot be accomplished from looking at only one sentence. Likewise, in dealing with generic structure, the students would need to look at a whole text. Basically, in this study, I wanted to give the students a short line of text and did not want them to have to look at a large amount of co-text in a corpus. Therefore, I discarded syntactic complexity, conjunction, and generic structure and restricted my study to items of grammar and lexis. More specifically, I wanted to have errors that are identifiable from one sentence or partial sentence only, so I needed to have

more specific information and I chose only sentence-level errors for the students to work on. As a result, only grammar of structure, grammar of class, and lexis were selected. The grammar of orientation was excluded from this experiment because it was assumed that it would be more difficult to identify the use of tenses or determiners from concordance lines. Appropriate use of tenses and determiners depends on the intended meaning. It is not always possible/easy to say that the tense used is wrong unless there is a clear signal word or the verb is in a wrong conjugation form.

After deriving the types of linguistic features to be tested, the next point to be considered is the number of items to include in the test and the amount of time to be allotted. In order not to exhaust the students, it was decided to have 20 items in the test with the expectation that the students would spend no longer than two hours to undertake it.

Step 2: Corpus analysis and concordance selection

After the decision on the types of linguistic features, the number of test items, and the amount of time to be allotted had been made, it was necessary to find sample sentences to represent the kinds of linguistic features to be tested. To prevent the student participants who were going to use the BNC in the study from seeing the sample sentences when they looked in the corpus, I needed to find the sample sentences from a different corpus. Thus the Corpus Concordance English (v.6.5), available online at http://www.lextutor.ca/concordancers/concord_e.html, was used to retrieve the sample sentence for each item, most of the sample sentences were taken from Brown corpus, and some were from the 2k Graded Corpus (920,000) and the 1k Graded Corpus (530,000) which is a subset of the 2k graded corpus. The selected concordance lines were stored in a word file to be used in the test designing stage.

Step 3: Test design

Once the concordance lines were gathered, the test and its format was designed. In this stage, the errors were made up and the examples were put together in a mixed order. All the errors were in the same format — they were underlined. To make sure that there were sufficient concordance lines useful for each of the target errors in the BNC, I conducted searches for each error in the BNC. Since the participants were required to do the test twice without using a corpus the first time and with the help of a corpus the second time, two blank spaces marked 'First correction' and 'Second correction' were given below the incorrect sentence.

Step 4: Piloting the test

The purposes of the pilot test were to find out 1) which test items work, 2) which items should be dropped because they are too easy or difficult, 3) how much time the subjects spend on the task, and to foresee the problems that might occur during the data collection stage. The test was piloted four times with Thai students in the UK.

The first pilot study revealed that none of the items were too easy and needed to be taken out as the students did not immediately correct them all without using the corpus. However, some given errors seemed difficult to be solved using a corpus and needed to be changed or replaced for the following reasons. Sometimes, the pilot students could not identify the errors underlined and made more errors in trying to correct them. To help them spot the errors more accurately, new ways to present the errors such as giving choices, spaces, or clues, needed to be given. Some items were difficult to understand because of the complicated meanings, unfamiliar words, and very long sentences which sound perplexing. These items needed to be replaced and new items and the pilot studies dealt with this test construction.

The language feature exemplified by each item in the final version of the test is shown in the table below.

Table 4.2: Kinds of language features exemplified in an error correction test

Item	Language features	Type of grammar	Example from a corpus
1	noun pattern	grammar of class (increase + in + noun)	therefore we anticipate an **increase** in the number of children
2	adjective pattern	grammar of class (difficult + to + verb)	He finds it **difficult** to describe his feelings
3	noun pattern	grammar of class (relationship + with + somebody)	His **relationship** with the kids is one between equals,
4	question tag	grammar of structure (positive statement +, negative tag?)	We had such fun, **didn't we**?
5	verb pattern	grammar of class (suggest + that-clause)	Unlike the other men, Peter did not **suggest** they meet again.
6	noun class	grammar of class (uncountable noun)	On either side **furniture** was piled in high, precarious heaps.
7	adjective pattern	grammar of class (wise + to + verb)	it would be **wise** to use it at any time flying above 10,000 feet
8	verb phrase	grammar of structure (ought to + verb)	This **ought not to** be the case.
9	clause structure: subordinate clause	grammar of structure (reduced adverbial phrase: while + V.ing	**While** deciding to stay as independent as possible, I contacted ACET who I knew provided practical care at home.
10	relative clause	grammar of structure (relative pronoun as object)	There is a Banker's Order form attached to this leaflet **which** you can use.
11	indirect question	grammar of structure (wonder + question word)	I thought, I **wonder** what it's like?
12	verb pattern	grammar of class (be + allowed + to)	One day in August 1973, without warning, visitors were not **allowed** to enter the prison.
13	noun phrase/word order	grammar of structure (adjective order)	His school was a **big red-brick Victorian** building to the east of Kilburn.

Item	Language features	Type of grammar	Example from a corpus
14	if clause	grammar of structure (if + past perfect, …would + have + past participle)	'**If it had been** a different time you'd have been a doctor or an engineer,' Rose said.
15	noun clause	grammar of structure	**What she means is that** this instruction should be borne in mind if at any time it starts raining.
16	word class	grammar of class (adjective + noun)	The Tornado would launch a Harm or Alarm at a **safe distance**.
17	word form	grammar of structure (who/ whom)	The numbers attending were usually small but on one occasion the minister (Mr Dwyer himself, it is implied,) found that 300 people, **some of whom he** had never seen before, had gathered to hear him.
18	verb pattern	grammar of class (get + object + V-ed)	He tried to **get it** marketed or patented but he never succeeded.
19	collocation	lexis (verb + mistake)	She would not **make that mistake** again.
20	collocation	lexis (adjective + loss)	This time many were braced for **heavy losses** again.

4.6 Methods

After receiving training in how to use the BNCweb software, the participants were given the error correction test to do in the computer room where they had access to the BNCweb. The intention was to ask them to use the BNC to gain language information helpful in correcting the items to find out which of those selected language problems are easily solved using a corpus. However, it was probable that some participants would have enough knowledge to correct some of the items without recourse to a corpus. Thus this task was broken down into two stages. To eliminate the items that each individual participant could answer without having to use a corpus, in the first stage of the study, the participants were given the test, labelled Version 1, and were asked to go through the 20 items and to attempt to correct as many errors as possible without looking at the

corpus. As expected (see Table 4.3 below), the participants differed widely in their ability to correct the items. None of them got all the answers right at this stage. Out of 20, the highest number of right answers was 15 and the least was 1. Also, some of the items were answered correctly more often than others. For example, items 17, 19, and 18 were answered correctly 31, 27, and 26 times respectively whereas items 1 and 7 were answered correctly 5 times and item 5 was answered correctly only once. The advantage of this pre-corpus study was that it gave a base-line for establishing the effect of corpus use on the students' ability to correct the items.

Table 4.3: Number of items for which the participants got the answers right the first time without a corpus

Item / Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	**Total**
01						✓				✓			✓			✓	✓	✓	✓	✓	8
02			✓										✓			✓	✓		✓		5
03		✓		✓		✓		✓	✓									✓			6
04			✓	✓			✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓			13
05			✓	✓	✓								✓			✓	✓		✓	✓	9
06						✓			✓	✓	✓				✓		✓	✓	✓	✓	9
08			✓														✓			✓	3
09									✓				✓			✓		✓	✓		5
10		✓						✓	✓	✓	✓	✓				✓	✓	✓	✓		10
14				✓									✓			✓	✓		✓		5
15			✓										✓			✓		✓			4
16	✓	✓	✓					✓					✓	✓		✓	✓	✓	✓		10
17				✓					✓		✓	✓				✓	✓	✓			7
19	✓	✓	✓					✓					✓	✓		✓	✓	✓	✓		11
20											✓		✓			✓	✓	✓	✓		6
21		✓	✓			✓	✓	✓	✓		✓	✓	✓		✓		✓	✓	✓		13
22				✓		✓											✓	✓		✓	5
23		✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		15
24				✓				✓					✓								7
26	✓	✓				✓				✓	✓	✓					✓	✓	✓		9
27		✓	✓	✓				✓					✓			✓	✓	✓	✓		9
28				✓		✓			✓				✓			✓		✓	✓		7

Item / Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	**Total**
29											✓					✓	✓		✓		4
30				✓					✓				✓				✓		✓		5
31																				✓	1
32			✓	✓		✓			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		13
33		✓		✓		✓	✓	✓	✓	✓	✓	✓			✓		✓	✓	✓		13
34			✓	✓			✓	✓			✓	✓			✓	✓	✓	✓	✓		11
35		✓		✓	✓	✓		✓	✓		✓	✓		✓	✓		✓	✓	✓		13
36	✓			✓																	2
37				✓		✓			✓							✓	✓	✓			6
38									✓								✓	✓	✓	✓	5
39				✓		✓							✓		✓	✓	✓		✓		7
40				✓																	1
41				✓					✓	✓		✓				✓	✓	✓	✓	✓	9
42			✓	✓					✓				✓			✓	✓	✓	✓	✓	9
43				✓					✓							✓	✓				4
45	✓	✓	✓	✓		✓		✓	✓		✓			✓	✓		✓		✓		12
Total	5	11	13	22	1	15	5	13	19	9	14	11	18	16	10	24	31	26	27	10	

In the second stage of the study, after the test Version 1, done in the first stage, was collected, the participants were given the same test, labelled Version 2. Both tests (Versions 1 and 2) were exactly the same and were conducted on the same day. The only difference between Version 1 and Version 2 was that the students did not have access to a corpus in Version 1. In conducting the test Version 2, the students were asked to go through the same 20 items again and use the on-line corpus provided to help correct the errors. The participants were encouraged to try to find answers to questions they had been unable to answer the first time, and to check their answers where they had attempted an answer the first time. At this stage, they could spend as much time as they wanted on the test, but they did not receive any feedback on which of their answers from both tests were correct. While accessing the BNCweb to search for the concordances for the target word to correct the errors in the second stage, the participants kept complaining about the slow link-up speed probably caused by a large number of users accessing the same website from

the same server at the same time. Consequently, some of the participants started talking to others and some asked their friends about the answers while waiting for the searches they made to be downloaded. This technical problem was beyond my ability to resolve and the participants ended up spending longer on this task than expected. However, it seemed that the participants could manage to record their interactions with the corpus quite well. Only a few of them made a mistake while saving the recordings into the computer, which will be discussed in the next section.

While doing a search and working with the corpus data in the second stage, the participants were asked to think aloud and record their on-screen activities and think aloud protocol using the Camtasia Studio 6, a screen-capture video tool. At this stage, they could also use a corpus to check the answers for the items they had corrected previously in Version 1 if they wanted. The participants could spend as much time as they wanted on the task. Once they finished the test, they saved the video files named after their code number and task number onto the computer they used. The original video files which were relatively large were, later, converted into more portable windows media video (wmv) files and were collected.

4.7 Data analysis and results

In order to find out what kinds of lexicogrammatical errors are most successfully solved using a corpus (RQ 1), the data from the test papers were analysed to see how many participants got the answer right for each item in each version of the test. The data from Version 2 are used as the main source to answer this research question. The types of lexicogrammatical errors that are most successfully solved using a corpus are those which the participants could not have corrected without using a corpus and to which they got the answers right after using a corpus. In other

words, these can be referred to as those for which the corpus made the most difference. To find out how many of the participants could answer each item in each version of the test, the correct answers for each item were counted and calculated into a percentage based on the number of respondents to each item in each version**.** In Version 1, where the participants were asked to correct the errors without using a corpus, it was assumed that all the participants had responded to all the errors even though they did not provide the answer to every item. Leaving an answer blank indicated that the participants could not correct that item. For this reason, the total number of respondents to each item in Version 1 was 38 (N=38), even though not all students gave an answer to each item.

In Version 2, on the other hand, the total number of respondents to each item varied because not all the participants responded to the same items, depending partly on their ability to correct the errors in Version 1. Therefore, in analyzing this version, the first step was to ascertain the number of respondents to each item based on the number of written answers supplied. However, it could not be assumed that they did not respond to or attempt to correct the items where the answers were not given. This was probably because they had provided the answers to the same items in Version 1 and did not want to check the answers again, or if not, they had used the corpus in an attempt to answer correctly, but they failed and left those items blank. The next step to make sure whether or not the participants who did not supply the answers to some items had attempted to use a corpus to correct those items was to check the video recordings and think aloud protocol of those particular participants. If it was found from the recordings that the participants had tried to correct those items, but were not able to figure out the answers and skipped them, those particular items were marked as done by the participants. Conversely, if it

was found out that any of the correct answers this group of participants provided were not based on the corpus data, but solely on their existing knowledge or from asking or talking to a friend while attempting to correct the errors or waiting for the results of the search, these items were not taken into account in the analysis. Then, the total number of respondents to each item was added up and the data were analyzed the same way as in Version 1.

The following is an initial analysis of the results which is based on the paper results only, without investigating the video recordings. It includes all 38 participants.

Table 4.4: The percentage of subjects who got the answers right for each item the first time by themselves and the second time using a corpus (based on the test papers)

Item	Language features	1st correction (Version 1)			2nd correction (Version 2)		
		N	Number of subjects getting the correct answers	%	N	Number of subjects getting the correct answers	%
1	noun pattern	38	5	13.16	37	22	59.46
2	adjective pattern	38	11	28.95	37	29	78.38
3	noun pattern	38	13	34.21	37	27	72.97
4	question tag	38	22	57.89	36	29	80.56
5	verb pattern	38	1	2.63	37	6	16.22
6	noun class	38	15	39.47	37	30	81.08
7	adjective pattern	38	5	13.16	38	29	76.32
8	verb phrase	38	13	34.21	37	24	64.86
9	clause structure: subordinate clause	38	19	50.00	38	34	89.47
10	relative clause	38	9	23.68	33	26	78.79
11	indirect question	38	14	36.84	35	27	77.14
12	verb pattern	38	11	28.95	36	23	63.89
13	Noun phrase/ word order	38	18	47.37	37	34	91.89
14	if clause	38	6	15.79	37	27	72.97
15	noun clause	38	10	26.32	32	26	81.25

Item	Language features	1st correction (Version 1)			2nd correction (Version 2)		
		N	Number of subjects getting the correct answers	%	N	Number of subjects getting the correct answers	%
16	word class	38	24	63.16	37	33	89.19
17	word form	38	31	81.58	37	36	97.30
18	verb pattern	38	26	68.42	37	27	72.97
19	collocation	38	27	71.05	36	32	88.89
20	collocation	38	10	26.32	38	25	65.79

Table 4.4 shows the percentage of subjects who got the correct answer for each item representing different language features in the test (task 1) both the first time by themselves using their existing knowledge and the second time having a corpus available to them. As shown, for all 20 items, the percentage of participants who got the answers for each item right the second time using a corpus was higher than the first time without using a corpus. Except for item 5, about 60–97 percent of the participants answered the item correctly after they were allowed to consult the corpus. The top five language problems that were most successfully solved by this particular group of students using a corpus were word form (item 17), noun phrase/word order (item 13), clause structure: subordinate clause (item 9), word class (item 16), and collocation (item 19) respectively. Some verb and noun patterns seemed to be least easily solved using a corpus as the percentage of students who could correct the errors with the help of the corpus is relatively low (see item 5), and also the percentage of students who could correct items 1, 8, and 12 is quite low (less than 70%) compared to other items. Table 4.4 also gives contradictory results showing that the same type of lexicogrammatical errors such as collocation errors can be classified as both most successfully and least easily solved using a corpus. Whereas the collocation error in item 19 is seen to be one of the lexicogrammatical errors that the students found most successfully

solved using a corpus, the collocation error in item 20 can be rated as one of the lexicogrammatical errors that is least easily solved by the use of a corpus as 65.79% of the students (less than 70%) answered correctly.

As stated, the data in Table 4.4 is based on the results of the test on paper only. Looking more closely into the video files of the participants who did not supply the answers to some items in Version 2 to check if they had consulted a corpus in order to try to correct those items or not, it was found that the results in this table are not totally accurate. Some of the answers these subjects provided were not based on the corpus findings, but they got the answers by asking their friends instead of searching the corpus themselves. This gave rise to the idea that this table needed to be revised by looking at the video file of each participant in order to discard the answers that were not based on the corpus data. However, on the day this data was collected, the video recordings of two students (S4 and S24) undertaking this task were missing. It was probable that the students made a mistake while saving the files. Moreover, while re-analysing this video data from S32, it was found that the video process went wrong. Only half of the recording could be played. Thus these three participants were taken out from this task. Only 35 participants were included, and the following table is based on 35 participants. The crucial difference is the results in the 2nd correction in Version 2.

Table 4.5: The percentage of subjects who got the answers right for each item the first time by themselves and the second time using a corpus (based on the video)

Item	Language features	1st correction (Version 1)			2nd correction (Version 2)		
		N	Number of subjects getting the correct answers	%	N	Number of subjects getting the correct answers	%
1	noun pattern	35	5	14.29	35	21	60.00
2	adjective pattern	35	11	31.43	33	26	78.79
3	noun pattern	35	11	31.43	32	23	71.88
4	question tag	35	19	54.29	22	18	81.82
5	verb pattern	35	1	2.86	34	6	17.65
6	noun class	35	15	42.86	25	21	84.00
7	adjective pattern	35	4	11.43	32	24	75.00
8	verb phrase	35	11	31.43	32	22	68.75
9	clause structure: subordinate clause	35	17	48.57	28	25	89.29
10	relative clause	35	7	20.00	21	18	85.71
11	indirect question	35	12	34.29	21	17	80.95
12	verb pattern	35	9	25.71	21	14	66.67
13	noun phrase/ word order	35	16	45.71	26	24	92.31
14	if clause	35	5	14.29	27	17	62.96
15	noun clause	35	8	22.86	17	13	76.47
16	word class	35	21	60.00	23	21	91.30
17	word form	35	28	80.00	16	15	93.75
18	verb pattern	35	23	65.71	15	13	86.67
19	collocation	35	25	71.43	22	21	95.45
20	collocation	35	10	28.57	28	21	75.00

Note S4, 24, and 32 were excluded

Compared to Table 4.4, the percentage of participants who got the answer for each item right after looking up a corpus in the second time is still higher than the percentage of participants who got the answer right without a corpus. Likewise, up to 60-95% of them could correct the errors

after looking up a corpus, except for item 5. The remarkable change is the order of the top five language problems that were most successfully solved using a corpus. In Table 4.4, word form was ranked the top, followed by noun phrase/word order, clause structure: subordinate clause, word class, and collocation. In this table, collocation (item 19) is at the top, followed by word form (item 17), noun phrase/word order (item 13), word class (item 16), and clause structure: subordinate clause (item 9). In addition, this table does not yield an obvious difference between the percentages of students who got the answers right for collocation errors in items 19 and 20 as found in Table 4.4. This indicates that a corpus is most useful for looking at collocation. Verb and noun patterns remain unchanged as the least easily solved lexicogrammatical errors using a corpus. If-clause structure becomes one of the most difficult lexicogrammatical errors to be solved with a corpus.

As mentioned earlier, the data that are used to answer RQ1 are the data from Version 2 which involved the use of corpora for error correction. To some extent, Tables 4.4 and 4.5 can answer this question. However, there is a big overlap between the number of participants who got the right answers the first time and the second time because some who got the right answer the first time also looked up a corpus the second time, but just to check their first answers in Version 1. For example, (see item 19 in Table 4.5) out of 35, 25 participants got the right answer for this item the first time. In the second time, 22 participants looked up a corpus and 21 of them got the right answer. What is still missing is the proportion of participants who could not get the answers and those who got the wrong answers the first time and got the right answers the second time after looking up a corpus. To illustrate this information, the answers the participants provided the second time were compared with their answers in Version 1. They were then classified into five

categories as follows: Categories A and B are evidence of the corpus searches working. Category C is evidence of the corpus searches being detrimental. Categories D and E show the corpus searches as having no effect.

Table 4.6: Classification of answers the participants got in the second correction in comparison with the answers they got in version 1

Classification	Meaning	Definition
A	= Nil/Right	(No answer is supplied the first time, and the participant got the right answer the second time.)
B	= Wrong/Right	(The participant got the wrong answer the first time and the right answer the second time.)
C	= Right/Wrong	(The participant got the right answer the first time and the wrong answer the second time.)
D	= Nil/Wrong & Wrong/Wrong	(The participant got either no answer or the wrong answer the first time and still got the wrong answer the second time.)
E	= Right/Right	(The participant got the right answer both the first and second time.)

The following is an example of classifying the answers. Twenty-two participants checked a corpus to answer item 19 the second time, and S17 got the wrong answer. On closer analysis, their answer to this item in Version 1 revealed that they could not answer this item either. Their answer to this question in Version 2 was, then, classified as D. After the answers to all 20 items were classified, they were calculated into percentage. The results are in Table 4.7 below.

Table 4.7: Proportion of participants who got the right and wrong answers the second time

Item	Language Features	Number of students who tried the item (N=35)	Nil/Right (A)		Wrong/Right (B)		Right/Wrong (C)		Nil/Wrong & Wrong/Wrong (D)		Right/Right (E)	
			N	%	N	%	N	%	N	%	N	%
1	Noun pattern	35	6	17.14	11	31.43	1	2.86	13	37.14	4	11.43
2	Adjective pattern	33	10	30.30	6	18.18	0	0.00	7	21.21	10	30.30
3	Noun pattern	32	8	25.00	5	15.63	2	6.25	10	31.25	7	21.88
4	Question tag	22	3	13.64	4	18.18	0	0.00	4	18.18	11	50.00
5	Verb pattern	34	2	5.88	3	8.82	0	0.00	28	82.35	1	2.94
6	Noun class	25	10	40.00	3	12.00	2	8.00	2	8.00	8	32.00
7	Adjective pattern	32	17	53.13	5	15.63	2	6.25	6	18.75	2	6.25
8	Verb phrase	32	11	34.38	1	3.13	1	3.13	9	28.13	10	31.25
9	Clause structure: subordinate clause	28	14	50.00	0	0.00	1	3.57	2	7.14	11	39.29
10	Relative clause	21	13	61.90	2	9.52	1	4.76	2	9.52	3	14.29
11	Indirect question	21	7	33.33	3	14.29	0	0.00	4	19.05	7	33.33
12	Verb pattern	21	5	23.81	5	23.81	0	0.00	7	33.33	4	19.05
13	Noun phrase/ word order	26	5	19.23	6	23.08	0	0.00	2	7.69	13	50.00
14	If clause	27	13	48.15	1	3.70	0	0.00	10	37.04	3	11.11
15	Noun clause	17	10	58.82	0	0.00	0	0.00	4	23.53	3	17.65
16	Word class	23	3	13.04	6	26.09	1	4.35	1	4.35	12	52.17
17	Word form	16	1	6.25	2	12.50	1	6.25	0	0.00	12	75.00
18	Verb pattern	15	1	6.67	4	26.67	1	6.67	1	6.67	8	53.33
19	Collocation	22	2	9.09	2	9.09	0	0.00	1	4.55	17	77.27
20	Collocation	28	1	3.57	14	50.00	3	10.71	5	17.86	5	17.86

Table 4.7 provides the proportion of participants who were successful and unsuccessful in using a corpus in order to correct different types of language problems. Although Tables 4.4 and 4.5 indicate that the greater percentage of the participants got most of the answers right after using a corpus, a more detailed analysis in Table 4.7 gives useful information about how these particular students responded to these particular items. Before looking at the types of language problems a corpus seems to be most useful for, it is worth examining the types of language problems that a corpus is found to be less useful for. In this table (see D), 82.35% of the participants looked in the corpus and failed to find the right answer to item 5 which is the verb pattern (*suggest*) as they still got the wrong answer. 37.14% of them were not successful in correcting item 1 (noun pattern) whereas approximately the same amount of 37.04% were unsuccessful in correcting if-clause structure in item 14. 33.33% could not answer item 12 (verb pattern), and 31.25% could not work out the answer to item 3 (noun pattern). This implies that for this group of participants, a corpus seems to be especially least useful for identifying patterns of both verbs and nouns, and also some grammatical structures such as the if-clause.

In some cases, it is quite surprising to find that some students who had successfully corrected the errors by themselves using their existing knowledge in the first test, changed their minds and got the wrong answers after looking up a corpus (see C). For example, three students (10.71%, see item 20) got the wrong answer for collocation, 2 got the wrong answer for noun class (8%, see item 6). The same number of two students got the wrong answers for the noun pattern, and adjective pattern (6.25%, see items 3, and 7). The other types of language problems for which at least one of the initially successful students changed the answer based on the corpus data and

gave the wrong answers were noun pattern, verb phrase, clause structure, relative clause, word class, word form, and verb pattern.

Nevertheless, by looking at the number of students who could correct the errors by themselves without a corpus available to them and changed their mind after looking up a corpus and eventually ended up getting the wrong answers in isolation, one would argue that the numbers in category C are so low that they could be treated as accidental. Adding together the numbers of items where corpus use is found to be detrimental in category C with the number of items where corpus use has no effect because it failed to lead to successful correction in category D would provide a more useful way of looking at the kinds of lexicogrammatical errors that a corpus is least useful for than solely looking at the data in category D. The combination of C and D is provided in Table 4.8 below, ranging from high to low.

Table 4.8: Number of students who failed to use a corpus to correct each item

Item	Language feature	Number of students who tried the item	Right/Wrong (C)	Nil/Wrong & Wrong/Wrong (D)	Total number of C+D	
					N	%
5	Verb pattern	34	0	28	28	82.35
1	Noun pattern	35	1	13	14	40
3	Noun pattern	32	2	10	12	37.5
14	If clause	27	0	10	10	37.04
12	Verb pattern	21	0	7	7	33.33
8	Verb phrase	32	1	9	10	31.25
20	Collocation	28	3	5	8	28.57
7	Adjective pattern	32	2	6	8	25
15	Noun clause	17	0	4	4	23.53
2	Adjective pattern	33	0	7	7	21.21
11	Indirect question	21	0	4	4	19.05
4	Question tag	22	0	4	4	18.18
6	Noun class	25	2	2	4	16
10	Relative clause	21	1	2	3	14.29

Item	Language feature	Number of students who tried the item	Right/Wrong (C)	Nil/Wrong & Wrong/Wrong (D)	Total number of C+D	
					N	%
18	Verb pattern	15	1	1	2	13.33
9	Clause structure: subordinate clause	28	1	2	3	10.71
16	Word class	23	1	1	2	8.70
13	Noun phrase/ word order	26	0	2	2	7.69
17	Word form	16	1	0	1	6.25
19	Collocation	22	0	1	1	4.55

From Table 4.8, where categories C and D are added together, we can see which items are most likely to lead to non-improvement. The highest number of students found it most difficult to search a corpus to correct errors concerning verb pattern (item 5), noun pattern (items 1 and 3), if-clause (item 14), and verb pattern (item 12), respectively. These results are the same as those found in category D in Table 4.7. The slight difference is that the students in category D found it more difficult to correct if-clause (item 14) and verb pattern (item 12) than to correct noun pattern (item 3). Students in categories C and D, added together, on the other hand, found it more difficult to correct noun pattern (item 3) than to correct if-clause (item 14) and verb pattern (item 12). The common feature shared between these two sets of data is that verb pattern (item 5), noun pattern (items 1 and 3), if-clause (item 14), and verb pattern (item12) are the lexicogrammatical errors that this group of learners found it most difficult to solve with a corpus. Taking into account the kinds of language features that a corpus is found to be least useful for, the most crucial part of Table 4.7 that gives more accurate results to RQ1 is the number of participants who changed their mind after searching a corpus and subsequently got the right answers. Therefore, the number of participants who could not answer the items the first time and

got the right answers the second time (A) and the number of participants who got the wrong answers the first time and got the right answers the second time (B) were added together. The results were then converted into a percentage from the highest to lowest as shown in the following table.

Table 4.9: Number of students who changed their answer for each item after looking up the corpus and got the right answer

Item	Language feature	Number of students who tried the item	Nil/Right (A)	Wrong/Right (B)	Total number of A+B	
					N	%
10	Relative clause	21	13	2	15	71.43
7	Adjective pattern	32	17	5	22	68.75
15	Noun clause	17	10	0	10	58.82
20	Collocation	28	1	14	15	53.57
6	Noun class	25	10	3	13	52.00
14	If clause	27	13	1	14	51.85
9	Clause structure: subordinate clause	28	14	0	14	50.00
1	Noun pattern	35	6	11	17	48.57
2	Adjective pattern	33	10	6	16	48.48
11	Indirect question	21	7	3	10	47.62
12	Verb pattern	21	5	5	10	47.62
13	Noun phrase/ word order	26	5	6	11	42.31
3	Noun pattern	32	8	5	13	40.63
16	Word class	23	3	6	9	39.13
8	Verb phrase	32	11	1	12	37.50
18	Verb pattern	15	1	4	5	33.33
4	Question tag	22	3	4	7	31.82
17	Word form	16	1	2	3	18.75
19	Collocation	22	2	2	4	18.18
5	Verb pattern	34	2	3	5	14.71

Close inspection of the number of students who changed their mind after looking up a corpus and got the right answer in Table 4.9 reveals that the high percentage of students who were not successful in correcting the error by themselves became most successful when they searched a corpus to check relative clause (71.43%, see item 10), adjective pattern (68.75%, see item 7), noun class (58.82%, see item 15), collocation (53.57%, see item 20), noun class (52%, see item 6), if clause structure (51.85%, see item 14), and subordinate clause structure (50%, see item 9). On the other hand, the students were least successful when they looked up a corpus to correct the errors concerning the verb pattern (14.71%, see item 5), collocation (18.18%, see item 19), word form (18.75%, see item 17), question tag (31.82%, see item 4), verb pattern (33.33%, see item 18), and verb phrase (37.50%, see item 8).

As seen, errors of the same types such as collocation and verb phrases could be classified as both most successfully solved and least successfully solved using a corpus. It, therefore, can be concluded that it is not possible to predict the type of language problem a corpus is most useful for. More discussion of these results is given in the next section.

4.8 Discussion

As the study seeks to identify what language problems the participants find it easiest to solve using a corpus, the results, to some extent, are beyond our expectation. Sometimes, the participants could find the answers to the items that are expected to be difficult for them to notice in a corpus. In other cases, the participants failed to notice the use of word or the pattern of word that can be simply found from concordance lines. In other words, some of the lexicogrammatical errors in the test are unexpectedly easy for the participants to solve using a corpus and some are found to be more difficult for them to solve than expected. For this reason, the results require

further analysis in order to better understand which items meet with or are contrary to my expectation. Some of the lexicogrammatical errors posed in the error correction test used in this study were expected to be easily solved with a corpus and some were expected to be difficult to solve. The following shows what was expected.

Table 4.10: Expectation of difficulty of each item

Item	Language features	Classification	Error	Expectation
1	noun pattern	grammar of class	an increase with	easy
2	adjective pattern	grammar of class	difficult for farming	easy
3	noun pattern	grammar of class	…relationship between his guess…	easy
4	question tag	grammar of structure	…We had a good time…, did we,…	difficult
5	verb pattern	grammar of class	suggested her to return	easy
6	noun class	grammar of class	The furnitures were	easy
7	adjective pattern	grammar of class	wise that you rent	easy
8	verb phrase	grammar of structure	did not ought to	easy
9	clause structure: subordinate clause	grammar of structure	While study	easy
10	relative clause	grammar of structure	which he made it	difficult
11	indirect question	grammar of structure	wonder why is the public always wrong	difficult
12	verb pattern	grammar of class	it isn't allowed to	difficult
13	noun phrase/word order	grammar of structure	white Victorian big house	easy
14	if clause	grammar of structure	…it would be very different if his wife had been with him.	difficult
15	noun clause	grammar of structure	What do they want to point out here is…	difficult
16	word class	grammar of class	…kept a *(save/safe/safety)* distance	easy
17	word form	grammar of structure	…some of *(who/whom)* he knew…	difficult

Item	Language features	Classification	Error	Expectation
18	verb pattern	grammar of class	…get it *(fix/fixes/fixed/fixing)*	easy
19	collocation	Lexis	I *(made/did/had)* the mistake…	easy
20	collocation	Lexis	…will be *(heavy/big)* losses…	easy

From Table 4.10, 13 items were expected to be easy and 7 items were expected to be difficult for the participants to find in a corpus. In principle, when designing the error correction test, I wanted to include items or language features that could be easily found in a corpus or language features that I believed a corpus would be most useful for. With this perspective, errors concerning noun pattern (items 1, 3), adjective pattern (items 2, 7), verb pattern (items 5, 18), noun class (item 6), verb phrase (item 8), subordinate clause structure (item 9), noun phrase (item 13), word class (item 16), and collocation (items 19, 20) which were believed to be easily found in a corpus were included. Other language features which were expected to be difficult like question tag (item 4), relative clause (item 10), indirect question (item 11), verb pattern (item 12), if clause (item 14), noun clause (item 15), and word form (item 17) were included in the test with the intention of testing the hypothesis that these features were difficult for the participants to find. With this expectation, Table 4.11 below shows to what extent the results match the expectation. Taken from Tables 4.8 and 4.9, the information in these tables represents the two things we need to know, the percentage of non-improvers and the percentage of improvers for each item. It is difficult to know how to interpret this figure because there is no obvious or correct way to do so. Some of the students deteriorated using a corpus while others improved. There is nothing to say, therefore, how the results should be interpreted. If we look at the percentage of non-improvers in item 19, it looks very easy because only 4.55% of the students could not get the right answer. On

the other hand, by looking at the percentage of improvers, it does not make a lot of sense to justify that this item is easy as it is as low as 18.18%, suggesting that the results are conflicting. Therefore, I took these two numbers together to decide whether the item is easy or difficult to correct with the help of a corpus, and these three possibilities of data interpretation were identified. If fewer than 50% of the students did not improve and 50% or more of them improved, the item is classified as easy. Conversely, if at least 50% of the students did not improve and fewer than 50% of the students improved, the item is labelled as difficult. If the number of both students who did not improve and those who improved is fewer than 50%, this means that the item is variable or uncertain, as follows.

<50% non-improved; ≥50% improved = easy

≥50% non-improved; <50% improved = difficult

<50% non-improved; <50% improved = variable

Item 1 is taken as an example. After checking in a corpus, 40% of the students still got it wrong or did not improve and the other 48.57% of them got it right or improved, so this item is regarded as variable.

Table 4.11: Comparison of expectation and results from Tables 8 and 9

Item	Language features	Expectation	% of non-improvers	% of improvers	Results
1	noun pattern	easy	40	48.57	variable
2	adjective pattern	easy	21.21	48.48	variable
3	noun pattern	easy	37.5	40.63	variable
4	question tag	difficult	18.18	31.82	variable
5	verb pattern	easy	82.35	14.71	difficult
6	noun class	easy	16	52	easy
7	adjective pattern	easy	25	68.75	easy

Item	Language features	Expectation	% of non-improvers	% of improvers	Results
8	verb phrase	easy	31.25	37.50	variable
9	clause structure: subordinate clause	easy	10.71	50	easy
10	relative clause	difficult	14.29	71.43	easy
11	indirect question	difficult	19.05	47.62	variable
12	verb pattern	difficult	33.33	47.62	variable
13	noun phrase/word order	easy	7.69	42.31	variable
14	if clause	difficult	37.04	51.85	easy
15	noun clause	difficult	23.53	58.82	easy
16	word class	easy	8.70	39.13	variable
17	word form	difficult	6.25	18.75	variable
18	verb pattern	easy	13.33	33.33	variable
19	collocation	easy	4.55	18.18	variable
20	collocation	easy	28.57	53.57	easy

Table 4.11 shows that the results are mixed. There are items that match expectation and items that are contrary to expectation. Unexpectedly, most of these items are classified as variable, meaning that it is difficult to decide whether they are easy or difficult because they are easy for some students but difficult for others. The four items that match the expectation are items 6, 7, 9, and 20 which were expected to be easy and appear to be consistently easy. The items that are contrary to expectation are items 5 which was expected to be easy but appears very difficult, and items 10, 14, and 15 which were expected to be difficult but turned out to be incredibly easy. 12 items found to be variable are items 1, 2, 3, 4, 8, 11, 12, 13, 16, 17, 18, and 19, most of which were expected to be easy except items 4, 11, 12, and 17.

It is also interesting to find that none of the 7 items that were expected to be difficult met expectation as three of them are easy for the students (items 10, 14, and 15) and the other four (items 4, 11, 12 and 17) can be variable. That the proportion of variable or uncertain items is

much larger than that of the easy and difficult items (12:7:1) is also worth noting. This greatest number of variable items suggests that all that can be concluded is that there is no clear evidence for the corpus to have helped a great deal. If we compare the variable items such as items 1 and 2, the result for item 1 is very variable because nearly half of the students deteriorated and about half of them improved. In item 2, however, a lot more students improved rather than deteriorated, so my conclusion would be to say that it is easier for them to find the answer to item 2 than to item 1 because the percentage of non-improvers is much lower. When all these variable items are consistently compared in order to identify which item functions better and put in order from functioning best, the order is 13, 16, 11, 2, 18, 12, 4, 19, 17, 1, 8, and 3, suggesting that item 13 functions better than item 16 and so on.

This table leads to the answer to the research question raised in this experiment. It indicates that, for this particular group of students, noun class (item 6), adjective pattern (item 7), clause structure: subordinate clause (item 9), relative clause (item 10), if-clause (item 14), noun clause (item 15), and collocation (item 20) are most successfully solved using a corpus. On the contrary, it is obvious that they found verb pattern (item 5) very difficult to observe in a corpus. These results, to a greater extent, are wholly surprising and beyond expectation. While we might predict an item to be easy, sometimes it is easy and sometimes it is difficult. When we expect an item to be difficult, sometimes it turns out to be surprisingly easy.

Given that they were mainly taught to conduct single word searches, some linguistic features like clause structure were expected to be difficult to find on a corpus and some simple collocation, verb and noun patterning, and word class were expected to be easy to find in a corpus. Yet the results are different, and this raises the question of why the students became so successful in

using a corpus to figure out an abstract concept such as subordinate clause structure (item 9), relative clause (item 10), if-clause (item 14), and noun clause (item 15) which is expected to be more difficult to check on a corpus, and found it difficult to identify simple errors like verb pattern in items 5 and 12 and word class in item 16. For example, they were able to correct the relative clause error in *"Good Time Jazz has released a nice two-record album which he made it."* (item 10) and the noun clause error in *"What do they want to point out here is that…"* (item 15), but were very unsuccessful in correcting the verb pattern error in *"Maude suggested her to return to New York."* (item 5) and the word class error in *"All bottles must be kept a (save/safe/safety) distance away from the pool…"* (item 16).

This is probably due to the fact that these students had been taught grammatical structure previously, and they already had the concept of relative clause and noun clause, so they were actually able to find the answer from a corpus. Also their failure to understand the use of *suggest* and class of the word *safe* might have been caused by a lack of awareness or knowledge of word usage which might have been less explicitly taught than grammatical structure in Thailand.

In this chapter, I have reported experiment 1 and discussed the results. The next chapter will discuss Reseaech Question 2: Language Features the Students Look for from a Corpus.

4.9 Conclusion

The purposes of this experiment are twofold: 1) as a preparation for the subsequent task, 2) to give a comparability of the students. It is necessary that I give them a task that is not varied in a controlled way. I want to have control over what the students are doing in order to get comparability of learners. In the other tasks that follow, the students will have full control over

their use of concordances, so the results will be very varied. My expectation as a teacher is that the students would find some errors easy to be solved using a corpus and they would find some other errors difficult to be solved using a corpus. However, the results suggest that it is impossible to predict what errors are easily solved using a corpus and that language concepts play an important role in investigating corpus data to find out how language works. If the students have concepts of the language, they will find it with ease.

Chapter 5

Language Features the Students Look for from a Corpus

5.1 Introduction

In the previous chapter, types of lexicogrammatical errors Thai learners of English find it easiest to solve using a corpus were discussed. This chapter gives an explanation of a writing and corpus-based editing task and presents the results. The aim of the task is to answer research question 2: When students are writing essays, what language points are they most likely to check in a corpus, and how do they go about it? The method of data collection, data analysis, and results are explained below.

5.2 Method

As mentioned in the methodology chapter, by the time the data for this question was about to be collected, the participants had already written their first drafts of an academic paper and they were in the process of doing a peer review to refine the drafts. Therefore, instead of asking the participants to write in class for my own research purpose, I decided to use this piece of writing in my research and asked the students to identify their own errors by themselves and edit them by checking a corpus (task 2 of my study). As reported in chapter 3, I observed that the students had problems identifying their own errors and were not able to find enough language problems to look at in a corpus, so I scrutinized their drafts and located some of the errors for them to correct again using a corpus. In total, language errors in this work were identified twice: the first time by the students themselves and the second time by me. The main problem I found in reading these pieces of writing and locating the errors for the students to check in a corpus was that this work

did not reflect their own writing ability. I suspected that the students had copied it from the original sources. Thus, I tried to provide a solution for them to write something that would reflect their genuine writing ability but would not give them too much writing burden. Most importantly, the work needed to somehow benefit their course.

Fortunately, during the last week of the course Academic Writing in English, the course instructors planned to have the students write in class about what they had learned or gained from the course. In addition to task 2, where the students wrote as homework and they were allowed to consult different resources available, I decided to include this writing task in the study because it would reflect the students' true ability in writing as they had no access to other resources other than a corpus. To undertake the task in an environment that would contribute to my study, both instructors agreed to have the students complete the writing task in the computer room under my control. The aim of the task was to have the students use a corpus to check lexis and grammar while writing in English or editing their writing before submitting their writing to the instructors. The students were given the topic to write about in class for 90 minutes. The topic for the students enrolled in section 1 (morning class) was *"What I have learned from the course Academic Writing in English"* and that of section 2 (afternoon class) was *"What I have gained from the course Academic Writing in English".* Both topics were assigned by the instructors. Within the time allotted, the students who participated in my study were asked to use the corpus to check the words or phrases which they did not know how to use either while writing or editing their writing. They were also asked to underline the errors or words they had checked from the corpus.

As in task 1, while searching the corpus and editing their writing, the students were asked to think aloud, capture the computer screen in progress, and record their think-aloud protocol using the Camtasia Studio6 program. At the end of the task, they were asked to save the files using their code number in the file name. Also, they wrote their code number on their paper work so that both the paper work and electronic files had their code number and could be easily matched up.

5.3 Data analysis and results

In order to identify what language features the students were looking for in a corpus or what language problems they were trying to solve using a corpus while writing in English, the videos the students made while doing the task were used to investigate the kinds of linguistic features the students sought to learn from a corpus to assist their writing in English. In particular, in this task, I needed to find out what the students were trying to do, how they did it, what they were thinking , and whether or not they accomplished what they wanted to find. In cases where the searches did not lead them to a successful outcome, I suggest what searches would have been more appropriate.

To obtain these data, I looked at all 38 videos and transcribed them. In the transcripts, I also added the details about how the students navigated the searches and processed the corpus data. From the video transcripts, I listed all the queries each student made, noted down what question or hypothesis the student posed or what information the student wanted to find about each query, and identified what the student found in each search. Then, to judge whether or not the students were successful or able to find what they wanted to find, I went through each student's writing to identify the target form that the student wanted to find. For example, in making sense of how the student (S1) dealt with the query *applying*, I watched the video to discover what it was that this

student wanted to find out and found that they wanted to know if the verb *apply* is followed by the preposition *in*. After looking at the concordance lines for *applying*, they concluded that *applying + in* was correct. Then, I went through their writing to check the target form that they actually wanted to find. Based on their writing *'I have learned many things for applying in every type of writing.',* I discovered that in this context the target form of *applying* that they wanted was *apply + to.* Therefore, I decided that they were not successful because what they found does not match the target form in the student's writing context even though *applying + for* is correct in the sentence like *'You're still applying for the wrong jobs.'*

While watching these videos, it was found that one student (S2) did not think aloud at all, so their work was taken out from the analysis of this task. Out of 38, 37 videos were used in analysis. In these useable videos, it was found that, sometimes, the students did not constantly think aloud or did not state clearly in the think-aloud protocols what they wanted to find and what they found about the search while making a query and interpreting the results for some certain queries. As a result, it is difficult to make sound judgement about what they wanted to learn from that query or what they found from the search results. I had to go back to the writing to check what it was that the students wanted to find and to the videos to look at how the students handled the searches and processed the corpus data to gain as much detail about the search as possible. The following is a summary of the number of problems the students looked up in a corpus and the number of queries they made. See Appendix 6 for full listings of queries made by individual students.

Table 5.1: Number of problems and queries the students looked up in a corpus

Subject	Number of problems	Number of queries
01	5	6
03	8	13
04	4	4
05	6	7
06	2	2
08	3	6
09	3	3
10	2	5
14	3	3
15	2	2
16	1	1
17	3	3
19	2	2
20	2	3
21	6	6
22	5	5
23	2	2
24	6	6
26	4	4
27	3	8
28	3	4
29	2	2
30	3	3
31	3	3
32	4	5
33	2	3
34	3	7
35	2	2
36	12	12
37	5	5
38	5	5
39	4	5
40	6	6
41	4	4
42	1	3
43	8	9
45	2	5
Total	**141**	**174**

Table 5.1 shows that, while undertaking this writing task, the students (N=37) tried to solve 141 problems and they made 174 queries in total. It can also be seen that the number of problems did not match the number of searches that the students attempted because sometimes the students tried several queries per problem. The average number of problems per student is 3.81, ranging from a minimum of 1 to 12. The majority of students had between 2 and 3 problems. Generally, the students made only one attempt to solve most problems either because the attempt was successful or because they gave up, and the highest number of searches related to a single problem is 6. The high number of searches or attempts per problem indicates that the students were initially unsuccessful in devising and performing the searches. Therefore, the initial attempts were unsuccessful and more searches had to be made.

In terms of searches, it was found that sometimes the students performed a single-word search and sometimes they performed a string search, putting in two or more words in the string. This gives rise to the question: what class of words did the students look for in a corpus when they did a single-word search and what word classes were combined together when the students did a string search? The results grouped by word class are as follows.

Table 5.2: Classes of words the students looked up in a corpus while writing/editing their writing

Class of words	Number of searches	Percentage
Noun	57	32.76%
Verb	56	32.18%
Adjective	28	16.09%
Adverb	11	6.32%
Determiner	8	4.60%
Preposition	5	2.87%
Conjunction	4	2.3%

Class of words	Number of searches	Percentage
Modal	3	1.72%
Pronoun	2	1.15%
Total	**174**	**100%**

As shown in Table 5.2 (see Appendix 7 for full listings), classes of words that the students searched in the corpus while writing in English, ranking from most to least frequent, are noun, verb, adjective, adverb, determiner, preposition, conjunction, modal, and pronoun. The number of nouns (57=32.76%), verbs (56=32.18%), adjectives (28=16.09 %), and adverbs (11=6.32 %), vastly outnumbers that of determiners (8=4.60 %), prepositions (5=2.87 %), conjunctions (4=2.30 %), modals (3=1.72 %), and pronouns (2=1.15%), indicating that when searching a corpus for linguistic investigation, the students tended to search for content or lexical words rather than grammatical words.

Moreover, it is clear (see Appendix 7) that when checking content words in the corpus, in addition to checking a single word, the students also looked at strings of words or compound words, ranking from 2-5 word strings. This is quite obvious when they looked up nouns and verbs. As shown in Appendix 7, the number of compound nouns the students looked up in the corpus is greater than the number of nouns as a single word (35:22). However, when looking up a verb, it seems that the students looked up verbs as a single word rather than a combination of verbs with other words like nouns or pronouns as the number of verbs as single words is slightly greater than the number of verbs combined with other words.

When comparing the total number of single word searches with the total number of string searches, it was found that the percentage of single word searches (55.17%) is far greater than

that of the word cluster searches (44.83%). This implies that when using a corpus to check how a language works, the students tended to search for a single word rather than a string of words.

After having gained the information about the classes of words the students looked in the corpus as in Table 5.2, the next step is analysing the data in order to investigate the language features the students attempted to look for in the corpus while searching those words. To answer this, only the queries made while the students were thinking aloud were taken into account as the think aloud protocol would have guided the questions the students posed while conducting searches. The queries that the students made without thinking aloud were not taken into account as there was no clear evidence to support what the students wanted to find out about those searches. The results are shown in Table 5.3 below.

Table 5.3: The language features the students look up from a corpus while writing in English

Language features	Total	Percentage
Colligation	51	29.31%
Collocation	28	16.09%
Acceptability of strings	23	13.22%
Agreement	15	8.62%
Word class	8	4.60%
Nouns in the plural	5	2.87%
Position	5	2.87%
Lexical word + to	4	2.30%
Form	4	2.30%
No information	31	17.82%
Total	**174**	**100%**

Table 5.3 illustrates the types of language features the students investigated from a corpus (see full listings in Appendix 8). In this study, it was found that the students looked up a corpus to

check the following language features. These language feature categories emerged inductively from the participants' think-aloud protocols.

- colligation (a co-occurrence of a grammatical word or class and a lexical word e.g. *could* + verb, *advanced* + *in*)
- collocation (a co-occurrence of two lexical words e.g. *obvious difference*, *major subject*)
- acceptability of strings (an occurrence of the specified string in the corpus e.g. *in my research, knowledge of how to*)
- determiner-noun agreement (a co-occurrence of a determiner and class of nouns e.g. *another* + singular noun, *every* + singular noun)
- word class (part of speech of the target word)
- nouns in the plural (the plural form of noun)
- position (the position where the string occurs in a sentence)
- forms (the form of words based on their functions in the sentence e.g. the adjective form of *bore* is *bored*)
- lexical word + *to* (an occurrence of the target word + *to* e.g. *procedure* + *to*)

The most common language features that the students sought to learn from the corpus are colligation, collocation, acceptability of strings, and agreement. A more detailed analysis of how the students did the searches to find out about these lexicogrammatical errors is given in tables 5.4-5.14 below. The tables consist of four columns. The first column shows a list of queries the students typed in. The second column gives an idea of what the students wanted to find out from the searches. The third column shows what the students found from the searches. The data in the first three columns are based on the videos and the think-aloud protocols. The fourth column is

guidance on what the correct English or real English is based on the students' context of writing and my knowledge of English. The categorisation below will never be watertight and the purpose is not to offer a watertight categorisation of the types of searches because there are too many variables.

Table 5.4: Word-Preposition colligation

Query item	Student's hypothesis	Student's finding	Target form
applying (S1)	applying + in	applying in	to apply to
emphasize (S1)	emphasize + PREP	emphasize + noun	emphasize something/a point
appropriate (S9)	appropriate + for	appropriate for	appropriate for (I choose a topic appropriate for an academic paper.)
study of (S10)	study + PREP	study of + subject	study (V) + subject
study in (S10)	study + PREP	study in + year	study + Noun
write (S17)	write + PREP	write + with/in (If the writer writes in incorrect grammatical structure,…)	write ungrammatically
skill (S22)	skill + PREP	skill level level of skill (…show your [level skill] skill level of English)	skill in English level of English skill
plagiarism	plagiarism + by	plagiarism + by	avoid plagiarism by + V-ing
pay attention	pay attention + to	pay attention to	pay attention to
educated	educated + about (…my teacher taught me to educated about a research paper.)	educated + at	educated about
be advanced in	advanced + in	advanced + in	be advanced in
advanced in	advanced + in	advanced in + V-ing	advanced in + V-ing
successful	successful + PREP	successful + in	successful in
emphasize (30)	emphasize + on	emphasize + in	emphasize something/a point
majoring in (S32)	major + in	majoring + in	major in
advantageous	advantageous + to	advantageous + to	advantageous to (somebody)

Query item	Student's hypothesis	Student's finding	Target form
taught (S35)	taught + PREP	taught + Noun	taught + somebody + Noun
title (S36)	title + of + paper	title paper	paper title
learned	learned + about	learn something	learn something
learn (S24)	learn + PREP (learned new skill)	learn + to (learned to new skill)	learn something
learnt (S24)	learn + PREP (I learnt about how to write…)	learn + from (I learnt from how to write reference)	learn how to
rewrite	rewrite + in + their own words	rewrite + in + their own words	rewrite in their own words
feel	feel like (This class make me feel like writing because I am more effective writing…)	feel like=ok	feel that (…makes me feel that I become a more effective writer)
paraphrase (S40)	paraphrase + PREP	No conclusion	paraphrase a point

Table 5.4 shows that when looking for word-preposition colligation to find out if the target word needs any preposition after it, the students worked this out in three different ways. The first group of students simply typed in one single word such as *emphasize* (S1), *write*, *skill*, *successful, taught, learn, learnt,* and *paraphrase* and tried to figure out the preposition that tends to occur after these keywords or whether or not each of these words needs any preposition after it. The second group of students typed in the keyword to check if a certain preposition they thought of occurs after the keyword. For example, they searched the words *applying*, *appropriate, plagiarism, pay attention*, *educated, emphasize* (S30), *learned*, *advantageous*, *rewrite*, and *feel* to see if *applying in, appropriate to, plagiarism by, pay attention to, educated about, emphasize on, learned about, advantageous to, rewrite in,* and *feel like* occur in English. The last group checked the word colligation by typing in both the keyword and the preposition. For example, they typed in *be advanced in, advanced in,* and *majoring in* to check if *advanced* and *majoring* are used with the preposition *in*.

From these results, there are two main distinctions that can be drawn. First, when the students do not actually know what preposition the target word would take or whether or not the target word needs any preposition, they tended to be more open-minded and made a more general query by putting in only the keyword. In a sense, this is an effective way of using a corpus to find information about word-preposition colligation, and the students are expected to do this when looking for such information. However, it was found that the first group of students who employed this strategy was not satisfactorily successful in doing so. Some of them were successful in searching for word colligation of *emphasize* (emphasize something), *successful* (successful in), and *taught* (taught something). The other students in this group, on the other hand, failed to look up word colligation of *write*, *skill*, *learn*, *learnt*, and *paraphrase*. It is quite obvious that the failure is mainly caused by the mismatch between what the students found and what they were expected to find to match the context in their writing. For example, in making a query *write*, the student (S17) wrote '*If the writer writes with incorrect grammatical structure, the reader might don't get the point*.' They wanted to check if the word *write* would need any preposition and they discovered that it could be followed by *with* or *in.* They then changed their sentence to *'If the writer writes in incorrect grammatical structure,…'* Although their first version '*write with incorrect grammatical structure'* sounds more acceptable, they were expected to discover the use of *write + adverb* and changed their sentence to *write incorrectly* or *write ungrammatically*.

The other two examples that best illustrate this point are the queries *learn* and *learnt* by S24. This student wanted to find out if the words *learn* and *learnt* would need any preposition. They found that *learn* could be followed by *to*, an infinitive marker, and preposition *from*, so they

changed their sentence from *'...I learned new skill in writing…'* to *'...I learned to new skill in writing…'* and from *'...I learnt about how to write reference.'* to *'...I learnt from how to write reference.'* Though it is true that *learn* can be following by *to*, as in *'learn to write'* or by *from* as in *'learn from experts'*, it is not followed by a preposition in the context of '*I learned a new skill'*. Second, when the students want to check if the target word usually occurs with a specific preposition they thought of, they tended to make a more specific hypothesis about the query they made by simply asking if the target word occurs with that specific preposition. As mentioned above, to find this information, they made queries in two different ways by typing in only the keyword (group 2) and by typing in both the keyword and the specified preposition (group 3). The students who put only the keyword tended to be more flexible and came up with valid interpretation correctly applicable to the contexts in their writing like *appropriate + for, pay attention + to, learned + something, advantageous + to, rewrite + in + their own words,* and *feel + like.* However, a problem can occur when the target word has varied phraseology, and the students sometimes came up with an acceptable interpretation but not applicable to their writing contexts. For example, they found *applying + in*, *educated + at*, and *emphasize + in* whereas they were expected to find *applying + to, educated + about*, and *emphasize + noun*. One student (S23) also accidentally found that *plagiarism* is followed by preposition *by* without being aware that *by* in their writing belongs to *avoid xxx by* and that *plagiarism by a man* is different from *avoid plagiarism by paraphrasing*. Being unable to parse the sentence either in the concordance lines or in their own writing can hinder the students from interpreting the results of their searches successfully. The strategy by the third groups of students who typed in both the keyword and the target preposition is more specific and tends to work if the target word has only one phraseology like *advanced in* and *pay attention to*.

Table 5.5: Lexical-word class colligation

Query item	Student's hypothesis	Student's finding	Target form
almost (S16)	almost + Noun (almost countries)	almost + NOUN	many/most + Noun
help (S19)	help + someone + Verb or to infinitive	No conclusion	help someone + Verb
become (S23)	Can I say 'help me become without 'to'?	Search given up	help me become
these brought me (S3)	Can I say 'These brought me take the course Academic Writing in English? pattern: bring me + noun or verb,	No hits	This led me to take the course Academic Writing in English.
brought (S3)	how to use it?	Search given up	led
brought me (S3)	bring me + word class	bring me + Noun bring me + Noun so I'd better change the verb into a gerund (These brought me taking the course Academic Writing in English)	This led me to take the course…
usually	usually + word class	No conclusion	usually + Verb
useful (S36)	How to use it	Be + useful	Be + useful
academic (S29)	more + academic	Search given up	more academic
look over	look over + word class	(S30) look over + Noun/NP	look over + Noun
thought on	thought on+ word class	(S31) thought on + gerund/noun	thought on + Noun
realize	realize + word class	realize + wh-word	realize how
help me improve (S32)	Can I say help me improve without 'to'?	No hits	help me improve
help improve (S32)	Can I say help me improve without 'to'?	help + me + improve	help me improve
study	study + word class	study + Prep	study + Noun
believable (S36)	Its function in the sentence	To modify a noun: believable + Noun [believable websites])	believable + Noun

Query item	Student's hypothesis	Student's finding	Target form
perfectly (S36)	perfectly + word class	Verb + perfectly Perfectly + Adj + Noun	perfect (to write a perfect research paper)
afraid (S40)	Is 'I am afraid' correct?	Yes: S + Be + afraid	Be + afraid
not only (S3)	How to use it? not only…+ but also…?	not only…+ but also…	not only…+ but also… …

If the students were looking for lexical-word class colligation, they most frequently asked a general question about the word class that comes after verbs, adverbs, and noun + preposition. See table 5.5. For example, a student looked at *help* to find if it is followed by to infinitive. One of the students looked for *almost* to check if it can be followed by a noun. An example search for noun + preposition is the search for *thought on* to find out the class of words that comes after preposition *on*. One student searched for the adjective *believable* to find out if it can be used as an attributive adjective like *believable websites*. In other less frequent examples, the students might have had a very specific question or a very general question about the search. An example of a very specific question of phraseology is students asking if it was correct to say *'I am afraid'* and *'more academic'*. A very general question asked is how to use *useful*.

To work this out, the students conducted the searches in three ways: (a) by typing in one word e.g. *almost, help, usually, become, perfectly,* (b) by typing in two words e.g. *brought me, thought on, help improve,* (c) by typing in three words or more e.g. *these brought me, help me improve, what are they like.* The successful queries from which the students were able to draw correct conclusions are *useful (useful + Noun), look over (look over + Noun), thought on + (Noun), realize (realize + wh- word), help improve (help + Verb, without to), believable (believable + Noun), afraid (Be + afraid),* and *not only (not only… + but also).* This suggests that successful

searches were likely to consist of one or two words. However, as we shall see, searching for one or two words is not a guarantee of success.

Failures in this, to some extent, are caused by wrong input. The students searched for a word or a sequence of words that was incorrect and did not recognise that the concordance lines did not match their own writing. For example, they searched for *almost, brought me*, and *perfectly* whereas they were expected to use *many, led me*, and *perfect* in their writing. Another cause of failure is from the corpus itself. This happened when the students searched for a string that is correct but which through happenstance does not occur in the corpus used i.e. *these brought me* and *help me improve.* However, by far the most important cause of failure is a lack of expertise in interpreting concordance lines. For example, some students were unable to interpret concordance lines for *help* and *usually* while others drew incorrect conclusions from the concordance lines and gave up interpreting the output of *become, brought* and *academic*.

Table 5.6: Grammatical-word class colligation

Query item	Student's hypothesis	Student's question	Target form
never + before (S1)	never + Verb form + before	(never + V3 + before)	have + never + V3 + before
after	after + VERB form	after + V-ing	after + V-ing
could (S38)	could + VERB	could + Verb	could + Verb
must (S17)	Is it correct to say 'must don't'?	must + not	must not
during process (S10)	during + process	No hits	- During the process of + V-ing - While processing

Query item	Student's hypothesis	Student's question	Target form
process_N* (S10)	PREP + process	in + process during processing (during processing the paper, there was a peer review activity)	in the process of writing the paper
for	Is it correct to use *'for'* in *'... I have to check my peer's writing carefully for the best text.'?*	No conclusion made	- to (to make it the best writing) - so that (…carefully so that it works out the best)
with (S31)	with + Noun	with +Noun/gerund	the same way with + Noun

It can be seen from table 5.6 that in looking for grammatical-word class colligation, the students sought to find a connection between a particular grammatical word and a word class that follows. Basically, they were asking what kind of words they could use after these grammatical words: *after, could, must, during,* and *with*. Except for *during*, which the student typed in as two words: *during process*, the students often worked this out by typing in one single word: *after, could, must,* and *with*. They were able to observe the correct forms of *after + V-ing*, *could + Verb*, *must + not*, and *with + Noun* and *gerund*, but the student who looked for *during process* to find out if *during* could be followed by the verb *process* obtained no concordance lines for it. This student then tried a new query *process_N** to find out if *process* can be a noun and which preposition comes before it, sorted the concordance lines to see the prepositions on the left, and found that it is often preceded by *in, of,* and *with*, not *during*. For this reason, they said it might be correct to say *during processing* and they wrote *'...during processing the paper, there was a peer review activity'*. Instead, to check what word class occurs after *during*, the student should have tried the query *during* and noticed the pattern *during + Noun* or *during the process of + Noun/gerund* as found in *'during the process of economic reform'*, and '*During the process of learning*', etc. An

alternative search to this is to type in the query *processing* and observe the pattern *while + processing*, e.g. *'while processing a claim from a former lover'* and *'while processing your transaction'*.

The student who searched for *for* to find out if it is correct to say *'... I have to check my peer's writing carefully for the best text.'* was also unsuccessful because they were unable to draw a conclusion from the concordance lines. In this case, the student had a problem concerning word choice and meaning, and it seems to be complicated to consult a corpus about this. Instead, they should have been encouraged to search for *to* or *so that* used to show the purpose of doing something.

Unlike other students, the one who looked for the verb form occurring between *never* and *before* (…never + Verb + before) got the correct answer by typing in two words: *never + before*. This search led them to observe the pattern *never + past participle + before* and they used it correctly in their writing.

Based on this table, it is likely that successful searches for grammatical-word class colligation are those that search for grammatical words only and that interpreting concordance lines for this information is quite straightforward for the students.

Table 5.7: Collocation

Search term	Type	Search	Type
Verb + Noun			
order	Verb + order (Outlining one of the important steps, it shows the order of the main ideas for your paper.)	No conclusion	show the order of the (main) ideas

Search term	Type	Search	Type
idea (S22)	Verb + idea (write a lot of ideas)	No conclusion	organize ideas
benefit (S22)	Verb + benefit	No conclusion	have benefit (Word not found in writing)
organize (S22)	organize + ideas	No conclusion (then organize the order and the groups of them)	organize ideas
search	search (the information from) the internet	search the Internet	search information from the internet
organize (S36)	organize a research paper	Acceptable	organize research paper
knowledge (S40)	Verb + knowledge	No conclusion	apply knowledge
finishing course	finish + course	No hits	finishing course
finishing	finish + course	finish course (after finishing this course)	finish + course
Adj +Noun			
mistake point	mistake point	No, no hits	weak/incorrect point, mistake
miss point	miss + point	No hits	weak/incorrect point, mistake
mistaking point	mistaking + point	No hits	weak/incorrect point, mistake
mistaked point	mistaked + point	No hits	weak/incorrect point, mistake
point	ADJ/Noun + point	No conclusion	incorrect/weak points, mistakes
referencing (S34:1)	referencing + format	Not found so not acceptable	reference format
academic language	academic + language	academic language	academic language
academic field	academic + field	academic field	academic field
major subject	major + subject	major + subject	major subject
filed point	f[a]iled + point	No hits	weak/incorrect point, mistake
correct format	correct + format	correct format	correct format
obvious difference	obvious + difference	obvious difference	obvious difference
Noun+Noun			
bunch (S33)	bunch + of + knowledge	Not found together so cannot be used together	I have gained (some) knowledge which is…

Search term	Type	Search	Type
knowledge (S33)	bunch + of + knowledge	Not found, so I use some knowledge	(Some) knowledge
writing technique	writing + technique	writing technique	Writing technique
format (S34)	reference/referencing + format	Not found, so not acceptable	reference format
referencing format	referencing + format	No hits	reference format
Verb + Adv			
write (S39)	write + smoothly	write very well	write well
think freely	think + freely	No hits	think freely

When looking for collocation, it was found that the students looked for collocation of Verb + Noun, Adjective + Noun, Noun + Noun, and Verb + Adverb. See table 5.7. In looking for collocation of two words, the students basically tried two different ways. First, when they did not know the collocates of the target word and wanted to find out what words tend to collocate with the keyword, they typed in the keyword only. For example, they typed in *order, idea, benefit,* and *knowledge* to find out what verbs collocate with these nouns. One student typed in the noun *point* to look for the preceding adjective that means *'weak',* which is the target meaning. Second, when they hypothesised collocation or had the collocate of the target word in mind, they typed in two words, both the keyword and its hypothesised collocate. For example, they formed these queries: *mistake point, miss point, mistaking point, mistaked point, finishing course, academic language, academic field, major subject, filed [failed] point, correct format, obvious difference, writing technique, referencing format, think freely* to find out if they occur in the corpus. Sometimes, the students checked the hypothesised collocation by typing in the keyword only to find out if it occurred with the hypothesised word: *write* (*write + smoothly), knowledge (bunch of + knowledge),* and *format (reference/referencing + format).* Sometimes, as opposed to this, they did it by typing in the hypothesised collocate only to check if it occurred with the target word

they used in their writing: *organize (+ idea), organize (+ research paper), search (+ the Internet), finishing (+ course), bunch (+ of knowledge), and referencing (+format).*

Both typing in one word and typing in two words, the target word and its hypothesised collocate, can sometimes work well, but more often these strategies do not work. The successful single-word searches are *search, organize (+ research paper), finishing,* and *write*. It is interesting to find that a single word search for collocation of these words is successful simply when the students typed in the query item only to see that it is used or occurs with the word or group of words they used in their writing or not. If they found any of the concordance lines that matched their hypothesis, they were confident that it is correct and used it in their writing. For example, they found *search + the Internet, finishing + courses,* and *organize + page* in the corpus, so they used *search the Internet, finishing course,* and *organize research paper*, which is thought to be similar to *organize + page* in the concordance line. On the other hand, typing in one word is not successful when the students asked a general question e.g. what verb goes with this noun? This is because they were unable to draw conclusions from the concordance lines. For example, they failed to find out what verbs they could use with the nouns *order, idea, benefit,* and *knowledge,* which adjectives to go with *point*, and what nouns go with *organize*.

Likewise, typing in both the keyword and the hypothesised collocate works when the strings are correct English and the students can see the presence of strings in the corpus without having to draw conclusions from the concordance lines: *academic language, academic field, major subject, correct format, obvious difference*, and *writing technique*. However, this method does not work well because there is a tendency that the students put in a query that is not a common collocation or not grammatically acceptable in terms of word forms, and the query leads to no hits, i.e.

mistake point, miss point, mistaking point, mistaked point, finishing course, filed [failed] point, referencing format, and *think freely*. Given that the strings *finishing + course* and *think + freely* are possible and correct collocations, their absence from the corpus misled the students into thinking that *finishing* and *course,* and *think* and *freely* do not collocate.

These findings suggest that it is relatively difficult for students to use a corpus to check word collocation in general. The students are not able to observe the collocation of words through an active investigation of the concordance lines. On the other hand, to some extent they can work this out only to prove their hypothesis about word collocation by looking for the evidence where the two specific words occur together in the concordance lines.

Table 5.8: Acceptability of strings

Query item	Type	Target item	Target form
at present	at + the + present or at + present	at present	at present
have benefits	have + benefits or have + the + benefits	have benefits	have benefits
branches of studies	Does it exist?	No hits	fields of study
into my research	into+my+research	No hits (I used all of these knowledge into my research paper.)	used…in my research, applied…into my research paper
native English speaker (S26)	native English speaker	native English speaker	native English speaker
into my	Into my research paper	No conclusion from examples	used…in my research paper
in the first place	in the first place	It is used in English	In the first place
weakness	Does it exist?	Yes	yes
knowledge of how to	Does it exist?	Yes	yes
is based on	Does it exist?	Yes	yes
either way	Does it exist? Does it mean both ways?	Yes	yes
write a stuff	Does it exist?	No hits	writing
write a text	Does it exist?	No hits	writing

Query item	Type	Target item	Target form
instructional channel	Does it exist (..they used video as a instruction channel.)	No hits	teaching aids, instructional media
instruction media	Does it exist?	No hits	teaching aids, instructional media
required subject	Does it exist?	Yes, 4 examples	required subject, compulsory course
what are they like	Is it acceptable to say 'what are they like' and can I say 'what are good topics like'?	(S21) 'what are they like' is present in the corpus, so I can say 'what are good topics like'.	what they are like (Noun clause: what good topics are like)
most useful	Is it acceptable?	yes	most useful
referencing (S34:2)	(S34) Does it exist?	Yes	reference format
"to be called"	Does it exist? (To be called a good paper, your paper must be strong and…)	No hits	as (As a good paper, your paper must be strong…)
be called	Does it exist?	Examples found but no conclusion	to be/as
be called a good	Does it exist?	Yes, 3 examples	to be/as
I quite love	Can I say this?	No hits	I quite love

As shown in table 5.8, the other reason for the students to use the corpus was to check the acceptability of words and strings of words when they were not quite sure whether the target words or strings exist or occur in the sequence as exactly typed in. In short, the students wanted to make sure if they could use those strings in their writing. If the strings were found in the corpus, the students thought that they were acceptable and correct. When checking the acceptability of the strings, the students mostly typed in 2-3 words. The two-word strings comprise *at present, have benefits, into my, either way, instructional channel, instruction media, required subject, most useful,* and *be called.* The three-word strings are composed of *branches of studies, native English speaker, is based on, write a stuff, write a text, "to be called",* and *I quite love.* Sometimes, the students typed in four-word strings i.e. *in the first place, knowledge of how*

to, be called a good, what are they like. In addition to checking the acceptability of strings, the students rarely checked the acceptability of a single word such as *weakness* and *referencing*.

As seen in this table, most of the searches the students made led to unsuccessful results. The main reason for this is because the students did not know the precise words or strings to convey the intended meaning, so they formed strings that are correct English, but the strings do not match the target ones and do not occur in the corpus. For example, they searched for *into my research, write a text,* and *instructional channel* while they are expected to use *in my research, writing,* and *teaching aids* or *instructional media* in their writing. As opposed to non-existing strings in the corpus used, the students found these strings in the corpus: *referencing, be called a good* and *what are they like* (comparable to *what are good topics like*), but they do not match the target ones: *reference, to be/as,* and *what good topics are like*. The searches on *into my* and *be called* were also present in the corpus, but the students were unable to interpret the concordance lines. The searches leading to success, present in the corpus, and found to match the students' target forms are *at present, have benefits, native English speaker, in the first place, weakness, knowledge of how to, is based on, either way, required subject, most useful.* It is obvious that some of these successful searches are fixed phrases i.e. *at present, in the first place, is based on* and common phrases or noun phrases like *either way, most useful, native English speaker,* and *required subject*. This suggests that checking whether the strings exist in the corpus is not helpful unless the students know exactly that the strings can be suitably used in their context of writing.

Table 5.9: Agreement

Query item	Student's hypothesis	Student's finding	Target form
another (S1)	another + sing or plural noun	another + sing noun	another + sing noun
other (S1)	other + sing or plural noun	Examples found but no conclusion	other +plural noun
the most important (S1)	the most important + singular or plural noun (thing)	The most important + sing noun	the most important + singular noun (the most important thing)
all of the (S3)	all of the + sing or plural noun	all of the+ plural noun	all of the + plural noun
all of the details (S3)	all of the details	all of the details	all of the details
every authors	every authors	No hits	every author
authors	every authors	No conclusion	every author
most skills	most skill or most skills	most skills	most skills
one of the easiest ways	one of the easiest way or ways	one of the easiest ways	one of the easiest ways
each part	each part or parts	each part	each part
a variety of	a variety of + sing or plural noun	a variety of+ plural noun	a variety of+ plural noun
vocabulary	many + vocabulary	Examples found but no conclusion	a lot of vocabulary
many vocabulary	many + vocabulary	No hits	a lot of vocabulary
lots of vocabulary	lots of + vocabulary	No hits	a lot of vocabulary
a lot of vocabulary	a lot of + vocabulary	a lot of + vocabulary	a lot of vocabulary

Table 5.9 indicates that the students also used the corpus to check agreement between determiners and nouns. The purpose of this is to find out if the specified quantifiers precede singular or plural nouns. In some cases, the students wanted to find out if it is acceptable to use a particular noun e.g. *vocabulary* with the quantifiers that they had in mind. In checking determiner-noun agreement, the students employed three different methods. One group of students typed in the quantifiers only such as *another, other, the most important, all of the,* and *a variety of* to see if these quantifiers are followed by singular or plural nouns. The other group of students typed in both the quantifiers and nouns to see if they occur together in the corpus: *all of*

the details, every authors, most skills, one of the easiest ways, each part, many vocabulary, lots of vocabulary, a lot of vocabulary. The last group of students typed in the nouns only: *authors* and *vocabulary*, to check if *authors* in the plural can be preceded by *every*, and if *vocabulary* can be used with *many*. Sometimes the students tried the same search more than once by putting in the quantifier only and by putting in both the quantifier and noun together such as *every authors* vs *authors* and *all of the* vs *all of the details*. In searching for *vocabulary*, the student used various methods to find out what quantifier is used with it.

As seen from the table, the students who were successful are those who typed in the quantifiers only: *another, the most important, all of the, a variety of,* and those who typed in both the quantifiers and nouns: *all of the details, most skills, one of the easiest ways, each part, a lot of vocabulary.* This means that typing in the quantifiers only and typing in both the quantifiers and the nouns that follow can work well. When the students typed in only the quantifiers, they saw the forms of nouns either in the singular or plural occurring after the query items and drew the correct conclusion. Similarly, when typing in both the quantifiers and nouns, the students saw these two elements occurring together in the corpus and concluded that the strings are correct. However, this strategy, typing in both the quantifiers and nouns, can be sometimes risky as the students may not get the hits if the strings that are correct English are not present in the corpus and the students misinterpret that they are wrong i.e. *lots of vocabulary*. Sometimes the students were not successful in obtaining concordance lines of the query items because they put in strings which are incorrect English like *every authors* and *many vocabulary* and got no hits. They then assumed that the strings are wrong or not acceptable just because they are not present in the corpus consulted, but they had no indication of what the correct ones would be. The students

who typed in only one word, either the quantifier to see what type of nouns come after it or the noun to highlight the preceding quantifiers, were also unsuccessful. Although this seems to be an effective way of checking determiner-noun agreement, the students were not able to process the corpus output. For example, they were unable to draw conclusions from the searches on *other, authors* and *vocabulary*. Therefore, failure to check determiner-noun agreement is mainly caused by incorrect input, which leads to no concordance lines, and an inability to interpret corpus data.

Table 5.10: Word class

Query item	Student's hypothesis	Student's finding	Target form
research	Can I use it as a verb?	It is mostly used as a noun, but can also be used as a verb.	Verb and noun, not found to be used as a verb in her writing
advantage	Its part of speech	Its part of speech (N)	advantageous (It is advantageous to students.)
complete (S36)	Its part of speech	part of speech (N)	Adj (a research paper is not complete if…)
part	Its part of speech	part of speech (N)	NOUN
worthwhile	Its part of speech	Part of speech (adj)	ADJ
beneficial	Its part of speech	Part of speech (adj)	beneficial Adj
researched information	Form –ed research or researched(adj) infornation	Error (no examples found)	research information (I learned how to organize my research information.)
researched	Its part of speech	part of speech (Adj)	research information

In looking up word class or part of speech of the target word, table 5.10 shows that the students typed in the target word only e.g. *research, advantage, complete*. Occasionally, a student typed in two words like *researched information* to check if *researched* in the –ed form is an adjective. One group of students was completely successful in looking up *part, worthwhile* and *beneficial*. They identified the correct word class of these words and used the words correctly in their writing. Some students were partially successful. They were able to identify the correct word

class of the target words but those words do not match the target form and were wrongly used in their writing. For example, one of the students looked for *advantage* and correctly found that it is a noun, but the target form that would make their sentence '*…it is advantage to students'* correct is *advantageous.* Similarly, the student who wrote *'I learned how to organize my researched information'* looked for *researched* to check if it is an adjective and found that it is. However, the target form that is well suited to their sentence is *'research' information.* The student who searched for *research* also found that it can be used as a verb as they had hypothesised, but the use of *research* as a verb was not found in their writing. The queries that lead to a complete failure are *researched information* as there were no hits obtained and *complete* as it was mistakenly interpreted as a noun. After trying another attempt by typing in *researched,* the student who had failed to do a search on *researched information* was able to find that it can be an adjective although they were not aware that it was not the correct word to use in their context. It is quite clear that when looking at word class, typing in one word often works better than typing in a string because there is a significant likelihood that the string will not occur in the corpus.

Table 5.11: Nouns in the plural

Query item	Student's question	Student's finding	Target form
sentence structures (S3)	Can I add –s after sentence structure?	Yes	sentence structures
sentence structure (S3)	Is sing form acceptable?	Yes	sentence structures
knowledges	Can I add –s after knowledge?	Yes	knowledge
different perspectives	Can I add –s after perspective?	Yes	different perspectives
essay	Can I add –s after essay?	No	essays

From table 5.11, it can be seen that the students were not sure if they could add –s after *sentence structure, knowledge, different perspective, and essay* and use the plural form of these nouns in their writing. Therefore, they sometimes searched for the plural forms: *knowledges* and *different perspectives*. Occasionally, they searched for the singular form only: *essay*. One student looked for both singular and plural: *sentence structure* and *sentence structures* to see if the singular form was also used in the examples.

The searches for the plural of these nouns which the students performed led to three different outcomes. Firstly, the searches led to success when the students typed in the target words in the plural, saw the plural forms, and drew the correct conclusion, i.e. *sentence structures* and *different perspectives*. Secondly, the searches led to failure when the student typed in the singular form of the words i.e. *sentence structure* and *essay*; therefore, they got no information about the presence or absence of a plural form. Thirdly, the search led to failure because of an oddity in the corpus i.e. *knowledges* which is unexpectedly found in the corpus of social science studies of epistemology, where *'knowledge'* is treated as a quantifiable set of entities, constructed in different contexts. In those studies, knowledge can be treated as something constructed (*created new knowledges, the production of knowledges, the construction of knowledges*) and as something divisible (*the construction between knowledges, the inter-relation between new scientific knowledges and …, an imaginative reworking of dominant knowledges*) and the word *'knowledge'* acquires a plural form.

As shown in this table, in checking if the nouns have the plural, some of the students did not enter the plural form of the nouns, so they did not know whether the plural form exists. Therefore, to do this effectively, the students needed to enter the input that matches the question by typing in

the plural form of nouns. It is also significant that the corpus used needs to be appropriate to the learners, not to include the material that is outside what the students need to know, for example, *knowledges*.

Table 5.12: Position

Query item	Student's question	Student's finding	Target form
etc	How to use it?	, etc.	, etc.
others	How to use *others* with *'s* to show that something belongs to someone?	others' + Noun	others' + Noun
hence	Can I use it at the beginning of the sentence?	yes	Hence,… Clause; hence, clause
previously	Can I use it to begin a sentence?	Yes, followed by a comma	Previously,
because of	Can I use it at the beginning of the sentence?	Yes, followed by a noun (Because of this class help me improve my writing skill.)	Because + clause

From this table, it can be seen that when the students wanted to check the position where the words appear in a sentence, they sometimes asked a question of how the target words are used in general and sometimes they hypothesised the position where the target words occur in the sentence. The students asked a general question when they searched for *etc* to find out how it is used in a sentence and for *others* to find out its possessive form: *others* + apostrophe *(-'s).* On the other hand, the students raised a hypothesis when they searched for *hence, previously,* and *because of* to check whether or not they could use these adverbs or transition words at the beginning of the sentence. In both cases, the students typed in only the keyword and noticed how it appears in the examples or if the word occurred in the position that matched their hypothesis. This is a successful strategy as the students were able to come up with the correct interpretation,

although the student who looked up *because of* was not aware that *because of* in their sentence is followed by a clause, not a noun, and that they should have replaced it with *because* to make their sentence correct. This suggests that the students were successful in using the corpus to check simple things that do not involve complicated interpretation of concordance lines.

Table 5.13: Lexical word + to

Query item	Student's hypothesis	Student's finding	Target form
process (S9)	process + to + Verb (the complex process to draft an outline)	process + to + Verb (*for there are fewer opportunities for the diligent to 'make work', the reverse process to easing*.)	process of + V-ing
procedure (S9)	procedure + to	procedure + to + Verb	procedure for + V-ing
afraid to	afraid + to	afraid + to (afraid to write essay)	afraid to + Verb
important	important + to	important + to + Verb	important thing to + Verb

Table 5.13 shows that, in looking for whether or not the target word needs to be followed by *to,* the students did it in two ways. In most cases, they generally typed in only the keyword: *process, procedure,* and *important*. In a rare case, the student typed in both the keyword and *to*: *afraid to*. Both of these two strategies work. When typing in one word: *important*, the student was able to draw the correct conclusion that it is followed by *to + Verb*. When typing in two words: *afraid to*, the student saw the concordance lines for *afraid to* and concluded that it is correct and is followed by a verb. Although, for effective corpus use, one might think that a good way to find this information is to type in the keyword only and observe if it is followed by *to + Verb*, this table shows that this strategy does not always work for this group of students and the difficulty lies in interpreting concordance lines as the students were basically unable to draw correct or

sensible conclusions from the concordance lines. For example, in searching for *procedure,* the student saw one concordance line: '*The procedure to find out.'*, and immediately concluded that *procedure + to + Verb* is correct. They did not try to understand how the words in the sequence relate to each other and were not aware that in this line the word *procedure* is the object of the verb *find out.* They also failed to notice the target form *procedure + for + V-ing* in the lines such as *'Libet set up a procedure for allowing subjects to report the time…'* and *'a sophisticated procedure for helping you solve difficult problems'*. Another unsuccessful search is the query *process*. The student wrote *'I have learned the complex process to draft an outline.'* and they conducted a search on *process* to confirm that *process + to + Verb* was correct. They correctly found that *process* is often followed by *of + Noun*, but it did not match their hypothesis, so they looked for the concordance lines where *process* is followed by *to + Verb*. They saw one example *'…for there are fewer opportunities for the diligent to 'make work', the reverse process to easing'* and confirmed that *process + to + Verb* was correct. Therefore, in testing the hypothesis that the target word is followed by *to* or not, success or failure does not depend on the input, but on how the students interpret the output.

Table 5.14: Form

Query item	Student's hypothesis	Student's finding	Target form
have learned	Is it correct to use *have + learned*?	have + learned	have + learned (past participle)
never bored	Is it correct to use *bored* with –ed?	Yes: is never bored	is never bored
a lots	a lot or a lots?	a lot	a lot (learn a lot)
grammatical	Should I use *grammar* or *grammatical* in *'It is necessary to use correctly sentence structure and grammatical.'*?	grammatical	grammar

From this table, it can be seen that the students hypothesised the correct form of the verb phrase *(have learned)*, adjective *(bored)*, adverb *(a lots)*, and noun *(grammar)*. In dealing with these hypotheses, the students raised two specific questions, which are more or less the same. First, they asked whether it is correct to say this: *have + learned* with –ed and *bored* with –ed as adjective. When the students searched for these in the corpus, they typed in two words: *have learned* and *never bored* to see if these two strings occur in the corpus. These query items led to the correct answers that both *have learned* and *never bored* are correct as they were found in the corpus although the student might not have been aware of the verb form and pattern in the perfect tense. A more general and productive way to find this is to search for *have* and look at the lines where *have* is followed by past participles. Likewise, when wanting to find out if *bored* is a correct form of adjective, the students should have typed in *bored* instead of *never bored* and observe its occurrence after verb *BE* to avoid the possibility that the string *never bored* does not occur in the corpus. The other question that the students asked is whether it is correct to say X or Y. These students wanted to check if they should use *a lot* or *a lots*, or *grammatical* or *grammar* in their writing. They tried this by typing in only one option like *a lots* and *grammatical*. This method does not work very well as the question raised does not match the mechanism used. If the student wanted to find out if they should have used *a lot* or *a lots* in their sentence *'you can also learn a lots from those who have gone before you'*, they should have searched for both *a lot* and *a lots* to see if both strings occur and how, given that both queries appear in the corpus, they are differently used in the examples. Although the result in this table shows that this student was able to find that *a lot* is the correct answer, the think-aloud protocol reveals that they accidentally made an illogically correct interpretation that *a lot* is the right answer based on the concordance lines below.

pent only a few minutes under the lights and had	a lots	of praise and Bonios afterwards. They love their work! U
TROL , launched this month, will be packed with	a lots	of ideas and activities for you and your Patrol. Regular fea
? It should n't. W-- er I, I've had	a lots	of problems with soft boots, but that's cos I use
now, in financially and the the came on that gate	a lots	of times with us. They were on that gate with us
d its started off oh [pause] I know when I've got	a lots	of, if I do one lot tonight you'll get it
. Was it? Was he erm in the He was into	a lots	of things, he was er in the sea scouts, he
e lots of shops about it's er sports shoes . There	a lots	lots of shops about that spell sell sports shoes Yeah And

From these concordance lines, which sound like incorrect spoken English, they noticed that *a lots* is followed by *of*, so they thought that when *a lot* follows a verb like *learn a lot*, it does not need –s at the end and she wrote *'…learn a lot…'* without looking for *a lot* in the corpus. Again, this query raises an issue of a corpus problem where inappropriate English is found and learners are not able to make sense of this. Similarly, the student who wrote *'It is necessary to use correctly sentence structure and grammatical.'* might not recognise whether *grammatical* or *grammar* is a noun and failed to find out which one is correct to use. This is because they searched for *grammatical* only and got no information about *grammar* and about how these two lemmas appear in the sentences. Therefore, this student's lack of success lies with the input which is incomplete or insufficient for obtaining a complete set of concordance lines to draw conclusions from. In this case, if the student was not aware of the different word class, a more effective search would be to search for both *grammar* and *grammatical* and observe how they are different in terms of usage.

5.4 Conclusion

This chapter has discussed types of lexicogrammatical errors the students were trying to check from the corpus in order to correct errors in their writing: colligation, collocation, acceptability of strings, determiner-noun agreement, word class, nouns in the plural, position, lexical word + to,

and word forms. In seeking to learn about these features from the corpus, the students employed three main strategies: a) typing in one word, b) typing in two words, and c) typing in three to five words. The results suggest that the searches that lead to the most successful output is typing in one or a maximum of two words. However, what determines success or failure is not the form of the search input, but how students interpret what they find in the corpus. Therefore, in addition to typing in an effective query item, the key thing that makes students succeed or fail in using a corpus is the ways in which they deal with the concordance output. Pedagogically, the results of this study confirm that training students to find the right string and to interpret concordance lines is of vital importance to promote an effective use of corpora for language investigation. The findings in this chapter suggest that students need to be linguistically intelligent enough to make effective queries that match their hypotheses and they need to be aware of varied phraseology the word can have when interpreting concordancing output. For example, they need to know that *apply + for* and *apply + to* are followed by a different group of nouns (e.g. *apply for a job, apply to the university*) and that the prepositions *for* and *to* cannot be used interchangeably. The next chapter will discuss key points about what the students in this study have done in dealing with corpus consultation and how learners should be trained to use a corpus more effectively for language learning.

Chapter 6

Searching and Interpreting Search Results

6.1 Introduction

The previous chapter outlined the kinds of linguistic information the students look for in a corpus while composing a writing task. In this chapter, I discuss the approaches to conducting searches and interpreting concordance output adopted by the students while searching the corpus for linguistic information for self-correction. The research question posed in this chapter is: what do the students do when they perform a linguistic investigation using the corpus?

This chapter comprises three main sections. The first section details the method of the study. The second section is about the number of searches the students conducted in total. The third section reports what the students did and what they found while searching and interpreting output from the corpus. The last section is the conclusion of the chapter.

The main question raised in this chapter is how the students conducted corpus searches and interpreted the information gained from a corpus and what should be done to help them use a corpus more effectively. To simplify the results of an overview of nearly 700 individual searches, I have selected eight specific points of interest. Examples of the students' searches and interpretation of the search results which comprehensively account for all searches are offered under each heading. The purpose of this is to gain a more in-depth understanding of how the students use corpus resources and to evaluate to what extent the strategies they employed were successful. In addition, I intend to initiate a discussion of what help students need in order to

make them become more productive corpus users. The findings show that it is quite a challenging task for this group of students to use corpus resources as a reference tool for language learning and self-editing.

6.2 Method

This section describes how the data for this study were collected and analysed. As described in Chapter 2, in this task, the students were asked to search a corpus to find linguistic information to correct their own language problems they had identified in their drafts of academic writing. All 38 students took part in this task and they were asked to record their use of corpus resources and their think-aloud protocols while interpreting the results as in doing tasks 1 and 3. However, during the analysis of the data obtained from the video recordings, it was found that two of the recordings were defective. In the first one by S2, there was no sound in the video and in the second one by S40, the picture repeatedly stayed frozen with the sound being played. As a result, the data from these two students were excluded from the analysis of this task. In total, the data gained from 36 students were used in this study. All the think-aloud recordings (≈ 19 hours) were transcribed to allow reliable interpretation.

In analysing the data, all videos provided by the 36 students involved in this task were examined carefully. The key data to be focused on from the videos, which comprise the think-aloud protocols, are the searches the students carried out and the interpretation they made based on the information gained while trying to make sense of the search results. To gather these data, first, all the searches or query items each individual student attempted were listed. At the same time, a record was compiled of the purpose of conducting each search – the information the student wanted to find out by undertaking the search – and the transcripts of the think-aloud protocols

(see Appendix 11) were examined to identify the ways each student had navigated the BNC software and interpreted the results of each search. In cases where it was not clearly stated in the think-aloud protocols what the students hoped to learn from the searches, I had to interpret this from what I had seen and heard from the videos. Based on the raw data, all the students are varied and the searches they have carried out are also varied. Therefore, it is not really possible to establish a defined taxonomy and to quantify the strategies they adopt or to identify objectively their levels of success. What I have done instead is to select representative examples of some observations about what the students have done while searching and interpreting the output. Then, sample searches that are thought to be particularly interesting have been chosen to make and illustrate specific points. The number of searches by each student and the observations made are given in the following sections.

6.3 Number of searches

The number of searches each individual student conducted in total is shown in the following table. See Appendix 9 for a list of searches each student did.

Table 6.1: The number of searches by each individual student

Student	Number of searches
S1	9
S3	37
S4	55
S5	14
S6	16
S8	10
S9	8
S10	21
S14	37
S15	7
S16	4
S17	17

Student	Number of searches
S19	7
S20	31
S21	38
S22	17
S23	13
S24	26
S26	17
S27	23
S28	24
S29	17
S30	15
S31	13
S32	11
S33	12
S34	42
S35	16
S36	12
S37	8
S38	12
S39	24
S41	9
S42	21
S43	30
S45	19
Total	**692**

As can be seen from table 6.1 above, the 36 students carried out 692 searches in total. It is apparent that the number of searches by individual students vary a great deal, from 4 to 55 searches, and based on the data shown in the videos, not all the searches led to satisfactory results or fruitful examples for the students to look at. In many cases, the search terms led to no results – no concordance lines were obtained. This calls for a calculation of the proportion of searches that got at least one concordance line and the ones that got no examples at all. It is useful to know this information as it reveals to what extent the students could do a search that leads to helpful

results and what they need to do to help them do better when searching a corpus for linguistic information. The results are in the table below.

Table 6.2: The proportion of searches that got the concordance lines and searches that got no concordance lines

Types of searches	Number of searches	Percentage
Searches getting concordance lines	541	78.18
Searches getting no concordance lines	151	21.82
Total	**692**	**100**

Table 6.2 shows that the proportion of searches that produced the concordance lines and the searches that produced no concordance lines is 541:151, or 78.18% and 21.82% respectively. In other words, the number of searches that generated results is about four times higher than the number of searches that produced no results. This suggests that on average one of the five searches these students did led to no results or no concordance lines (see a list of searches or search strings which returned no results in Appendix 10). This result shows that generally the students were able to devise search strings that led to concordance lines. However, at times, they were relatively unsuccessful in obtaining relevant information from the corpus, although it must be remembered that a null result – no concordance lines – may in some circumstances be informative. In particular, if the search string represents incorrect English (that is, it is a sequence not found in standard English) then one might expect that no concordance lines will be returned.

An analysis of these two sets of searches is carried out and discussed in section 4 below to establish how the students in this study use a corpus.

Observations on student use of corpus resources

Although the majority of the students in this study were able to form the search terms that return concordance lines from which they can investigate the language features, it is observed that these students have difficulties in using a corpus and interpreting results of the searches. Here are the main observations about their use of corpus resources.

6.4 Mechanical issues in student searches

As devising the search strings is one of the basic but technically challenging strategies for using corpus resources, it is intriguing to learn that quite a number of searches that can potentially occur in English led to unsuccessful search results – no concordance lines found – despite the fact that the corpus used was large enough to predict some lines. This raises the question of what types of search strings the students entered while searching the corpus, so that they ended up getting no search results. To deal with this issue, problematic searches which produced no results were examined and it was found that one obvious problem the students had is making mechanical mistakes, which include the following.

- **Spelling mistake**

Spelling can cause problems in using a corpus. This problem occurs when the students put in the wrong spelling of the lemma or string. Only one mistake of this kind was found when one student (S15) looked up *focus*, but they typed in the wrong spelling *focuss*, which returned no hits and they were not aware that it was wrong.

- **Wrong use of wildcards and tags**

The other kind of mechanical mistake is the incorrect use of wildcards and tags in the search terms. This problem occurs when the students form the strings with the wildcards or tags that technically do not conform to the BNC software. Therefore, the strings they put in do not match or make sense to the search criteria and the problem remains because the students fail to change their input accordingly as found in the following examples.

Problematic search	**Suggested search**
cause \ that | cause	*cause + that, that + cause*
effect_V	*effect_V*, {effect/V}*
effect_VV	*effect_VVB, effect_VVI*
serious + n	*serious _N**
make+scare	*make + scare*
*flash + back_V**	*flashback*
response / information	*(response | information), response + information*

- **Slow link up speed**

When a large group of corpus users attempt to download corpus data online at the same time through the same server, the link up speed can be slowed down. This, in turn, makes the students become impatient with the speed of the Internet, and they tend to give up the searches because it takes a very long time to download the data. For example, S35 looked up the string *the +*

number + of, which is in fact perfectly correct English; if they had persisted in that search, they would have come across eight examples in the BNC.

Though this mechanical issue is not the most interesting point to make, it can occur at any time and it irrationally prevents the students from gaining access to the concordance lines that exist in the BNC or other concordance software if they are not aware of it.

6.5 False negatives

In addition to not getting the results of the searches due to the mechanical issues, the students often find false negatives, thinking that a string which is actually acceptable and correct is wrong simply because it does not occur in the corpus consulted. They enter a string which would be considered correct English and just look for the occurrence of the string, but the string does not occur in the corpus. One example from my data is the string *doing hard work.* It has emerged that the search is too specific in terms of the sequence of individual word forms. For example, although the verb *DO* collocates with *hard work*, so that phrases such as *'they were required to do very hard work'* are present in the corpus consulted by the students, the precise string *'doing hard work'* does not occur. The students, therefore, think that their string is incorrect. What the student needs to do is to explore further by trying a more generalisable search on *hard work* and notice what verbs come before it. As expected, this student, after unsuccessfully searching for *doing + hard work* for the second attempt, tried the query *hard work* and discovered that it could be used with the verb *DO* as found in the example *'Like Miss if you don't do hard work you'll get a flat face!'.* Alternatively, if this student really wants to find out if the verb *do* collocates with *hard work*, she can try the string *do + hard work*, in which the wildcard (+) will permit intervening words, and they will see examples such as *'You do the hard work – I will look on.',*

'with those who do the hard work in other countries' and *'We do the hard work together during the week.'* in the BNC. Another query that this student should try is the string *do hard work* and they will find the following examples.

they were ready to undergo discipline and do hard work,

and he wants to do hard work on our behalf and on the countrys' behalf.

Like Miss if you don't do hard work you'll get a flat face!

How do you get a flat face if you don't do hard work?

More examples of searches that lead to false negatives are as follow.

people meet others

people meet other people

nothing comes for free

has been developed for a long time

may faint

teach easily

how they are going to get

the stuff that is needed

carry wrong messages

women use cosmetics

treat their skin

stored food

continuously improved

who worked all day

what have been mentioned

what was mentioned

couldn't have walked

without diseases

realistic movie

not all the students

recover fast

solemn problem

violent problem

A brief look at these examples suggests that they are acceptable. However, the students searching for these strings found no examples because the precise strings do not exist in the corpus consulted by the students, but not because they are wrong.

For example, the string *recover fast* was not found in the BNC, even though it might be considered correct English. It is, however, less common than the near-synonym *recover quickly* of which there are 12 examples in the BNC. In this case, the student would have been more successful if they had looked up *recover* as a lemma and looked at the adverb that comes after it from the examples or from a list of collocations. This suggested search is found to be effective when the result of further analysis shows that this student did in fact look up *recover* after having found no examples of *recover fast*. When they got the concordance lines for *recover*, they sorted the lines to look at the adverbs on the right of the keyword and noticed that *recover* collocates with *quickly* and *rapidly*. However, they observed that *quickly* is more frequent than *rapidly*, which occurs only once with *recover*, so they concluded that *recover quickly* is more acceptable.

As is well-known, however, judgments of acceptability can be difficult. One student searched for *solemn problem* and *violent problem*. Exploring more extensive search strings such as *solemn + problem, solemn + problems, solemn + problem,* and *solemn + problems* confirmed that *solemn* and *violent* do not co-occur with either *problem* or *problems* in the BNC, and a further search in the Bank of English led to the same result. However, a search in Google led to 12,000 hits for *solemn problem* and 33,700 hits for *violent problem,* suggesting that *solemn* and *violent* might be used with *problem,* though probably rarely in NS English. In this case, to find out from a corpus what adjective to use with *problem,* the students should have looked at the lemma *problem* and looked for the adjective that preceded it as an alternative to *solemn* or *violent*.

A more extreme example is the string *realistic movie,* which is not found in the corpus. A check with Google returned approximately 84,900 hits. This suggests that it is a correct collocation, and intuition confirms that it is acceptable. Its absence from the BNC seems to be due to the makeup of the corpus, which is not very current, rather than the acceptability of the string. This is clearly problematic if students are being encouraged to trust the information they find in a corpus.

A similar example is the string *stored food*. The student (S4) wrote *"When your excretory system is clean from not having any stinking stored foods in your intestinal lining, there will not be any bad smell come out from your body."* They wanted to check if *stored food* (noun phrase) was acceptable English, so they did a search for *stored food* and got the following two examples:

*Farmers who **stored food** from a good harvest were shot for hoarding.*

*She also **stored food** in there like a hamster, and sometimes, when least expected, her head would snake back inside her covering and the crunch of crisps or the slurp of a boiled sweet could be heard.*

Based on these examples, they said that *stored food* (adj. + N) did not exist in the corpus. Only *stored* as a verb followed by *food* as an object was found. Though they did not make any correction to their sentence, and it could be inferred from their protocol that they were wrongly advised by the corpus evidence that *stored food* (adj + noun) was not acceptable.

6.6 False positives

The previous section gave some examples of 'false negatives', where the absence of a string from the corpus would lead to the incorrect conclusion that the string is incorrect English. The study shows, however, that students also find false positives. One type of false positive is where a sequence of words is found in the corpus, and is assumed to match the student's own writing, but this similarity is incorrect. In the other case, they accidentally make a correct interpretation of the results despite the fact that the string occurs in a different context. In both cases, the students misunderstand what they are seeing but the interpretation is accidentally correct. The examples below illustrate this point.

An example of a false positive where the word sequence is found to match the student's writing is the string *with lots*. S4 wrote *"…a Thai Director-General of Public Health Department urged that eating foods frying with lots of oil continually can cause heart disease and obesity…"* and

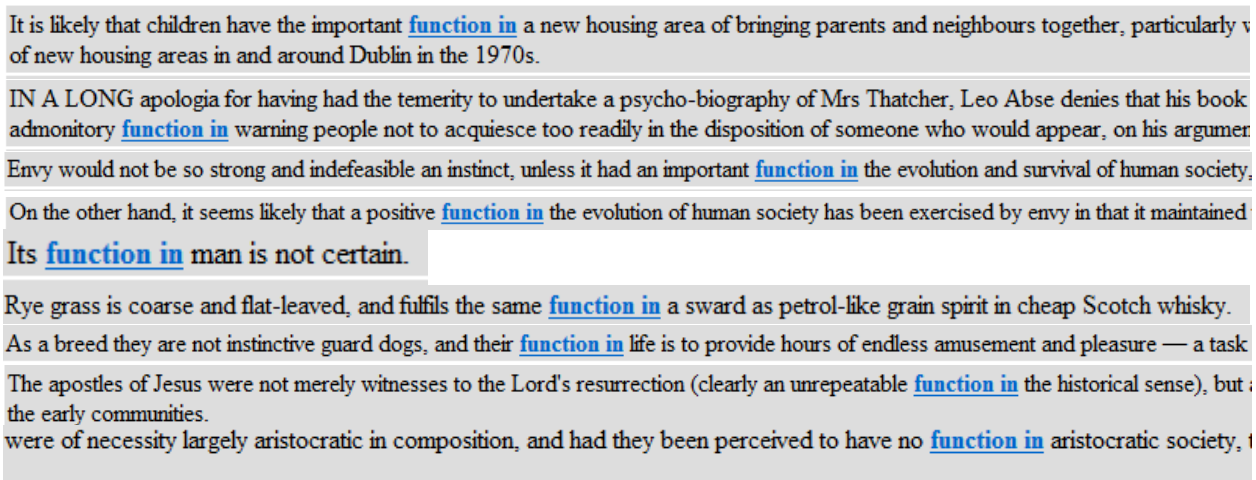
they wanted to check if *with* can be used with *lots* like *with lots of*. They made concordances for *with lots* and found the following concordance lines.

Smack-a-Mac! is an idea **with lots** of potential.
We apparently like large stores **with lots** of choice and a car park.
Incommunicado, which Hogarth saw fit to sing from the top of a heap of loudspeakers, blustered along **with lot**
It's still pretty difficult to understand why someone **with lots** of other opportunities (that BSc in psychology, the should be a porno model.
Johnson is about to explain **with lots** of big words but it's agreed he should shut up.
In addition to catching errors, object-oriented systems can shelter the user from the complexity of dealing **with l**
Backed up **with lots** of advertising aimed at upmarket buyers, the formula worked like magic.
Labour-intensive crops like maize and cotton are sometimes more efficiently produced by humble farmers **with**
Flexibility may need redefining, if armies are going to combine smallness with the punch needed to go on the offe
'Are you energetic **with lots** of ideas?
With lots of beer and home-made bread and cheese and ham and some of my mother's special cakes and — '
She is extremely well behaved, and travels everywhere with me — to open days where she has to cope **with lo** dog, friends' houses, talks I give, dog clubs and work engagements.
It may lack a certain flair, it may have a garish badge, it may still be cheap and cheerful inside (there are now so easy-to-drive family car **with lots** of equipment as standard and rock bottom prices.
A light touch leavens the proverbial 'riot of colour' **with lots** of fresh greens
'Make a fire **with lots** of hot embers.

They randomly looked at these lines and found the examples like *with lots of choice, with lots of ideas, with lots of fresh greens,* and *with lots of hot embers* and agreed that *with lots* is acceptable. This example shows that they did not realise while writing and forming the search string that *with* can be followed by and belongs with a noun (*with + Noun*) and that *lots of* is a quantifier that goes along with and belongs to a noun, *oil.* If they had been aware of this, they would not have formed the search string in this way and made the unsound interpretation that *with* could be used with *lots*. Instead, they would have checked if *lots of* could be used with *oil* by putting in *lots of* and seeing what kind of nouns comes after it. More specifically, they would have put in *lots of oil* to check its frequency in the corpus rather than exploring if *with* collocates

with *lots*. Therefore, forming a search string without prior or well-informed knowledge can lead to a wrong/invalid interpretation that the words in the string collocate with each other despite the fact that they do not belong to each other, but to the words that come before and after the string. In this case, *with* belongs to the verb *fry* and *lots* belongs to *oil (lots of oil)*.

An example of a false positive where the student accidentally made a correct interpretation of the concordance output, but the context was inapplicable to the student's writing is the string *function in*. S24 looked up *function* in the corpus to check if they used it correctly with *in* in their writing: *He performs his original function in harness most capably, carrying a light load at a moderate speed over great distance.* They found that *function* is often followed by *of* in the examples, and that *function in* is followed by V+ ing such as *function in warming people*. At this point they were not sure if *function in harness* in their writing was correct, so they put in the string *function in harness* in the corpus but found no examples. As they saw that *function in* is a possible string in a corpus, they entered this string again to check what type of words comes after *in* and found in the following lines, where the word *function* is a noun, that it was followed by a noun.


It is likely that children have the important function in a new housing area of bringing parents and neighbours together, particularly v
of new housing areas in and around Dublin in the 1970s.

IN A LONG apologia for having had the temerity to undertake a psycho-biography of Mrs Thatcher, Leo Abse denies that his book
admonitory function in warning people not to acquiesce too readily in the disposition of someone who would appear, on his argumen

Envy would not be so strong and indefeasible an instinct, unless it had an important function in the evolution and survival of human society,

On the other hand, it seems likely that a positive function in the evolution of human society has been exercised by envy in that it maintained

Its function in man is not certain.

Rye grass is coarse and flat-leaved, and fulfils the same function in a sward as petrol-like grain spirit in cheap Scotch whisky.

As a breed they are not instinctive guard dogs, and their function in life is to provide hours of endless amusement and pleasure — a task

The apostles of Jesus were not merely witnesses to the Lord's resurrection (clearly an unrepeatable function in the historical sense), but
the early communities.
were of necessity largely aristocratic in composition, and had they been perceived to have no function in aristocratic society,

They then put in *harness* to check if it is a noun, but the tag showed that the instance they were looking at was a verb. They sought help from an online dictionary and found that *harness* is a noun, so they confirmed that their writing was correct.

In this example, the student sees the correct form *function + in* but it is not the meaning that they need. However, they did not realise that the word *function* in their writing is not that kind of *function* which needs the preposition *in + Noun or V-ing* and that it is not that kind of *in* which is a grammatical word.

6.7 Students need to be aware of word class and meaning

There is a set of words in English that are identical in form, but differ in meaning and belong to different word classes. For example, the words *cost* and *sentence* can be either a noun or a verb, and the words *fine* and *patient* can be classed as nouns or adjectives. When learners are presented with this group of words, they need to be aware of the word class and the meaning that the word conveys. A lack of this awareness or failure to distinguish between word classes can lead to an interpretation that is correct in form or pattern, but incorrect in meaning and can cause confusion over word class. These phenomena are illustrated in the following sections.

6.7.1 Correct form but incorrect meaning

Even though forms and meanings of the language are closely related, a corpus is found to be more powerful in helping the students to notice language forms, in isolation from meanings. The students are able to induce the correct forms of the target word correctly, but they might not always notice the meaning it conveys. In other words, they recognize the patterns of the target word from the concordance lines and are able to use the word correctly in the sentence, but they

do not actually understand its meaning. As a result, the word is used in an odd and unnatural way. For example, S26 looked up the word *urge* in a corpus to check if they had used it correctly in their sentence: *Finally when the climate has been rising, it urges bacteria grow rapidly,…* and found the following concordance lines.

Leonard's first two books of poetry 'the thirties	urge	to relate the imagery of poetry to the world we live in
better, in fact, that he could no longer resist the	urge	to go and see Amanda. He rang the bell to her
he had taken me to a pub, and I felt this	urge	to have a hold on all the different sides there were to
ecognition. At risk of seeming a killjoy, I would	urge	readers to think through the implications of all home-made
apropos the Petrashevsky Circle, one feels an	urge	to smoke Dostoevsky out with the question, who's talking?
n their blind side: in other words that he had an	urge	towards crisis and clarity which he could only satisfy by yie
throughout these novels, that the true gambler's	urge	to lose is as strong as and not ultimately separable from its
of my dicta: Dostoevsky could only satisfy his	urge	towards crisis and clarity by surrendering it to the enemy. I
, the possessed state of the book, and feels an	urge	to thrust aside the irresolute self-contradictory narrator and
found 'chief character'. His first and continuing	urge	was to drive him towards crisis and clarity, which involved
a fresh turn. And a new twist is given to the	urge	to be a Napoleon. In the immediate context of The Posses
my analysis a surrender of the crisis-and-clarity	urge	, and an unloading of all the consecutions of theory upon ot
tantly to the actual facts of persecution, that we	urge	the support of Labour's leaders and membership by restor
ign Minister, has been calling upon the MPs to	urge	them to put the future of their country before any factional

From these concordance lines, they concluded that *urge* could be used in two ways. One is *urge + to + Verb* like *urge to go and see.* The other is to be followed by a noun + to + Verb as found in the following concordance lines.

> *At risk of seeming a killjoy, I would **urge** the readers to think through the implications*
>
> *has been calling upon the MPs to **urge** them to put the future of their country*

Though this discovery works in allowing this student to be able to correct their sentence to *…it urges bacteria to grow rapidly…*, it seems that they were not aware that *urge* is used in two different ways because it has different parts of speech. The pattern *urge + to + Verb* is correct

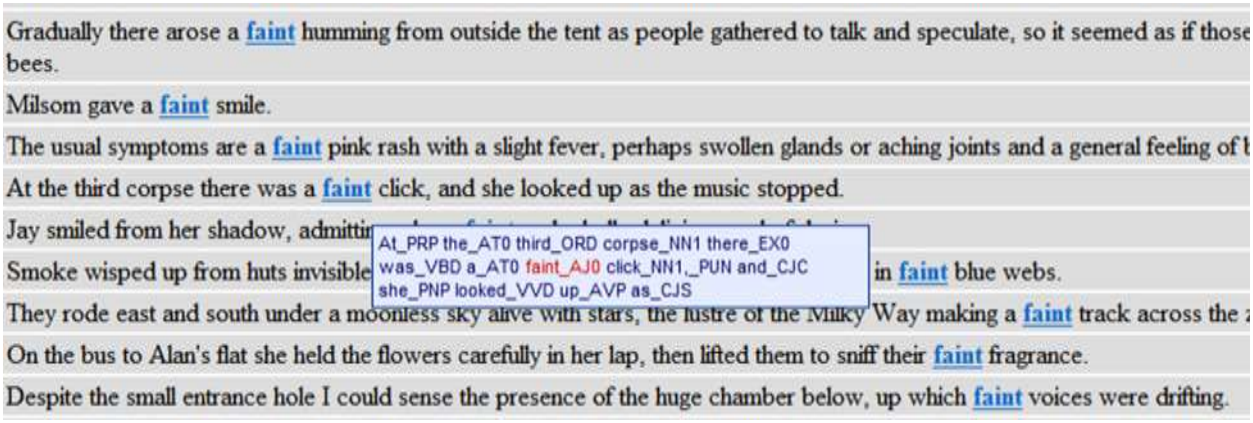
when *urge* is used as a noun like *the urge to go and see Amanda* and *the urge to be a Napoleon*. The pattern *urge + Noun + to + Verb* is true when it is a verb such as *I would urge readers to think through …* and *to urge them to put the future….* The other thing that this student did not notice is that even though they got the grammar of that point right, *urge* is still a wrong choice of word in this context. We normally urge somebody or something through speech and the addressee has to be able to understand our speech. If we look at the two examples above, we will find that the objects of *urge* when it is a verb are people: *readers* and *them*. Therefore, it is unnatural to write *it urges bacteria to grow rapidly* as bacteria would not understand human speech or gesture.

Considered to be successful in some way, this example of corpus investigation suggests that it is more feasible for this group of students to use a corpus to check aspects of language usage related to grammar or patterns of the words. In relation to meaning and word choice, it is not entirely helpful when looking at word use that goes beyond grammar as the students do not notice in what sense the target word like *urge* is used or which group of words the target word tends to collocate with. Although the students can work out its usage, they are not aware that they have made the wrong choice of word because they are looking at the examples in a very surface or linear way.

6.7.2 Confusion over word class

Word class can cause confusion. In interpreting concordance output, the students appear to have some confusion over word class, especially when they are dealing with the target words that have more than one part of speech as found in the following examples.

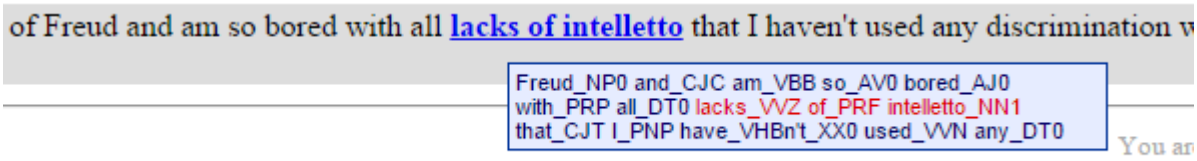
One example is the query *faint*. S3 wrote *...if you are playing football at noon which has high temperature, you may faint.* They searched for *faint* in the BNC and found the following examples.



While working out how *faint* is used in the examples, they looked at the part of speech tag to see if it *(faint)* is a verb and found that in all cases it was tagged as an adjective. Without realising that *faint* in the examples in the concordance lines and in their writing has different parts of speech and it is used in a different context, they concluded that their sentence was wrong and needed to be corrected as *...you may be faint,* which is incorrect in relation to meaning.

In this example, a lack of awareness of word class or different parts of speech the word can have hinders the student from interpreting corpus data more intelligently and effectively. When the target word has more than one part of speech, the student does not notice the different part of speech and does not realise that what they have discovered, which is correct in some way, does not correspond with their intended meaning. Thus, their sentence correction does not improve the sentence because it changes the intended meaning.

Another example of word class confusion is the query *lacks*. As S28 wrote *People live here so crowded that's why it lacks of infrastructure*, they made concordances for *lacks* to find out if it, as a verb, needed to be followed by *of*. They found one of the examples with *lacks* tagged as a verb and followed by *of*: *Have always been interested in intelligence, escaped the germy epoch of Freud and am so bored with all lacks of intelletto that I haven't used any discrimination when I have referred to 'em…* (see tagging below).



The tag *lacks_VVZ* led this student to conclude that *lacks* used as a verb is followed by *of* and that their writing was correct. In this case, they were misled by the wrong tag in the BNC in which *lacks* in this example is used as a countable noun, which is unusual, but wrongly tagged as a verb.

From these two examples, it is quite obvious that tagging can be a distraction, especially if students are not sufficiently linguistically competent. In the second example, it is even worse for the student (S28) when they come across and rely on the wrong tag in the corpus, which is a possible mistake in the process of tagging in a corpus. This student was not successful because they based their judgement about word class on the concordance line in which the keyword *lacks* is wrongly tagged as a verb, instead of a noun. Moreover, used as a noun in the example, the word *lacks* itself is a non-standard use of plural form of *lack*, which is actually an uncountable noun, and it occurs with a lot of non-standard elements like *have* and *am* with no subjects, and *germy* and *intelletto*, non-standard words which do not make any sense in terms of meaning to the

reader. This non-standard language use can cause another difficulty for the student, and it is sometimes unwise for the students to base their judgment on one line of examples like this. This has called into question whether it would be better for students to make their own judgement about word class.

6.8 Students need to parse a sentence

Being unable to understand or be aware of the grammar of a particular word when it is used in a concordance line or grammar of a sentence is one of the major causes that prevents the students from interpreting the concordance lines accurately. Sometimes, when checking if the target word needs to be followed by a preposition or not, the students focus only on the string and the preposition that occurs immediately after it. If the target word or the string is found to co-occur with a specified or other preposition, they make a hasty decision that it is right without careful thought about which word in the sentence the word belongs to. The following examples best describe this point.

S6 wrote *Before you adopt a pet, you have many things to consider*. They wondered if *before* needed to be followed by a gerund and if they needed to write *before adopting a pet*. They did a search for *before* and observed that it is followed by a noun like *before their arrival* and by an –*ing* form of a verb like *before returning, before leaving,* and *before being.* Based on the information that *before* can be followed by a noun, they thought that they did not need to change their sentence to *Before adopting…* as it was already correct. However, they felt that *you* was overused and decided to change their sentence to *Before adopting a pet,* to make it less wordy. Although their inductive reasoning that *before* could be followed by a noun or a gerund was right and they made a correct change to *Before adopting …,* again it is interesting to find that they were

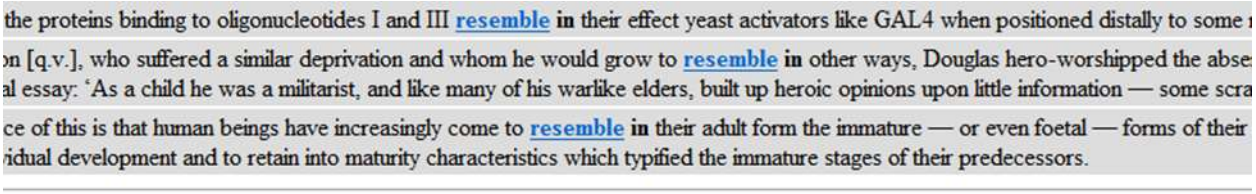
not entirely right. In their own sentence *Before you adopt a pet, ...,* the word *before* was followed by a clause, another use of *before*, which they were not aware of and did not find in the corpus. Taking *you* as a noun, they thought it was correct as it complied with the pattern *before + noun*. Though this is not actually wrong, it shows that they did not understand the sentence properly and they got it right by accident.

A more complicated search to deal with is the string *they face*. S21 wrote *"The word "mai pen rai" which means "It doesn't matter" is frequently used by Thais when they are face with situations involving conflicts."* They were not sure if they needed to say *when they are faced with* or *when they face with*. They first looked up *are faced* in the corpus, looked at some examples and assumed that it could be correct. Then, they did a search for *they face* and found that it is also possible, but none in the examples are followed by *with.* To prove if *they face with* is also correct, they used the sort function to check if *they face* occurs with *with.* Once they found one example as follows, they concluded that both *they are faced with* and *they face with* are acceptable.

> *for the Environment as environmentally benign, certainly over the problems that* ***they face with*** *the new standard spending assessment announced in the past few days*

In this case, the corpus data wrongly tells the student that *they face with* in their writing is correct because this student did not read the concordance line carefully enough to notice that *they* is the subject and *face* is the finite verb of the relative clause, and that *with* belongs to *the problems*, not *face*. They only looked at the key words without paying attention to the structure in which they occur.

The other example that needs the students' effort to understand how the target word is used is a search on *resemble*. S42 wanted to know how to use *resemble*. They thought that this word is used with a preposition, so they wanted to find out which preposition is usually used with it and if *in* was a common one as they wrote *…the spectators are expect to see how the movie resembles in a real situation, …* Herein, they meant the movie that resembles a real situation. They started off by putting in *resemble* and sorted its collocation. Once they saw *in* listed the 58th, they clicked to see the following three examples and concluded that *resemble* needs to be used with *in*, which is actually wrong when talking about something that resembles another thing.


the proteins binding to oligonucleotides I and III **resemble** **in** their effect yeast activators like GAL4 when positioned distally to some

ɔn [q.v.], who suffered a similar deprivation and whom he would grow to **resemble** **in** other ways, Douglas hero-worshipped the abse

al essay: 'As a child he was a militarist, and like many of his warlike elders, built up heroic opinions upon little information — some scra

ce of this is that human beings have increasingly come to **resemble** **in** their adult form the immature — or even foetal — forms of their

idual development and to retain into maturity characteristics which typified the immature stages of their predecessors.

Again, in this case, the student was not aware that in these concordance lines, the preposition *in* is not dependent on *resemble*, but it is part of the prepositional phrases *(in their effect yeast activators, in other ways,* and *in their adult form)* after the verb *resemble*. Also they did not observe the two patterns such as *X and Y resemble* (example 1), *someone resembles somebody* (example 2), and *something resembles something* (example 3). To come up with these patterns, the student, to a great extent, may need to read carefully in order to understand sentences which are quite complicated.

In this additional example where several searches were made, S43 wrote *"Children who play the online games may get strained or angry from the games because they want to win them."* They were not sure if they could say *get strained*, so they looked it up in the corpus to check if *strained*

could be used as adjective in this way. Having found no search results, they put in *get strain* but still got no results and they thought that *get strain* was not acceptable. They then did a search for *get strained* again and inferred that it was not a possible string as they had found no examples. To carry on, they looked for a single word *strained* and found the examples like *is still strained, looked so strained and tired.* They noticed that *strained* could be used as adjective after the verbs *be* and *look* in these examples but not *get strained.* They started to look for a noun form of *strained* by putting in *strain* in an online English-Thai dictionary and found its synonyms: *stress* and *tension*. They wanted to know if they could use *get tension* to convey their intended meaning. They typed in *get tension* and got the following concordance line.

> *if it's not prepared to support moves to try and* ***get tension*** *reduced between the two blocks? Well I think that we have….*

This example assured them that *get tension* was acceptable, but they were not quite sure if *tension* was the right word in their context. They looked up *get stress* as an alternative in the corpus but no concordance line was found. To make the search more general, they carried out a search on *stress* in the corpus using the collocation option to see the verbs that frequently go along with it. They found that the verb *cause* matches the noun *stress,* so they made concordances for *cause stress* to check if it occurs in the BNC. They found *can cause stress* in the examples so they said it fitted the meaning they wanted to convey better than *get tension,* so they changed *get strained* in their sentence from *"Children who play the online games may get strained …"* to *cause stress,* which is inferior in terms of meaning.

This student tried a series of searches but got the wrong information from the corpus as it wrongly indicates that the string *get tension* is a possible one. This student did not notice that, in the example, *get* and *tension* are not dependent on each other but they occur together in the span of three words: *get tension reduced*. The student saw *get* and *tension* occurring next to each other, so they made a quick judgement that *get* collocates with *tension*, without examining the left or right contexts of the string. In fact, they occur together in the pattern *get something done (get + N + V3)*, but the student did not notice it.

It can be seen from these examples that contexts to the right and left of the string are important for interpreting concordance lines. These students were not successful because they did not pay attention to the words in the right and left contexts and did not understand how the words in the string are related to the contexts. To interpret concordance lines more effectively, students need to be able to parse the sentence to see which of the words in the sentence the string or the keyword belongs to so that they will not make a quick judgement about what they are seeing and just accept it without enough careful thought.

6.9 Entering one word or shorter strings often works better than entering a specific long string.

As mentioned earlier at the beginning of the chapter, around one fifth of the searches the students conducted returned no examples of concordance lines, and one cause of this problem is the mechanical mistakes the students made such as spelling mistakes and incorrect use of wildcards. It is also found from the false negatives (see 6.5) that entering a specific long string often returns no search results.

Another type of string the students often searched for in a corpus that did not work well is a string that is considered to be grammatically incorrect English. This kind of string does not occur in the corpus of English because the words in the string are combined together without conforming to the rules of the language or because the grammatical words like a determiner or an article in between the words in the strings are missing. Some examples of grammatically incorrect strings the students formed are given below.

Incorrect string	**Correct string**
eat suit	*eat suitably*
when arrive at the party	*when arriving at the party*
get marry	*get married*
what was mention	*what was mentioned*
make you scare of marriage	*make you scared of marriage*
receive deposition	*receive a deposition*
everything include	*everything includes*
smell come	*smell comes*
carry wrong message	*carry a wrong message/wrong messages*
many condition	*many conditions*
punishment are soft	*punishment is soft/punishments are soft*
reduce feeling	*reduce + feeling/reduce the feeling*
decrease feeling	*decrease + feeling/decrease the feeling*

Most of these strings do not occur in the corpus because they are grammatically wrong. The strings *receive deposition, reduce feeling, and decrease feeling* do not sound wrong, but it seems

that the students did not get the search results because something is missing between them. My attempts to search for *receive + deposition* from the BNC (that is, permitting intervening words) resulted in one example of *receive a deposition*, and *reduce + feeling* returned two examples of *reduce the feeling* which is followed by *of* like *reduce the feeling of isolation* and *reduce anti-US feeling*, whereas *decrease + feeling* returned no results. This confirms that the words *receive* and *deposition* and *reduce* and *feeling* do collocate, but they occur together in a span of three words or more. To get these search results, the students may need to expand the searches by using the wildcard + between the words. Alternatively, if they are wondering if they can say *reduce* or *decrease the feeling* and want to find out what verb is supposed to be used with *feeling* to express that, they need to look at *feeling* in the corpus and look at the verbs that come with it.

It can be inferred that entering one word often works better than entering a string. This is because there is a high tendency that the strings the students make do not match any occurrence in the corpus as they are either too specific or grammatically incorrect. If a long string search does not work, the searcher needs to be intelligent enough to plan and try a new search and needs to think carefully if the search does not work before judging that the string is not acceptable or incorrect. The following is a good example of corpus searches to mediate unsatisfactory search results.

S35 wrote *"Human activities have recently increased the number of carbon dioxide, an additional greenhouse gas, which…".* They knew that *the number of* is used to indicate the quantity of something and wanted to find out what type of nouns goes with it. They ran concordances for *number + of* and found that it is followed by plural nouns such as *the number of troops*. They wanted to examine further if there were any instances where it is followed by an

uncountable noun but could not find any relevant examples as most of the results for the string *number + of* showed *number* and *of* occurring with other words in the span as the following.

greatest happiness of the greatest **number' of** people.
These shortcomings stem from the vagueness and dubiousness of the demand that a 'large **number' of** observations be
These show that every publication has a class **number consisting of** at least three digits, but the system permits further
The first **number there of** the two.
On each such line, there are on average some c*n possible middle points, where c is a **number independent of** n.
Fieldworkers used list A as a basis and the other two as a back-up resource; B and C were ordered in such a way that number 4 on list A, household **number 4 of** lists B and C would be of a similar type in the same locale.
What's, what's **number three of** profitability mean?
Erm all those I've mentioned before I think the **number of of** people who were convicted of petty offences, and maybe quite high.
A **number were of** the same place, some old village with nobody in it, the first a shot down a dusty road and the rest of fencing.
So, each size and style of font is given a unique **number out of** the 32767 available.
We expected that a number — perhaps an embarrassingly large **number — of** students would decide that they simply transfer.
He showed no surprise when Blanche told him the **number plate of** the car was false, and Nowak emphasised he did n
In addition to owning 'any **number' of** yachts, Cumming acquired six motor boats.

Not having found *number of + singular nouns*, they thought of running concordances for *the amount of* to compare the results. They typed in *amount + of* but did not look at the examples shown because they instantly changed their mind to search for the word *amount* instead. They used the sort option to specify the results occurring with *of* as shown below.

Hits 51 to 100 Page 2 / 58		
well known and popular with local residents and therefore a vast	**amount**	**of_PRF** acute work … is und
careful about how we entered this market. There's a significant	**amount**	**of_PRF** adaptation that has to
is just not feasible. In strictly layman's terms, the	**amount**	**of_PRF** additional engine-wei
yees, but what appears unusual from a Western perspective is the	**amount**	**of_PRF** administrative suppor
a 'wrap' Increased activity for several hours. Increases the	**amount**	**of_PRF** adrenalin in the body.

They saw *amount of acute work,* which is an uncountable noun, so it was clear to them that *amount of carbon dioxide* is better compared to *the number of carbon dioxide*.

It is interesting that this student searched the strings *number + of* and *amount + of,* and *amount.* In fact, the strings that would lead them to the answer quicker and more reliably is to put in *the number of* and *the amount of* and see what type of nouns comes after each of these strings. Alternatively, if they do not know whether they need to say *the number of carbon dioxide* or *the amount of carbon dioxide,* a good search would be *carbon dioxide* or *of carbon dioxide*, which provides more examples with *amount* in the BNC.

6.10 Knowing where the problem lies

To use a corpus more effectively for solving language problems in writing, it is not enough for students to be able to formulate search terms that will produce fruitful results and to interpret the concordance lines accurately. The students also need to be informed about where exactly the language problem lies in their writing. If the students are not able to find something that might be wrong in their writing, they probably cannot correct the sentence.

For example, S3 wrote *…, it can be bad and cause the health problem if you don't pay attention to the correct or proper ways to do.* They were not sure if the word *cause* was used correctly in their sentence, so they looked it up in the corpus and found *cause of, cause in*, and *cause for*. They looked further and found one concordance line which reads *…because a wing can so easily go down and cause a bad swing as you slow down after landing.* They then said that *cause* can be followed by a noun, but mostly it is followed by a preposition. Without showing that they had observed the difference between *cause + noun* and *cause + preposition*, they concluded that their sentence was correct. This example indicates that the student jumped to the conclusion that

cause could be followed either by a preposition or by a noun straightaway without realising that *cause* followed by a noun is a verb and *cause* followed by a preposition is a noun and that these two patterns cannot be used interchangeably.

Although this student knows that *cause* collocates with the noun *problem*, it seems that they do not know that the issue with their sentence is the noun phrase *the health problem*, which should be used as *a health problem* or *health problems* to refer to health problems in general rather than a specific health problem that might have been mentioned before. Therefore, in this case, the student could look up *cause* and see if it is followed by a noun phrase with or without a definite article or a noun phrase in a generic plural form when it is used to refer to that noun in general. All of this relies on the student being able to identify something that may be incorrect in their writing. If they could identify the correct problem, they could probably correct the sentence.

6.11 Conclusion

According to the findings to this chapter, it can be seen that the students go through three main stages in their corpus searches. The first stage is to identify the language problems in their writing. At this stage, the students need to be able to spot where the problem lies in the sentence and they need to decide if it can be solved by consulting a corpus. However, the findings show that the students sometimes are not able to identify their own errors. The second stage is to plan and enter the search to get concordance output for investigation. In addition to conducting a single word search, the students often put in strings that are either linguistically incorrect or too specific in terms of word sequence and obtain no concordance lines because they do not occur in the corpus. Very few students have technical problems conducting a search, but these problems are not so serious as to warrant immediate attention. The third stage is the interpretation of the

output from the corpus. The findings indicate that sometimes the interpretation, whether or not the search is found in the corpus, can be wrong. Wrong interpretations are caused by false negatives, inability to parse the sentence, and confusion over word class of the words with different parts of speech.

The next chapter will discuss the findings of the research and its pedagogical implications. .

Chapter 7

Implications for Learners and Teachers

7.1 Introduction

This study of learner use of corpora in self-editing aims to find out what learners actually do when encouraged to use corpus resources to identify and correct their own errors in writing. The study has addressed the following research questions.

1. What kind of lexicogrammatical errors do Thai learners of English find it easiest to solve by using a corpus?
2. When students are writing essays, what language points are they most likely to check in a corpus?
3. What do the students do when they perform a linguistic investigation using a corpus?

The three previous chapters showed the results in detail. Chapter 4 identified the kinds of lexicogrammatical errors Thai learners of English find it easiest to solve by using a corpus. To some extent, the results of this chapter are surprising as they suggest that it is difficult to predict the kind of lexicogrammatical errors this groups of learners can easily solve using the corpus. Whether or not the students will find it easy to solve the lexicogrammatical errors using a corpus largely depends on their familiarity with the language concepts. Chapter 5 reported the kinds of language points the students checked from the corpus when writing essays in English as well as the strategies they used when undertaking corpus searches for those features. Learners checked various kinds of language features from the corpus, and the features that they most frequently

checked were colligation, collocation, and the occurrence or acceptability of particular strings in the corpus. Chapter 6 dealt with the ways the students conducted searches and interpreted the results. Observations of what the students did when undertaking corpus searches which were successful and unsuccessful were made. The results of these three chapters can indicate to what extent Thai learners of English at this level are able use a corpus for error correction on their own and what needs to be done to prepare them to use a corpus more effectively.

This chapter discusses the key points found. It also discusses implications for pedagogy, from the point of view of the learner and the teacher.

7.2 Evaluations of learner use of corpora

In this study, the students used the corpus to correct two sets of errors – made-up errors in the error correction test and errors in their own writing. The error correction test was designed to investigate if the errors included therein would be easily solved. When the students used the corpus to correct the different kinds of errors in the test, the results were not entirely as expected. Although the students found simple collocation easy to identify, more abstract grammar patterns presented them with greater difficulty. Some clause structure features, expected to be difficult, appeared to be easier for the students to correct using the corpus resources. It appears that a key factor is the students' familiarity with given language concepts. They will find a familiar concept easy to investigate using a corpus. In contrast, an unfamiliar concept will be difficult to investigate, even when it is something that corpus researchers find very salient.

When the students carried out corpus searches to correct their own errors in writing, most of them were successful because correct usages they searched for were confirmed and incorrect usages

were corrected. In some cases, the students failed in using the corpus because they gave up before reaching a conclusion and sometimes they got the wrong information. An analysis of the queries they made and their interpretation of concordance lines reveals that, in many cases, the ways the students carried out the searches are not as expected. Some of them simply entered a string of words to check whether it occurred or not rather than making more sense from the corpus information.

Overall, it can be said that the students in this study were able use a corpus on their own because most of the complete searches led to useful conclusions. On the other hand, if individual problematic searches are taken into account, the results would be that the students are not satisfactorily successful. They still need a lot of help in planning more useful searches and interpreting the lines.

7.3 Some key points about learner use of corpora

These are four things I have found, and they are going to feed into my pedagogical implications.

1. Concepts of language are important.
2. A corpus can provide wrong information about language.
3. Students need to know about phraseology, for example, collocation and pattern, to avoid interpreting it as string matching.
4. Searching for one word is more likely to be successful than searching for a string.

7.3.1 Concepts of language are important.

As reported in chapter 4, it was unexpected to find that the students could look into the corpus and be able to use the corpus data to help correct errors concerning the structure of clauses such

as a subordinate clause, a relative clause, an if-clause, and a noun clause. Indeed, devising a corpus search that would lead to the discovery of this clause structure seems to be challenging for learners. A probe into the students' ways of thinking while trying to correct these clause structure errors suggests that they used their existing-knowledge, that is, the concepts of language they already had for planning and devising queries and interpreting corpus data. This implies that learners' existing knowledge, or the concepts of language they have, enables them to find relevant concordance lines quickly. In addition, if they have a concept of language relevant to the search being conducted, they are able to make correct conclusions about the searches or are able to notice even patterns which are not easily seen in the corpus. On the other hand, if the students are not aware of language or do not have that concept about language, they do not notice it even when the form such as *...to **suggest** that **he reflect** further on...* is salient to them. Therefore, in teaching language, teachers need to direct learners' attention to language form so that they can develop language awareness and noticing skills.

7.3.2 A corpus can provide wrong information about language.

A corpus can contain wrong information or incorrect language, so the students can be misguided by the wrong information found in the corpus. In this study, I came across two examples where the corpus can be either wrong or misleading. In the first example, the student carried out a search for *a lots*. The phrase *'a lots'* is of course incorrect English; the correct version would be *'a lot'*. Although it is not correct, the BNC corpus the students were using returned 7 hits. The student might have justifiably believed that *'a lots'* was correct. In the second example, the student conducted a search for *knowledges* to check if *knowledge* could be a plural noun, and found 22 examples where *knowledges* was in the plural form. In this case, *knowledges* is not

incorrect, but its use is restricted (see chapter 5). For a learner at this stage, the information is misleading, as the learner should use *knowledge* as a non-count noun, with no plural form. Therefore, it is important that students need to be made aware of this limitation of corpus data.

7.3.3 Students need to know about phraseology, for example, collocation and pattern, to avoid interpreting it as string matching.

Because the participants in my study are English-major students and they have been learning English for at least fifteen years in Thailand mostly through grammar-based instruction, especially during their first ten years of learning English where the rules are explicitly taught, it is assumed that they have already known grammar rules of English. However, they might not have much knowledge about other aspects of English such as phraseology of words which are not extensively emphasised and are not as explicitly taught as grammar. With regards to vocabulary and meaning, single words are generally taught in isolation from contexts where they are actually used. Although the students might have some concepts about phraseology, e.g. collocation and pattern, they are not often certain about word phraseology and the findings of this study confirm that the students need to know about this. This also implies that in teaching new vocabulary, a teacher should emphasise both its meaning that may vary according to the contexts and its form or pattern. This is to make the students aware of the pattern of the word and the different meanings the word can have. Once the students become conscious about word patterns and meanings, they can look for this information from a corpus themselves when learning about or using the target word.

7.3.4 Students did not fully understand 'phraseology' and tended to interpret it as string matching.

Students need to know that 'phraseology' refers to something more than simple strings. When they do not know this, they tend to conduct string searches, for example, *mistake point, mistaking point, finishing course, academic language, obvious difference, writing technique, think freely*, just to check whether the string occurs in the corpus instead of running concordances for the keywords *point* and *difference,* for example, and observing what adjectives are frequently used with them and which of those adjectives can be used to convey the target meaning.

There are four things that the students did not understand about phraseology, as follow.

1. There is a link between pattern and meaning.

The students were not aware that words in English have their own patterns of usage and did not perceive a link between pattern and meaning. This can be a problem when they look for word colligation information to see the occurrence of a lexical word with a grammatical word such as adjective + preposition. They were not aware that the target word can be used with different prepositions and that when the word takes a different preposition, the meaning changes. For example, in searching for the lemmas *learn* and *learnt* to check if the verb *learn* takes any preposition, S24 (see chapter 5) noticed that the verb *learn* is followed by *to*, which is, in fact, an infinitive marker, and *from*, so they changed their sentence from the correct sentence to the incorrect one by inserting *to* to the pattern *learn* + noun (learn + to + noun: learned to new skill) and changing *learn + about + how to* to *learn + from + how to*, which is wrong in terms of meaning conveyed. They did not understand the phraseology of *learn* and were not able to differentiate the change in meaning between *learn something*, *learn from,*

learn about, learn to, and *learn how to.* The student was not also aware of the functions of *to* as a preposition and an infinitive. They could observe something which is the correct form in the sentences that they saw but they did not get the correct meaning because what they wanted to say is *learn a new skill* and *learn (about) how to.* Other examples of the words that the students looked for from the corpus and can be followed by different prepositions are the queries *applying* (applying in vs applying to), *eduated* (educated at vs educated about), and *emphasis* (emphasise in vs emphasise + noun). A lack of awareness of the pattern-meaning link results in the use of words which are correct in form but incorrect in meaning.

2. Phraseology is variable.

There are words in English that tend to occur or be used with a particular group of lexical items. Some words are usually used with words which are negative in meaning *(commit, cause)* while other words tend to occur with words with a positive meaning *(provide).* When searching a corpus for the use or usage of the target word, the students did not observe the semantic association between the target word and a set of words that frequently occur with it, which is referred to as semantic prosody (Stubbs, 1996). For example, in investigating concordance lines for the verb *cause*, the student sees a group of nouns such as *death, disease, cancer, problems, argument, and allergies* being used with *cause*. However, the exact noun, such as *dispute*, that the student wants to use with the verb *cause* in their writing is not found in the corpus even though it would still be correct because it has a negative meaning. Thus the student is not certain whether the word they want to use can be used in that slot. In this case, *cause* has a negative semantic prosody because it often occurs with words with a negative meaning. Therefore, phraseology is not necessarily about form or exact form, but it is also about a set of things with shared features that may not occur in the

corpus. If the student writes *'The decision made by the committee may cause dispute among workers.'*, and finds in the corpus no instances of *cause* followed immediately by *dispute*, but does find examples where *cause* is followed by *argument*, the student should know that it is acceptable to say *cause dispute*. Therefore, what the students need to understand about phraseology is that it often occurs with a set of particular items and they need to know whether the word they want to use in their writing belongs to the set or not.

3. Phraseology is discontinuous.

One thing that makes it difficult for students to observe phraseology is that they do not understand that phraseology is discontinuous. For example, in looking for subject-verb collocation or verb-noun collocation, there might be something in between the subject and the verb or between the verb and noun that follows it. One of the examples of searches by the student that can illustrate this point is the string *receive deposition*. An expert user of English knows that there must be something missing in between *receive* and *deposition*, but in this case the student simply typed in *receive deposition*, and found that it is not in the corpus. My search for the string *receive + deposition*, allowing for intervening words, returned one example of *receive a deposition* with the article *a* in between. Another example of this is the search for *learn* to discover if it can be followed by a noun. When the student saw examples of *learn to*, they concluded that the verb *learn* must be followed by *to* + noun *(learn to new skills)*, which is wrong. In searching a corpus for studying phraseology, students need to understand the structure and be able to induce valid patterns even though the actual string may not occur in the corpus. When they look at the verb *learn* and see examples such as *learn a few tricks, learn a foreign language, learn how to manage, learn to avoid*, it is important that they know what it is about the search they are looking for. They need to know,

for example, that the article *a* in *a few tricks* is not important because it is not part of what they are seeing. What they have to see is the pattern *learn* + noun phrase or *learn* + *to* + verb and *learn* + *how to* + verb.

4. Phraseology can be about class.

Phraseology is something about word class and knowing about classes of words is important for interpreting concordance lines. If students see examples of the verb *advise* such as *advise you to book early* and *advise you on particular health issues*, they need to know which of the words in the patterns found is important. In these two examples, they need to be aware that the word *you* is completely relevant in the sense that there has to be a noun in that slot, though that noun is not necessarily *you*. It could be *him, John, my friend, any female competitor,* etc. To induce the patterns of *advise*, students have to understand the whole classes of words in its contexts. In the first example, the exact word *book* is completely unimportant because any verbs can be used in that position. However, *to* is particularly important because only the word *to* can be used there. Likewise, in the second example, *particular health issues* is not very important because the position can be occupied by any noun, noun phrase, gerund, or compound noun. The word *on*, on the other hand, is relevant and more important because it, in this case, can be only *on*. In another example of *suggest that the two poets resemble one another,* none of the word class is important except the word *that*, and even that can be omitted. What matters here is the fact that the verb *suggest* is followed by a clause and it is necessary that students observe this.

These are the things that students need to know about phraseology in order to conduct effective corpus searches, but as shown, these are the things that at least some of the students found it difficult to understand. A lack of knowledge and awareness of phraseology can result in

students' failure to reach the conclusions about language use. On the other hand, as found in the think-aloud protocols, existing knowledge about language is useful for the students for interpreting concordance lines as some of the students referred to it when carrying out corpus searches and it led them to successful searches. From the think aloud, the students did understand some basic language features such as subject-verb agreement, class of nouns that can be singular or plural, or class of nouns that can be only one. They also understood quite well about collocation and colligation, which is part of phraseology, and conducted a lot of searches on these. Although they were quite successful in doing so, they might not see that in the context of phraseology because this term is not widely known to most of them.

7.3.5 Searching for one word is more likely to be successful than searching for a string.

As shown in chapter 5, the students in this study most frequently used the corpus for checking colligation and collocation, and they searched for this in different ways. For example, if they wanted to look at verb-noun collocation, sometimes they just put in the verb and noun. Sometimes, they typed the verb only to see what nouns occur after it. It seems that the strategy that works better is just to put in the verb. When they put in two words, the strings can be not found either because they are wrong collocates or because the two words do not occur next to each other. This leads to a strong conclusion that the search for one word is more likely to be successful than searching for a string. This is especially true for colligation. However, it is also problematic to search for a single word which has many uses without limiting the output because the search can bring up a large amount of corpus data which is difficult to process.

7.4 Pedagogical implications

The previous section has demonstrated that there are three issues that can make learner use of a corpus go wrong. Firstly, the corpus used may contain inappropriate texts, such as those that use language that does not form part of the target language for the students. Secondly, students lack awareness of language. Thirdly, students lack skills to conduct searches and interpret concordance lines.

The first issue which is a corpus problem is larger than can be solved by teachers and students. However, if we believe that using a corpus is useful for learners, something could be done to change the situation. One thing that providers of corpora could do to make life easier is to provide a corpus that is suitable for learners. It might be helpful if learners are provided with corpora that have in them texts that learners like to read and the language is not too difficult for them to understand. If learners read concordance lines that are retrieved from texts of their own interest, they would find concordancing enjoyable. An alternative solution could be to train students to read in a different way, i.e. 'vertically' rather than 'horizontally'. Tognini-Bonelli (2001) points out that concordance lines present the pattern of language vertically, so learners investigating concordances need to read the lines vertically, rather than horizontally. Since reading concordance lines is different from reading books which are read horizontally, it is important to teach learners to read concordance lines vertically.

Learner preparation is the key to success in classroom concordancing. Boulton (2008a) suggests that learners at a low level can gain benefit from corpus data if they are properly prepared. One of the effective ways to prepare learners to learn in a DDL situation is to start with prepared

paper-based materials. By using paper-based materials, the teacher can select the lines suitable to the learners' level. The purpose of this is to get learners familiar with corpus data which are new to them. When learners are familiar with concordance lines, they can be gradually introduced to a new way of inductive learning, to induce or notice language patterns from concordance lines. Once learners are well-equipped with the skills necessary for coping with corpus data, they can be trained to have hands-on experience with concordancing software. Boulton (2008a) argues that it is not a good idea to introduce learners to using corpus software at the beginning because they are new to corpus data, new to the learning style, and new to using corpus software.

The other two issues are relevant to learners and teachers and need further discussion.

7.4.1 Learners/independent (learners and using a corpus independently)

As pointed out by Johns (1991) and Bernardini (2000), learners become more independent when they use a corpus to check their immediate language needs. Thus, the true benefit of this method of corpus consultation is that it fosters learner independence and helps learners develop skills in learning how to learn. In using a corpus to correct their own errors in writing, learners will learn to identify errors, plan corpus searches, and correct the errors by themselves. Hopefully, doing this kind of corpus work not only makes the students' writing better, but it also helps solve their immediate language problems. By investigating corpus data regularly to discover facts about language use and usage by themselves, it is also hoped that students will gain some gradual awareness of English as a whole.

On the other hand, it can be argued that it is too much hard work for students to use a corpus to solve their immediate lexicogrammatical errors, and looking into a corpus may not be an effective way of doing that because students can probably do it better by using other resources

familiar to them such as a dictionary or grammar book. However, the virtue of investigating a corpus is that students will gain more understanding of the phraseology of English, rather than purely solving their immediate language problems. With the availability of the World Wide Web, there might be a question as to whether students need to look into a corpus, rather than simply 'googling' their query. The answer may be that Google might not be a good source of language reference. As I have demonstrated in chapter 6, my searches for *solemn problem* and *violent problem* in Google return thousands of hits where *solemn* and *violent* are used with *problem*, but most of these examples sound odd from the native speaker's view. Many of the examples are quotations from old-fashioned books or their writers are probably second language users of English. On the other hand, one good thing about having learners use a corpus of native speakers of English such as the BNC is that they have access to standard English by native speakers rather than a mix of standard and non-standard English as found in Google.

While the ideas discussed above presuppose that learners will become independent if they use a corpus to help correct their written errors, a practical question posed by this is whether learners at this level are able to use a corpus independently. Based on the results from most of the previous studies, the answer might be 'Yes, they can'. However, the results of my study suggest that it can be challenging for this group of learners to use a corpus independently, and I am not going to be too optimistic about this. Being able to use a corpus independently does not literally mean learners are able to technically use the corpus software to access corpus data and make concordance lines, but it is much more to do with how to input queries that lead to useful data and how to process the corpus data. Therefore, my answer to this question is 'Yes, IF they are

well-prepared.' The following section discusses what teachers should do to prepare students to use a corpus more productively and effectively.

7.4.2 What use can teachers make of this research?

A question that arises from previous research on corpus use for language pedagogy is why are corpora not more routinely used by teachers? Most of the researchers agree that using a corpus is a good thing. In principle, scholars such as Sylvia Bernardini and Tim Johns, who are the pioneers of data-driven learning or discovery learning, agreed that if learners can use a corpus, they will be more independent, they will have a lot of language data available to them, and they will be able to find out more about language that they do not know before. In that case, why do teachers not encourage students to use it more than they do? One of the issues that might determine classroom use of language corpora is practicality. How can teachers fit corpus work in the classroom? If teachers try to have a class of 40 students do corpus work, they cannot have them all doing independent learning simultaneously at computers. However, partly from what I did when I made the students use the corpus for this study, I discover that the results are not wholly negative, but not wholly positive either. I will argue it is rather a heuristic suggestion that learners at the level I am talking about can simply go up and benefit from the corpus use. The participants in my study, to some extent, can represent learners of English in general. They are quite a high level of learners, studying English at a university. They may not be near the native speaker's competence, but in the world of English learning, they may be good. The results of this research, therefore, are useful for language teachers as they can be grounds for integrating corpora to language classroom and teaching students concordancing skills so that learners can use corpora on their own outside the classroom.

7.4.3 Preparing students to use corpora

The ultimate goal of learner concordancing is for the learners to be able to use a corpus independently to find out facts about language use. Johns says that it is impossible for teachers to teach every single thing about language to learners and learners should be trained to learn how to learn by themselves. One thing that Johns advocates about learner concordancing is they can gradually learn about language by themselves. However, as found in the literature review, learners need help in using a corpus effectively, especially in interpreting concordance lines. Unfortunately, it is not entirely clear how learners should be prepared to use a corpus and very few people suggest how to train learners to interpret concordance lines so that they will possess concordancing skills needed for effective corpus use. Sripicharn (2010) provides practical suggestions on how to prepare learners to use language corpora. His ideas include getting started by surveying the students' background knowledge about language corpora and concordancing as well as providing them general information about corpora, what learners can use corpora for, preparing corpus data by using ready-made corpora and creating their own corpora, formulating queries, familiarising learners with concordance software, and interpreting concordance lines. This information is genuinely useful for giving teachers guidelines on how to prepare learners to use a corpus in general. However, as the results of my study have shown, in using the corpus independently to check language use for error correction, the students did not encounter serious technical problems that need special and immediate attention. The practical difficulty lies in the ways learners formulate and type in the queries and more in the ways they deal with concordance output, or the ways they interpret concordance lines. This implies that, in training students to use language corpora, the teacher needs to prioritise what it is that students need to know or to master and to think about how they should be taught. My study suggests that the most important thing

that students need to be able to do and need help with is how to interpret the search results. Concordancing tasks aimed at guiding corpus users through the process of reading concordance lines are given by Sinclair (2003). However, his exercises may be too detailed and require too close attention to individual lines, making the process time-consuming and slow. In reality, when looking at concordance lines, a corpus user may read them quickly – scanning them in fact. A set of concordance lines that does not immediately suggest usable results may be discarded and replaced with another. Sinclair's book does not replicate this experience. Thus, the main issue raised by this study is how to train learners to use a corpus if we still believe that using a corpus for language learning is a good thing.

7.4.4 Raising students' language awareness

A revolution in language study by means of corpus-driven research has had some influence on language teaching. Most major monolingual dictionaries and some coursebooks are now based on corpus data. This also brings traditional views on language such as grammar and vocabulary into question and a new view of language which is phraseology, or the notion that each word has its certain grammar and meaning, has been proposed. In a phraseological view, meaning belongs to the whole phrase. Therefore, corpus research tends to prioritise words and their phraseology and meaning, etc. and language teaching tends to prioritise grammar and the structure of sentences. While phraseology is very important in corpus linguistics, there is a question of how learners who have been trained to view a language in a traditional way and have perceived that learning language is to learn its grammar and vocabulary should be prepared to use a corpus. What should teachers do to get them ready for corpus use before training them to use the corpus software?

As my study has shown, one thing that enabled the students to see what they wanted to find is their familiarity with the given language concept. If they have the concept in mind, they tend to find it in the corpus, no matter how hard it is. Most of the students in my study carried out a lot of searches for collocation and colligation and they were able to find the answers from the corpus because they had the concept of word collocation. The students knew that some lexical words need to be followed by a preposition. However, it seems that the students did not know much about phraseology. They were less certain of it and needed to develop language awareness. When an adjective or verb is followed by a different preposition, they need to think that it may have a different meaning, for example. They need to know what phraseology is and what it means.

How can teachers help students develop language awareness or knowledge of phraseology? Phraseology is something that can be seen in context, so an effective way of practicing noticing phraseology would be through reading. Students have to notice how words occur when they encounter the words in reading. As Schmidt (1990) points out, learning occurs and students learn more when they consciously notice language pattern. To help students develop noticing skill, the teacher may need to point out to them what is important about the way that word is being used. For example, when the students encounter the verb *suggest*, the teacher needs to point out that the thing that follows *suggest* is a that-clause and links it to other words that the students know that are also followed by that-clause such as *say, argue* etc. At the end, the students need to know a class of words that is followed by that-clause. The teacher can link this to a group of nouns or verbs which is followed by a to-infinitive clause such as *advise me to*, and *tell me to*. It is also important that teachers raise the problem of *suggest* which logically might

belong to the same class as *advise*, though in fact it does not. One way of raising this awareness is to do it as part of reading and as a result of reading, the teacher can bring out this meta-linguistic knowledge about the class of word. Tim Johns, the pioneer of DDL, said that what the students need to learn from a corpus is how to read, so what a corpus does is to make the students aware that there is phraseology going on (Hunston, personal communication). When they read, they are more likely to notice and acquire knowledge of phraseological use of words by themselves. Teachers cannot teach them everything and will never teach them everything, so they have to be able to learn by themselves from reading. When they read, they come across the word and they think that is how that word is being used: that is the phraseology of that word, and the way they can develop this skill is through corpus work. But all the corpus work is not teaching them specific items, it teaches them to notice the patterning of words, so the teacher can give them reading comprehension and this is the sort of things that the teacher might draw to the attention of the students. One way of doing that is to underline the word and describe how the word is being used, and link it to other words they know that are used in the same way.

7.4.5 Training students to interpret corpus data

The key things that make the students fail or succeed in using a corpus are being able to enter into the search something that will discover what they want to find, and being able to interpret the corpus output. These, then, can be used as a criterion for making a decision about how to teach students to use a corpus. In teaching students how to do this, the starting point may not be introducing them to the corpus software and teaching them about how to formulate search queries. Instead, the teacher can start off by showing students concordance lines and showing them questions that guide them to interpreting concordance lines. Before introducing students to

a corpus, the teacher should show them a sentence which contains a language problem that most of the students find it difficult to solve. Then, the teacher gives them a set of concordance lines relevant to the problems and helps them to interpret what they see. The following two lexicogrammatical errors I came across in my study can be good examples to demonstrate how teachers can do this. These examples have been chosen because they illustrate the points where the students 1) fail to notice the verb form and 2) are able to induce the correct form but not the right meaning. In both cases, the sentences involve the use of a verb and the to-infinitive clause.

The first example is taken from the error correction test. The students were asked to use the corpus to correct the sentence *Maude <u>suggested her to return</u> to New York*. The problem with this sentence (as noted above) can be articulated in two ways: either the form following the verb is incorrect, or, if the form (noun + to-infinitive) is retained then the verb needs to be replaced. The results (see chapter 4) show that most of the students failed to notice the pattern of *suggest* as the number of non-improvers is much greater than the number of improvers (82.35%: 14.71%) , indicating that this item is very difficult for most of these students to solve using the corpus. This is interesting and surprising because this item was expected to be easily solved and the students were expected to be able to find the complementation pattern of *suggest* in the corpus, but they were not able to find it. To enable the students to observe the pattern of *suggest*, the teacher can provide them a set of concordance lines for *suggest* set out in three columns with words to the left of the keyword *suggest* including the subject (column 1), the verb *suggest* (column 2), and that-clause (column 3) as shown in figure 7.1. This allows the students to see what comes after *suggest* and to see the pattern *suggest (that) someone (should) do something.* Then the teacher takes the word *advise* and gives some concordance lines for *advise* in the table with four columns

as shown in figure 7.2. The first column is about words in context to the left of the keyword *advise*. The second column is the word *advise*. The third column shows the object of *advise.* The fourth column shows the to-infinitive verb. By presenting the concordance lines in this way at the beginning, the students can see what comes after the keyword *advise* and this is also to encourage the students to notice the to-infinitive, which does not occur immediately after the verb *advise*. The teacher, then, ask the students to contrast *suggest* and *advise* and say what the difference between *suggest* and *advise* is (suggest (that) someone (should) do something vs advise someone to do something).

and I	suggest	you read some of his early short plays
She could ring Jamie and	suggest	that he write something for it.
was screwing up the courage to	suggest	she cook him a meal over the weekend
Peter did not	suggest	they meet again.
What do you	suggest	he should do?
I	suggest	you shouldn't use words you don't understand.
Matilda was going to	suggest	she wait for the good father's return.
it doesn't	suggest	that you should eat less
her FedPol friend, Chertro, to	suggest	that he contact Ardakke and offer his services.
she waited for him to	suggest	she should come up to Liverpool in the New Year.
He also had the cheek to	suggest	I should move into a flat nearer town.

Figure 7.1 Concordance lines for *suggest*

but we strongly	advise	you	to think about another career
I	advise	you	to piss off as soon as you can.
it was possible for me to	advise	him	to do something
be cruel and thoughtless to	advise	them	to look after their spiritual life
is sufficient for them to	advise	government	to seek ways of restricting
We	advise	readers	to check that all parts are still available
I	advise	you	to call upon them as soon as you may.
You might as well	advise	them	to win a gold medal
How would you	advise	me	to deal with this problem?
Professional people should not	advise	parents	to use physical punishments.

Figure 7.2 Concordance lines for *advise*

In the second example, I came across in the student's own writing, the student consulted the corpus in order to correct the sentence *'Finally, when the climate has been rising continually, it urges bacteria grow rapidly, and mosquitoes reproduce easily'*. The problem that the student was trying to solve is the pattern following *urge*. The results indicate that the student was able to identify the pattern *urge + noun + to + verb* and they corrected the problematic sentence to *…it urges bacteria to grow rapidly,…* In the corrected sentence, verb form is no longer a problem, but the problem is the meaning of the word. The form of the sentence is correct, but the problem is the mismatching collocation between *urge* and a non-human object.

In helping the students discover the use of *urge* that is appropriate in meaning, the teacher needs to decide whether it is more useful to show the students a noun or a verb *urge* because the class of *urge* can be a noun or a verb. If the students simply search for *urge* with no restriction, the

search results would be a mix of concordance lines for *urge* used as a noun and a verb. My advice would be to start with *urge* as a verb in order not to confuse them because what they want to do is to use the verb *urge.* My suggestion for this is to present the students two sets of concordance lines for the verb *urge* and for the alternative verb *encourage* in the table with four columns of subject, verb, object, and to-infinitive as shown in figures 7.3 and 7.4. The teacher then draws the students' attention to the to-infinitive in column 4 to remind them about the fact that both *urge* and *encourage* involve the to-infinitive. After that, the teacher asks whether they see a difference within column 3 which is the object of *urge* and *encourage*. This brings the students to the kind of nouns within column 3 for *urge* and for *encourage*. At the end, the students will realise that *urge* is followed by a human object and *encourage* can be followed either by a human object and a non-human object and that they should correct *'it urges bacteria to grow'* to *'it encourages bacteria to grow'*.

MPs can ask, and we	urge	them	to ask.
a killjoy, I would	urge	readers	to think through the implications of
these posters would I	urge	them	to ignore them as the show is still
He will	urge	people	to look after their hearts,
He said he would	urge	Mr Gorbachev	to end the arms supply
I simply	urge	you	to be reasonable and not go over the
I would	urge	readers	to think through the implications of all
So I	urge	anyone with a garden	to visit Sainsbury's
confidence, and will instead	urge	locals	to build the territory
we will run a campaign to	urge	all students	to boycott all banks

Figure 7.3 Concordance lines for *urge*

informed about our work and	encourage	people	to consider becoming volunteers
his humanity, and sought to	encourage	his followers	to do so too
You should	encourage	employees	to seek help voluntarily
rather than left to rot, or they will	encourage	pests and diseases	to build up in the
a pleasant cool temperature and	encourage	the yeast	to form a hard ring at the
different places as possible and	encourage	Him	to meet many more people
despite a brave attempt to	encourage	candidates	to stand at last week's
Winning Words' initiative, to	encourage	the tourism industry	to improve the service
were erected at selected sites to	encourage	the crows	to use them instead
Hard cheese isn't moist enough to	encourage	bugs	to grow, and tomatoes are rarely

Figure 7.4 Concordance lines for *encourage*

The follow-up exercise would be to take the word *urge* when it is a noun and has different phraseology because it is followed immediately by to-infinitive. However, the problem of doing that is that it has nothing to do with the verb *encourage*. What the teacher is trying to do is to show the students that *urge* is followed by a human noun and *encourage* is followed by either a human noun or noun. To do this, the teacher needs to think about how to present concordance lines to learners. One way would be simply to present concordance lines as they might appear on the screen. The problem of that might be it is difficult to guide the students towards what they are going to see. Therefore, a way of encouraging the students to look at what the teacher wants them to look at is to put the concordance lines in the table with columns as shown. This allows the teacher to draw the students' attention to column 4 where all is about to-infinitive.

When the students become more familiar with the idea of interpreting concordance lines, the teacher can give them a more demanding exercise for interpreting concordance lines. The teacher can make the task more demanding in a way that reflects independent corpus investigation by presenting the lines in a KWIC (keyword in context) format instead of putting them in columns. For example, in searching for *urge*, the students will obtain concordance lines for *urge* as a noun and *urge* as a verb. In interpreting these lines, the students need to be aware of the word class of *urge* and its phraseology. Otherwise, they will not be able to differentiate its phraseological use or use the word correctly. The teacher can do this by giving them a set of mixed concordance lines for *urge* used as a noun and verb (see figure 7.5) as they will see on the screen and asking them questions such as the following to enable to see the pattern of urge.

1. Is *urge* always followed by *to*?
2. When *urge* is followed by *to*, what kind of word is it?
3. Now, highlight in **green** when *urge* is followed by *to* and to highlight in **red** when it is not followed by *to*.
4. Look at *urge* followed by *to* and see what words always come before it.
5. When *urge* comes after the words *the* or *an*, what kind of word is it?
6. When *urge* is not followed by *to*, what words come before it and what words come after it?

After that, the teacher can give the students an exercise and ask them to fill in the gap with *urge* or *urge to* for practice. For example, the students can fill in the following sentences with *urge* or *urge to*.

1. He could no longer resist the _____ go and see Amanda.

2. I have this _____ show you my childhood stamp collection.
3. They _____ me to come to Paris.
4. Attlee was not content simply to _____ the need for economies.
5. Sir Rhodes _____ Mr Patten to reduce poll tax levels in London.

After this exercise, the teacher can ask the students to focus on the concordance lines where *urge* is used as a verb (lines 3, 5, 6, 7, 9) and asks them what they notice about the subject of the verb *urge*. This is to draw the students' conclusion that the object of *urge* is a human noun. At this point, the teacher can take the concordances for *encourage* (see figure 7.6), which is a verb very much like *urge*. The teachers can ask the following questions to guide the students through the use of *encourage*. The purpose of these questions is to enable the students to understand that the verb *encourage* can be followed by both human and non-human objects while the verb *urge* is followed by a human object only.

1. Is *encourage* used as a noun or verb?
2. Look at the subject of *encourage*. What kind of thing is it? Is it a person or thing, or both?
3. Look at the object of *encourage*. What kind of thing is it? Is it a person or thing, or both?
4. What conclusion can you draw about the use of *encourage*?
5. Is *encourage* used in the same way as the verb urge? If 'No', how are they different?

At this stage, the teacher can ask the students to correct the sentence *…it urges bacteria to grow rapidly,…* and explain that when we use it in a sentence, we use *encounter* rather than the verb *urge*. To familiarise the students with the ideas of concordancing, the teacher can provide them

different kinds of things they need to demonstrate before letting the students have hands-on experience with the corpus.

1	As Okely (1987: 67) observes, the	**urge**	to create publications is not always as
2	natural result was a strong peasant	**urge**	to supplement agricultural income by
3	At risk of seeming a killjoy, I would	**urge**	readers to think through the implications of
4	the Petrashevsky Circle, one feels an	**urge**	to smoke Dostoevsky out with the question
5	has been calling upon the MPs to	**urge**	them to put the future of their country
6	the ground, and up they come. So I	**urge**	anyone with a garden to visit Sainsbury's
7	supporters staged a rally last month to	**urge**	President Menem not to issue a pardon
8	the other portrays an unglamorous	**urge**	to stay alive. The heroic plot has Chubei
9	we are truly comfortable and	**urge**	you to try it.' The great dragon
10	also produce provocative insights. The	**urge**	to preserve these rural communities led

Figure 7.5 Concordance lines for *urge*

1	church informed about our work and	**encourage**	people to consider becoming volunteers
2	The canopy is light enough to	**encourage**	underplanting too, making
3	Floating mulches of spun fibre both	**encourage**	early crops and shield them from pests.
4	fruit trees or straw round their bases to	**encourage**	the insect. Earwigs are redeemed by
5	Use a proprietary bulb fibre which will	**encourage**	good root growth. • In some areas
6	tined fork or mechanical aerator to	**encourage**	vital air exchange at the roots.
7	But perhaps I should not	**encourage**	you. The business man can use you all
8	The instructor should	**encourage**	the pilot to talk through his thoughts aloud
9	co-operation they have met … and to	**encourage**	research by the police themselves'
10	But take courage, and try and	**encourage**	your children to talk about their feelings

Figure 7.6 Concordance lines for *encourage*

Exercises like these will guide learners through the process of interpreting concordance lines and applying what they find to error correction. Once students become familiar with reading and interpreting printed concordance lines, teachers can move them to a computer room where students can have hands-on experience with a concordancing program. At this stage, teachers may give them a set of sentences with the errors and discuss with them the problems needed to be solved. Teachers can lead the class to discuss about what searches to make that will be useful for the kind of errors. The purpose of this is to raise learners' awareness of the right kind of searches to make and to familiarise themselves with the software and the concordance output. It is necessary that learners become aware and informed of the nature of corpus and concordances so that they will not become frustrated if the search results are not as expected, as suggested by Kennedy and Miceli (2010). After students are able to interpret concordance lines and formulate

a query, teachers can allow them to use a corpus more independently. This idea of training learners to use a corpus which starts from interpreting concordance lines, conducting searches, and software training may work better than training beginning with using the concordance software as done in my study.

7.5 Conclusion

It is not entirely unexpected to learn from this study that learners find using a corpus surprisingly difficult. In this chapter, I have discussed why it is challenging for students to do corpus work. Nonetheless, enabling learners to use a corpus for their own language learning is not a hopeless task if teachers have the mindset that leads to the belief that learners can be trained to use a corpus. In order to make learner use of corpora feasible, what advice do we have to offer? One piece of advice that is greater than what teachers can address is we need the corpus that is appropriate for learners. Learners need a corpus that contains language suitable for their level and corpus software that is easy to use. Teachers, then, have three aspects of corpus use that are important and that can be practiced. These are finding a right string, interpreting concordance lines, and developing sensitivity to meaning, structure, and specific uses.

Chapter 8

Conclusion

8.1 Introduction

The aim of this thesis is to examine learner use of language corpora for self-correction of errors. The main focus of the research is on the process of carrying out searches and interpreting corpus data. The study has addressed three main research questions: 1) What kind of lexicogrammatical errors do Thai learners of English find it easiest to solve by using a corpus?, 2) When students are writing essays, what language points are they most likely to check in a corpus?, 3) What do the students do when they perform a linguistic investigation using a corpus? Answers to these questions have been reported in chapters 4-7. Thanks to an advance of technology that made possible an innovative way of collecting data for this research by means of think-aloud protocol, this research has raised the fourth research question: How effective are video-recorded think-aloud protocols in providing information about what learners are doing when they use a corpus? Can improvements to the procedure be proposed? These additional questions have not been answered yet and will be answered in this chapter.

This final chapter provides a conclusion to the study and discusses some of the points it raises. This chapter also raises limitations and gives recommendations for further studies.

8.2 Summary of the thesis

Chapter 1 provided the background and setting of the study I conducted in Thailand. It began with my inspiration for this research. I believe that language learners need exposure to real language and that mature learners should be trained to make use of rich and reliable sources of

language for language learning, especially for finding facts about language to serve their immediate linguistic needs and for correcting their own errors. This chapter also gave a general overview of corpus linguistics and briefly surveyed its influence on language pedagogy. The chapter ended with the research questions and thesis outline.

Chapter 2 justified the need for encouraging language learners to use a corpus and reviewed previous research on learner use of corpora. The most talked-about value of encouraging learners to learn from corpus data is that learners have an exposure to real language rather than the invented, intuition-based language often found in textbooks, that might not be used in the real world. Corpus evidence has also changed views on the nature of language itself, and corpus data have been fed into the production of many teaching materials, such as dictionaries and textbooks. By exposing learners to language in a corpus and training them to learn from a corpus, it is hoped that learners will learn about phraseology and lexico-grammar or other aspects of language that are not salient in other language sources and learners will be independent in their own learning. However, it seems that corpora are not widely used in the classroom and research that focus on learner use of corpora is limited. Most of the existing studies suggest that learners are able to use corpora and that they have positive attitudes towards using corpora for assisting their language learning.

Chapter 3 described the research methodology. In this study, the students received some training in how to use the BNCweb. At the data collection stage, they were asked to use the corpus to correct errors in the error correction test and to correct errors in their own writing. While they were doing these tasks, they were asked to think aloud and video-record their computer screen. These videos were used as data for the study.

Chapters 4-6 brought us to the results of the study and the key findings that will be discussed in the subsequent sections of this chapter. Chapter 7 addressed the major issues raised by the learners' use of concordancing and discussed implications for pedagogy. The key issue addressed by this study is that concepts of language and knowledge about phraseology are very important for investigating corpus data. This chapter suggested that one way teachers can develop learners' awareness of phraseology is through reading, and by pointing things out for learners to notice. Chapter 7 also suggested ways of training learners to use a corpus by starting from interpreting concordance lines in handouts to familiarise learners with the interpretation of concordance lines. When students become familiar with interpreting concordances, they can be trained to conduct searches themselves and to use the corpus software.

8.3 In this study, how did the learners make use of the corpus?

As concluded in chapter 6, the students in this study went through three stages in making use of the corpus. Stage 1 is to identify their own lexicogrammatical errors. Stage 2 is to plan and conduct the searches. Stage 3 is to interpret the search results. This section, therefore, discusses the errors the students were trying to solve by using a corpus, the searches they made, the effectiveness of their interpretation of the concordance output, and how well the students did in these stages.

8.3.1 What errors are easily solved?

As reported in chapter 4, the error correction test, designed to prove what types of errors can be easily solved by using a corpus, yielded a surprisingly unpredictable outcome. Sometimes, the students found it difficult to correct the errors that were expected to be easily solved by using a corpus, and vice versa. Most of the errors led to variable results. This means that it is difficult to

judge whether an individual item can be easily solved with the help of the corpus because it may be easy for some students but difficult others. Seven types of errors that this group of students can easily solve by using a corpus were identified; these are those concerning noun class, adjective pattern, subordinate clause structure, relative clause, if-clause, noun clause, and collocation. On the other hand, the complementation patterns associated with individual verbs (e.g. *suggest*) appear to be the kind of errors that the students find extremely difficult to solve using a corpus.

8.3.2 What searches do students make?

The kind of problems that the students experience in writing essays in English can be deduced from the kind of searches they undertake in order to solve those problems. Although it is not certain whether there are the problems the students actually have or the problems they perceive themselves as having, the searches conducted when the students were writing short essays in class indicate that the students, generally, have problems with the use of nouns and verbs as they search for these two word classes far more frequently than other word classes such as adjectives and adverbs. This is not surprising because nouns and verbs are classes of words that are used most to deliver the content of communication. To use nouns correctly and appropriately, students need to know a lot about their usage. For example, they need to know whether a noun is countable or non-countable and whether a singular or plural form is preferred. They also have to think about what determiner or quantifier to take. Some nouns need to be followed by a certain preposition, such as *interest in, attention to, focus on, knowledge of*. Students also need to know about collocations of nouns. Some nouns collocate more frequently with certain verbs and adjectives. For example, the word *losses* occurs more frequently with *heavy* (70 hits) than *big* (13

hits) in the BNC. There is a class of nouns that often occurs in the pattern the + noun + of + noun e.g. *the number of ways, the kind of things, the beginning of the year*. Students need to know what class the noun belongs to.

In using verbs, students need to know whether they are used transitively or intransitively or both. They also need to know about verb forms which vary according to the subject, tense, and sentence structure. Some verbs occur in only one pattern while other verbs are used with a variety of patterns. For example, the verb *advise* is used in the patterns: 'advise someone on something' and 'advise someone to do something'. All these things about verbs and nouns in English are important and not being aware of them can lead to problems. Sometimes learners know the concepts of collocation or of pattern, but they do not know the usage of specific nouns or verbs. Another word class that the students searched for is adjectives, but about half as frequently as nouns and verbs. A further analysis of the searches conducted as a result of this writing, as reported in chapter 5, reveals that the two features that the students most frequently looked for from the corpus are colligation and collocation. This confirms that the students conducted the searches for colligation and collocation information of verbs and nouns more frequently than other language features.

8.3.3 What strategies do students use for conducting searches?

In searching the corpus for examining the kinds of linguistic features they have problems with, the students employ different strategies. Most of the time, they type in only one word. Sometimes, they type in a string of between two and five words. Each of these strategies produces different search results. Typing in one word will produce many concordance lines, giving a large amount of information about the word, but may be difficulty for learners to

interpret. Occasionally, the students put in two different strings per search to compare their frequency. The risk of forming a string search is that the students sometimes do not obtain the output of the search either because they type in a grammatically incorrect string or because the string is correct and acceptable but does not occur verbatim in the corpus (see chapter 6). The absence of the string in the corpus often leads the students to put an interpretation on the search results that the string is not acceptable. Where a string actually is incorrect, and thus not found in the corpus, this is an accurate conclusion. In some cases, however, a perfectly correct string is not found in the corpus, and the student's deduction is in this case inaccurate.

It is generally true that the longer string, the less likely it is to occur in the corpus, even if it is correct. For this reason, it is suggested that a word search might be better than a long string search. In cases where the attempted string is not found in the corpus, the students need to try a different one and they need to be quite imaginative in forming the right kind of searches.

The use of these strategies may provoke classroom practitioners into thinking critically about how to teach students corpus skills. In practice, it is challenging to sit learners down with a corpus, simply teach them how to use a corpus and the software, and let them have hands-on experience with an expectation that they will be able to use a corpus effectively. If learners are trained to use a corpus in this way, they tend to use it just to check if the strings exist in a corpus, or to answer the question: Can I say this?, which is not fully useful. This kind of searches is especially unhelpful when the corpus data wrongly tell them that it is correct, for example, to say *"...if you are playing football at noon [which has high temperature], you may be faint."* because they see the word *"faint"*, which is actually an adjective in the concordance lines, occurs after the verb *be*. In another example, the corpus wrongly tells the students that the noun phrase *"stored*

food" is not acceptable because only the examples of *"stored food"* as Verb + Noun are shown in the corpus. This raises questions about what recommendations to make if teachers want the students to make better use of corpora. If we want students to be well-informed about the kind of information they can look for from a corpus, to be able to plan and devise a search wisely to obtain useful and relevant search results, and to be able to interpret the results accurately, how can we, as teachers, train, teach, or encourage them to use concordancing strategies that work. Teachers need to do something more than simply teaching the students concordancing techniques and letting them use a corpus on their own. It is difficult to teach students or ask them to use a corpus extensively for investigating aspects of language. As my study has shown, many things that the students have done do not work. It is the skills of searching and interpreting concordance output that do not work well, not the skills of using the software. The students need to develop the skill of knowing how to get the best information out of the corpus data they have. This is what the teachers need to do. Ideas about how learners should be trained have been discussed in chapter 7.

8.4 To what extent do learners interpret concordance lines effectively?

To what extent are learners of English, especially Thai students, able to use a corpus independently? My hope would be that learners would be able to use the corpus on their own outside the classroom, without the direct assistance of the teacher, when producing their own writing. My research suggests that, to some extent, they can use corpus resources to help them by themselves. However, when the students interpret concordance lines, there are things that can go wrong and things that work in that particular context.

When searching for concordance lines, sometimes the students are not successful in forming the searches in the sense that their search does not produce any concordance lines. Sometimes, when they obtain the concordance lines, they are unable to find the answer to the question they pose before conducting the search. Therefore, whether or not the students get the concordance lines of the search is not as important as how they interpret the search results.

8.5 To what extent does this research confirm or disconfirm previous research?

Most of the previous research on the use of corpora by learners related to this study is cautiously optimistic about learner use of corpora, suggesting that learners will be able to use a corpus to correct their own errors and to assist their writing in L2. Chamber's (2007) and Boulton's (2008b) surveys of previous DDL studies also suggest that the results are encouraging. One of the reasons for this might be that the participants in this group of studies are advanced postgraduate students who possess a high level of language proficiency. As my study has shown, the results of learner use of corpora independently for error correction are somewhat different, and I am less positive about learner use of corpora, especially for undergraduate students. On the whole, the participants in my study are able to use a corpus on their own, but their use is not always satisfactorily effective. As described earlier, they have problems in conducting searches and interpreting the results. They do not make optimal use of corpus resources, but tend to simply check if the string of words they have written is acceptable or not.

8.6 How well does think aloud work?

The focus of this experimental study is on the process of searching and interpreting concordance lines. To understand how the students conducted linguistic searches from a corpus and

interpreted the results, the students were asked to think aloud while undertaking concordance tasks at computers. Their think aloud and their work with concordancing software were video-recorded by themselves. In this study, think aloud protocols were the only way of accessing learner thought-processes. This has raised the fourth research questions: How effective are video-recorded think-aloud protocols in providing information about what learners are doing when they use a corpus? Can improvements to the procedure be proposed?

Think-aloud protocols are advantageous because they provide valid data which are not affected by the task difficulty (Guan et al: 2006, cited in Gray, 2015: 14). It might be anticipated that this data-collection method would be disruptive to the thinking process, and that many students would prefer not to think aloud when doing problem-solving activities. Therefore, it is worth investigating if this method of data collection is effective.

In my study, all the students completed the concordancing tasks at the same time in the computer room. They were seated in rows next to each other. Prior to the data collection stage, the students were trained to use the Camtasia Studio, screen recorder software, to record their concordancing work and their think-aloud. At the commencement of the data collection, they were given a set of headphones and microphones connected to the computers to record their voices.

Based on the analysis of their video-recorded think-aloud protocols, it may be concluded that for this group of students the method of data collection was reasonably effective. In terms of video quality, I greatly appreciated the technology that allowed me to gain access to all the students' thinking in a quick an easy way. Despite the fact that the students were sitting very close together and kept talking aloud at the same time, there was no noise disturbance from their peers

that might have had a negative effect on the quality of the think aloud protocols. Regarding individual students' ability to think aloud, I found that most of the students performed the task well and provided rich data for the study. However, some students appeared to keep silent when they were using the corpus. My deduction is that they might prefer not to think aloud as it could distract their thought process. This problem will be dealt with in the recommendation section.

My research shows that it is possible to use a think-aloud technique to probe students' thinking in real time for research. This computer-recorded think-aloud protocol analysis has some advantages over the traditional method, where a human researcher sits next to a research participant, observes them at work, and records and/or makes notes on what they are doing. First, with computers and appropriate software, this method can be carried out simultaneously with a large number of participants, so it helps save time. Second, the participants do not feel that they are controlled by the researcher, so they have more freedom and confidence to do the task. Third, the screen capture software can record the ways the students navigate the computer screens. This provides more concrete and more complete data than those obtained purely through human observation.

It is also true, however, that the think-aloud method may be unnatural for some participants and it is possible that they forget to think aloud when they become engrossed in their task. In this respect, the traditional method has an advantage over the computer-assisted approach used in this study, in that it allows the researcher who is monitoring the participant's thinking-aloud to prompt the individual participant to continue talking. The researcher can also probe into the way the participant is thinking to obtain relevant and useful data. However, this advantage could also be obtained through the use of computer-assisted think-aloud method by having the participants

work in pairs. They could take turns keeping their peers alert in a supportive way to prevent the thinkers from forgetting to speak and turning to the silent mode.

8.7 Limitations of the research

In analysing the data reported in chapter 6, time and space did not permit me to go through each of the searches. I examined about 692 searches all together, of which 541 resulted in concordance lines. Time and space did not permit me to analyse, categorise, and discuss them in detail. It is also difficult to know how to categorise them since there are many different ways of doing so. Therefore, I decided to take a qualitative approach instead of a quantitative approach to present these data. As described in chapter 6, I selected interesting examples of searches from all 692 searches and discussed them in detail. The rationale of choosing these examples is that, based on my extensive knowledge of watching and transcribing the videos of what the students had done while searching and interpreting concordance lines, I picked out those searches accompanied by clear think-aloud protocols to be a representative of the points to make about the student use of a corpus. To further the study, the next stage of this project might be to categorise all 692 in ways that are useful and feasible, for example, to identify and quantify a set of observed successful and unsuccessful concordancing strategies that the students appeared to adopt. The results would be useful for recommending how learners should use a corpus and how to help less successful learners to become more successful. It would also be interesting to research if successful strategies can be taught as well as to identify the kinds of search strings the students enter in a corpus.

8.8 Recommendations for further research

As noted in chapter 3, the participants in this study received short training in using the BNCweb and interpreting concordance lines, and it would be reasonable to recommend enhancing the study by providing the students with a more extensive amount of corpus training before collecting the data. Equipped with the skills necessary for conducting searches and coping with corpus data, the students would provide a clearer picture of how they use a corpus.

To obtain the data through the use of think-aloud method seems to be challenging. Some of the students did not think aloud as much as expected when performing concordance tasks for my study, and this made it difficult for me to interpret how they went about their searches. Thus, one of the issues to be resolved is how to get the students to think aloud as much as possible in a useful way to enrich the data. This could be achieved by having students working in pairs and taking turns encouraging each other to think aloud when necessary while they are doing concordancing tasks.

This study also addresses the question of what materials would help make the students become effective corpus users. Hence, the next step of this study might be to develop concordancing materials to equip students with skills required for effective corpus use. Ideally, the materials should be based on what each student is looking at, but in reality it is difficult to do one-to-one tutorial with the students by asking individuals to look at concordance lines and identify what is special about those lines.

8.9 Final remarks

I came into this study wishing to examine English-major students' use of a corpus. My expectation, that the students would make maximally successful use of the corpus, was not entirely borne out.

I have demonstrated that the students found the tasks they were asked to undertake surprisingly difficult. On the whole, I am not entirely optimistic about the potential for students to use a corpus on their own, but if they are given the right preparation and the right amount of help, they could use a corpus on their own more effectively. Besides the skills in conducting a search and interpreting the results, students need the skills to operate the corpus software. The software that is complicated to use or the one that operate on the web can make concordance tasks less encouraging and less supportive. At times, especially when performing complex searches, the students in this study became demotivated and gave up their searches because they failed to form and devise the searches using the wildcards and tags, etc. They also encountered the problems of slow download speed caused by the BNCweb. Therefore, it might be better if they are provided with specific corpus software that is more user-friendly such as the Sketch Engine. To some extent, having the different corpus software that is easy for students to use can make learner concordancing more enjoyable and better. My own experience of using a corpus to improve my own writing and to teach me more about English encourages me to share this expertise with students.

Appendix 6

A list of queries the students looked up in a corpus

Subject	Problem	Attempt	No.	Query
1	1	1	1	applying
	2	1	2	never+before
	3	1	3	emphasize
	4	1	4	the most important
	5	2	5	another
			6	other
03	1	1	7	not only
	2	2	8	all of the
			9	all of the details
	3	3	10	these brought me
			11	brought
			12	brought me
	4	2	13	sentence structures
			14	sentence structure
	5	1	15	have learned
	6	2	16	every authors
			17	authors
	7	1	18	etc
	8	1	19	knowledges

Subject	Problem	Attempt	No.	Query
04	1	1	20	at present
	2	1	21	most skills
	3	1	22	branches of studies
	4	1	23	or not
05	1	2	24	important
			25	important for
	2	1	26	writing
	3	1	27	good
	4	1	28	should
	5	1	29	therefore
	6	1	30	for me
06	1	1	31	usually
	2	2	32	after
08	1	3	33	to learn
			34	learn
			35	learn
	2	2	36	to known
			37	known
	3	1	38	It’s make

Subject	Problem	Attempt	No.	Query
09	1	1	39	appropriate
	2	1	40	procedure
	3	1	41	process
10	1	3	42	study of
			43	study in
			44	study
	2	2	45	during process
			46	process_N*
14	1	1	47	writing technique
	2	1	48	academic language
	3	1	49	different perspectives
15	1	1	50	smoother
	2	1	51	complete
16	1	1	52	almost
17	1	1	53	write
	2	1	54	others
	3	1	55	must
19	1	1	56	help
	2	1	57	systematically

Subject	Problem	Attempt	No.	Query
20	1	1	58	either way
	2	2	59	into my research
			60	into my
21	1	1	61	learned
	2	1	62	what are they like
	3	1	63	researched information
	4	1	64	one of the easiest ways
	5	1	65	researched
	6	1	66	academic field
22	1	1	67	skill
	2	1	68	order
	3	1	69	idea
	4	1	70	benefit
	5	1	71	organize
23	1	1	72	plagiarism
	2	1	73	become
24	1	1	74	learn
	2	1	75	other
	3	1	76	pay attention
	4	1	77	never bored
	5	1	78	educated
	6	1	79	learnt

Subject	Problem	Attempt	No.	Query
26	1	1	80	important
	2	1	81	native English speaker
	3	1	82	essay
	4	1	83	for
27	1	1	84	in the first place
	2	1	85	major subject
	3	6	86	mistake point
			87	miss point
			88	point
			89	mistaking point
			90	mistaked point
			91	filed [failed] point
28	1	1	92	most useful
	2	2	93	be advanced in
			94	advanced in
	3	1	95	have benefits
29	1	1	96	academic
	2	1	97	weakness
30	1	1	98	successful
	2	1	99	emphasize
	3	1	100	look over

Subject	Problem	Attempt	No.	Query
31	1	1	101	a lots
	2	1	102	thought on
	3	1	103	with
32	1	1	104	knowledge of how to
	2	2	105	help me improve
			106	help improve
	3	1	107	majoring in
	4	1	108	each part
33	1	2	109	bunch
			110	knowledge
	2		111	advantageous
34	1	2	112	finishing course
			113	fisnishing
	2	1	114	correct format
	3	4	115	referencing format
			116	referencing
			117	format
			118	referencing
35	1	1	119	you
	2	1	120	taught

Subject	Problem	Attempt	No.	Query
36	1	1	121	research
	2	1	122	search
	3	1	123	believable
	4	1	124	advantage
	5	1	125	teach
	6	1	126	grammatical
	7	1	127	organize
	8	1	128	title
	9	1	129	complete
	10	1	130	part
	11	1	131	perfectly
	12	1	132	useful
37	1	1	133	hence
	2	1	134	many more
	3	1	135	benefit
	4	1	136	worthwhile
	5	1	137	beneficial
38	1	1	138	could
	2	1	139	outlining
	3	1	140	choosing
	4	1	141	hedging
	5	1	142	the English major

Subject	Problem	Attempt	No.	Query
39	1	1	143	previously
	2	2	144	instructional channel
			145	instruction media
	3	1	146	required subject
	4	1	147	write
40	1	1	148	afraid
	2	1	149	because of
	3	1	150	paraphrase
	4	1	151	feel
	5	1	152	realize
	6	1	153	knowledge
41	1	1	154	obvious difference
	2	1	155	is based on
	3	1	156	a variety of
	4	1	157	rewrite
42	1	3	158	“to be called”
			159	be called
			160	be called a good

Subject	Problem	Attempt	No.	Query
43	1	1	161	English writing
	2	1	162	provide some benefits for me
	3	2	163	write a stuff
			164	write a text
	4	1	165	being a good writer
	5	1	166	attitude towards
	6	1	167	I quite love
	7	1	168	especially
	8	1	169	think freely
45	1	4	170	vocabulary
			171	many vocabulary
			172	lots of vocabulary
			173	a lot of vocabulary
	2	1	174	afraid to
Total	**141**		**174**	

Appendix 7

Classes of words the students looked up in a corpus while writing/editing their writing

Class of words	Searches	Total	Percentage
Noun		**57**	**32.76%**
• Single noun	authors, knowledges, procedure, process, process_N*, skill, order, idea, benefit (S22), plagiarism, essay, point, weakness, bunch, knowledge (S33), format, advantage, title, part, benefit (S37), knowledge (S40), vocabulary	**22**	
• Adj + Noun	academic language, different perspectives, academic field, major subject, mistaking point, mistaked point, f[a]iled point, correct format, referencing format, instructional channel, required subject, obvious difference	**12**	**6.9**
• Noun + Noun	sentence structures, sentence structure, writing technique, instruction media, English writing	**5**	
• Det + Noun	every authors, most skills, many vocabulary	**3**	
• Noun + Prep	thought on, attitude towards	**2**	

Class of words	Searches	Total	Percentage
• Verb + Noun	mistake point, miss point	**2**	
• Quantifier + of + Noun	lots of vocabulary, a lot of vocabulary	**2**	
• Adj + Noun + Noun	native English speaker	**1**	
• Prep + Noun	at present	**1**	
• Noun + of + Noun	branches of studies	**1**	
• Prep + Det + Noun	into my research (S20)	**1**	
• Det + Noun + Noun	the English major	**1**	
• Det + Noun + Prep	a variety of	**1**	
• Prep + Det + Adj + Noun	in the first place	**1**	
• Noun + of + Adv + to	knowledge of how to	**1**	
• Pron + of + Art + Adj + Noun	one of the easiest ways	**1**	

Class of words	Searches	Total	Percentage
Verb		**56**	**32.18%**
• **Verb**	applying, emphasize, brought, writing, learn (S8), learn (S8), known (S8), study, write, help, learned (S21), organize, become, learn (S24), learnt (S24), emphasize (S30), look over, finishing, taught, research, search, teach, organize (S36), outlining, choosing, hedging, write, paraphrase, feel, realize, rewrite	**31**	
• Verb + Noun	pay attention, have benefits, finishing course	**3**	
• Verb + Prep	study of, study in, majoring in	**3**	
• Verb + Verb	help improve	**1**	
• To + Verb	to learn (S8), to known (S8)	**2**	
• AUX + VERB	have learned (S3), be called	**2**	
• Verb + Pron	brought me	**1**	
• Verb + Adv	think freely	**1**	
• Verb + Det + Noun	write a stuff, write a text	**2**	
• Pron + Adv + Verb	I quite love	**1**	
• Prep + Verb	during process	**1**	
• Det + Verb + Pron	these brought me	**1**	

Class of words	Searches	Total	Percentage
• Pron + Aux + Verb	it's make	**1**	
• Verb + Pron + Verb	help me improve	**1**	
• Aux + Verb + Prep	is based on	**1**	
• To + Aux + Verb	"to be called"	**1**	
• Aux + Verb + Det + Adj	be called a good	**1**	
• Verb + Det + Adj + Noun	being a good writer	**1**	
• Verb + Det + Noun + Prep + Pron	provide some benefits for me	**1**	
Adjective		**28**	**16.09%**
• Adjective	important, good, appropriate, smoother, complete, researched, educated, important (S26), academic, successful, advantageous, referencing (S34) referencing (S34), believable, grammatical, complete (S36), useful, worthwhile, beneficial, afraid	**20**	
• Adj + Prep	important for, advanced in, afraid to	**3**	
• Adj + Noun	researched information	**1**	
• Adv + Adj	never bored	**1**	
• Adv + Adj	most useful	**1**	

Class of words	Searches	Total	Percentage
• Adv + Adv + Adj	the most important	**1**	
• Verb + Adj + Prep	be advanced in	**1**	
Adverb		**11**	**6.32%**
• Adverb	etc, usually, almost, systematically, perfectly, hence, previously, especially, a lots	9	
• Adv + Adv	never+before	**1**	
• Conj + Adv	or not	**1**	
Determiner		8	4.60%
• Determiner	another, other (S1), other (S24)	3	
• Det + Noun	either way (det), each part	2	
• Det + Pron	many more	1	
• Quantifier + of + Det	all of the	1	
• Quantifier + of + Det + Noun	all of the details	1	
Preposition		**5**	**2.87%**
• Preposition	with, because of	2	
• Prep + Pron	for me	1	
• Prep + Det	into my	1	
• Pron + Verb + Pron + Prep	what are they like	1	

Class of words	Searches	Total	Percentage
Conjunction	not only, therefore, after, for	**4**	**2.30%**
Modal	should, must, could	**3**	**1.72%**
Pronoun	others, you	**2**	**1.15%**
Total		**174**	**100**

Appendix 8

The language features the students looked up from a corpus while writing in English

Language Features	Searches	Total	Percentage
Colligation	applying, emphasize (S1), appropriate, study of, study in, write (S17), skill, plagiarism, pay attention, educated, be advanced in, advanced in, successful, emphasize (S30), majoring in, advantageous, taught, title, learned, learn (S24), learnt (S24), rewrite, paraphrase, feel, with, help, become, these brought me, brought, brought me, usually, useful, academic, look over, thought on, realize, help me improve, help improve, study, believable, perfectly, afraid, not only, never + before, after, could, must, during process, process_N*, for, with	51	29.31%
Collocation	order, idea (S22), benefit (S22), organize (S22), point, bunch, search, organize (S36), knowledge (S33), mistake point, miss point, mistaking point, mistaked point, filed point, finishing course, finishing, referencing, academic language, knowledge (S40), academic field, major subject, correct format, obvious difference, writing technique, format, referencing format, write, think freely	28	16.09%

Language Features	Searches	Total	Percentage
Acceptability of strings	at present, have benefits, branches of studies, into my research, native English speaker (S26), into my, in the first place, weakness, knowledge of how to, is based on, either way, write a stuff, write a text, instructional channel, instruction media, required subject, what are they like, most useful, referencing (S34:2), "to be called", be called, be called a good, I quite love	23	13.22%
Agreement	another (S1), other (S1), the most important (S1), all of the (S3), all of the details (S3), every authors, authors, most skills, one of the easiest ways, each part, a variety of, vocabulary, many vocabulary, lots of vocabulary, a lot of vocabulary	15	8.62%
Word class	research, advantage, complete (S36), part, worthwhile, beneficial, researched information, researched	8	4.60%
Nouns in the plural	sentence structures (S3), sentence structure (S3), knowledges, different perspectives, essay	5	2.87%
Position	etc, others, hence, previously, because of	5	2.87%
Lexical word + to	process (S9), procedure (S9), afraid to, important	4	2.30%
Form	have learned, never bored, a lots, grammatical	4	2.30%

Language Features	Searches	Total	Percentage
No infomation	important, writing, important for, good, should, therefore, for me, to learn (S8), learn(S8), learn(S8), to known, known, it's make, smoother, complete, systematically, other (S24), you, teach, many more, benefit (S37), outlining, choosing, hedging, the English major, English writing, provide some benefits for me, being a good writer, attitude towards, especially, or not	31	17.82%
Total		**174**	**100**

Appendix 9

A list of searches the students did

Subject	No.	Min	Search term	Type	Search	Type	Results	
01	1	1.16	often	ADV	often+VERB	ADV+VERB	Y	
	2		occur	VERB	often+occur	ADV+VERB	Y	
	3		take place	VERB	take place+PREP	VERB+PREP	Y	
	4		thing	NOUN	thing/things	sing/plu NOUN	Y	
	5	6.33	(thing | things)	NOUN	thing/things	sing/plu NOUN	Y	
	6		(make | makes)	VERB	make people has	VERB+NOUN+VERB	Y	
	7	10.55	that	CONJ	PUNC+that	PUNC+CONJ	Y	
	8		assist in	VERB+ PREP	assist+in	VERB+PREP	Y	
	9		recommend	VERB	recommend that+NOUN/ Subject	VERB+CONJ +NOUN/ Subject	Y	
02	10		strongly	ADV		No think aloud at all		
	11		instantly	ADV				
	12		encompass	VERB				
	13		ended	VERB				
	14		initiative	ADJ				
	15		in which	PREP+ REL PRON				

Subject	No.	Min	Search term	Type	Search	Type	Results	
	16		because	CONJ				
	17		most	ADV			9	0
	18		using	VERB				
	19		appropriately	ADV				
03	20		everything include	PRON+VERB	everything+include	PRON+VERB		N
	21		everything	PRON	everything+VERB	everything+sing/plu VERB	Y	
	22		also alike	ADV+ADJ	also+alike	ADV+ADJ		N
	23		alike	ADJ	also+alike	ADV+ADJ	Y	
	24		also	ADV	also	ADV+ADJ	Y	
	25		do	VERB	do+for	VERB+PREP	Y	
	26		cause	VERB	cause+PREP	VERB+PREP	Y	
	27		way	NOUN	ADJ+way	ADJ+NOUN	Y	
	28		exercise	NOUN	efficient+exercise	ADJ+NOUN	Y	
	29		exercise	VERB	exercise+ADV	VERB+ADV	Y	
	30		exercise	NOUN	ADJ+exercise	ADJ+NOUN	Y	
	31		protect	VERB	protect+from	VERB+PREP	Y	
	32		injury	NOUN	sing/plu NOUN	NOUN	Y	
	33		consort	VERB	consort+with	VERB+PREP	Y	
	34		make	VERB	make+from	VERB+PREP	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	35		exercise	NOUN	flexible+ exercise	ADJ+NOUN	Y	
	36		agile exercise	ADJ+NOUN	agile+ exercise	ADJ+NOUN		N
	37		agile + exercise	ADJ+NOUN	agile+ exercise	ADJ+NOUN	23	N =4
	38		agile	ADJ	agile+ NOUN	ADJ+NOUN	Y	
	39		flexible	ADJ	flexible+ NOUN	ADJ+NOUN	Y	
	40		exercise	ADJ	ADJ+ exercise	ADJ+NOUN	Y	
	41		flexible	ADJ	VERB+ flexible	VERB+ADJ	Y	
	42		prepare	VERB	prepare+by	VERB+PREP	Y	
	43		defend	VERB	defend+from	VERB+PREP	Y	
	44		make	VERB	make+you+ feel	VERB+PRON+VERB	Y	
	45		get along with	PHRASAL VERB+ PREP	get along+with+ NOUN	PHRASAL VERB+PREP +NOUN	Y	
	46		may faint	MODAL+ VERB	may+faint	MODAL+VERB		N
	47		faint	ADJ/VERB	part of speech	ADJ/VERB	Y	
	48		suitable	ADJ	suitable+for	ADJ+PREP	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	49		also	ADV	VERB/MODAL+ also	VERB/MODAL+ ADV	Y	
	50		should	MODAL	should+also	MODAL+ADV	Y	
	51		exercise	VERB	exercise+ADV	VERB+ADV	Y	
	52		rheumatic fever	NOUN	Does the word exist?	NOUN	Y=37	5
	53		safe	ADJ	safe+for	ADJ+PREP	Y	
	54		whatever	DET	whatever+ NOUN	DET+NOUN	Y	
	55		truly	ADV	truly+benefit	ADV+NOUN	Y	
	56		benefit	NOUN	ADJ+benefit	ADJ+NOUN	Y	
04	57		to being	to+GERUND	to+being	to+GERUND	Y	
	58		to being healthy	to+GERUND+ ADJ	to+being+ healthy	to+GERUND +ADJ	Y	
	59		to be healthy	To+AUX+ ADJ	to+be+healthy	To+AUX+ADJ	Y	
	60		someone do	PRON+ AUX	someone+don’t/ doesn’t	PRON+AUX	Y	
	61		Someone do n’t	PRON+ AUX	someone+don’t	PRON+AUX		N
	62		someone does	PRON+ AUX	someone+does	PRON+AUX	Y	
	63		someone do	PRON+ AUX	someone+do	PRON+AUX	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	64		no hundred percent guarantee	DET+NUMBER+ADJ+NOUN	Is it acceptable?	DET+NUMBER+ ADJ+NOUN		N
	65		no percent guarantee	DET+ADJ+NOUN	Is it acceptable?	DET+ADJ+ NOUN		N
	66		no guarantee	DET+NOUN	no guarantee+ PREP	DET+NOUN + PREP	Y =48	=8
	67		lots of	PRON+ of	lots of+sing/plu NOUN	PRON+of+sing/ plu NOUN (lots of can also be used with uncount noun)	Y	
	68		lots of danger	PRON+ of+NOUN	Is it acceptable?	PRON+of+ NOUN		N
	69		lot s danger	PRON+ s+NOUN	lots of danger	PRON+s+NOUN		N
	70		lot danger	PRON+ NOUN	lots of danger	PRON+NOUN		N
	71		a lot of	PRON+ of	a lot of+sing or plu NOUN	PRON+of+NOUN	Y	
	72		bring lots	VERB+ PRON	Bring lots+sing or plu NOUN?	VERB+PRON+ NOUN	Y 3 ex	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	73	18.50	as well	CONJ+ ADV	How it is used?	CONJ+ADV	Y	
	74	20.34	to eating	PREP+ VERB	to+GERUND	PREP+VERB	Y	
	75	23.28	with lots	PREP+P RON	Is it acceptable/do es it occur together?	PREP+PRON	Y	
	76	24.34	same oil	DET+N OUN	Is it acceptable/do es it occur together?	DET+NOUN	Y 1 ex =55	=11
	77	25.35	balanced nutrient	ADJ+N OUN	Is *nutrient* singular or plural? Is there a sing form?	ADJ+NOUN	Y 1 ex	
	78	26.11	balance nutrients	VERB+ NOUN	Is *nutrient* singular or plural? Can I put s at the end?	VERB+NOU N		N
	79	26.29	nutrients	NOUN	Is there a plural form of *nutrient*? Can I put s after it?		Y	
	80	27.15	or	CONJ	No think aloud		Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	81	28.45	have over	VERB+ ADV	Does it exit?	(For having oil over the need of your body)	Y	
	82	30.00	consume over	VERB+ ADV	Does it exist? Colligation?	ดีกว่า have over	Y 3 EX	
	83		cause_V*	VERB	cause+??	Verb pattern	Y	
	84		stink food	NOUN+ NOUN	Does it exist?	NOUN+NOUN		N
	85		stinking	ADJ	stinking+food	ADJ+NOUN	Y	
	86		by the way	PREP+ ART+NOUN	Meaning+use ใช้อย่างไรในแง่ความหมาย		Y	
	87		only necessary	ADV+ADJ	Or only as necessary?		Y=64	=13
	88		only as necessary	ADV+CONJ+ADJ			Y 1 EX	
	89		natural states	ADJ+NOUN	Is states sing or plural?		Y 1 EX	
	90		states	NOUN	Is states sing or plural?		Y	
	91		nutritional source	ADJ+NOUN	Does it exist?			N
	92		nutritional	ADJ	nutritional+ source	ADJ+NOUN	Y	
	93		needing	VERB	NOUN+ needing	NOUN+VERB	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	94		eat suitably	VERB+ ADV	Does it exist?			N
	95		eat suitable	VERB+ ADJ	Does it exist?			N
	96		eat suit	VERB+ VERB	Does it exist?			N
	97		suitably	คำนี้ไม่ค้น				
	98		eat	VERB	eat+ADV	collocation	Y	
	99		eating	VERB	eating+ADV or ADJ	collocation	Y	
	100		suitably	ADV	VERB+ suitably	collocation	Y	
	101		risk of being	NOUN+ PREP+ VERB	risk of being+NOUN	Risk of being+what type of word	Y	
	102		especially	ADV	especially+ punctuation	ADV+punctuation	Y=74	=17
	103		and especially	CONJ+ ADV	and especially+ punctuation	CONJ+ADV+ punctuation	Y	
	104		stored food	ADJ+NOUN	Does it exist? As adj+n	ADJ+NOUN	Y 2 EX	
	105		storing food	VERB+ NOUN	storing+food	ADJ+NOUN	Y 4 EX	
	106		smell come	NOUN+ VERB	NOUN+ VERB	collocation		N
	107		smell	NOUN	smell+VERB	collocation	Y	
	108		bad smell	ADJ+NOUN	bad smell+VERB	collocation	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	109		study about	NOUN+ PREP	study+about	colligation	Y	
	110		study of	NOUN+ PREP	study+of	colligation	Y	
	111		in a study	PREP+ ART+ NOUN	in a study+PREP	PREP+ART+ NOUN+PREP	Y	
	112		make	VERB	How to use it?	use	Y	
05	113		for instance				Y	
	114		vegetarian diet				Y	
	115		fiber				Y	
	116		fibers				Y	
	117		will help				Y	
	118		such as				Y	
	119		prevent				Y	
	120		prevent by				90	N=19
	121		prevent cancer				Y 4 EX	
	122		lots of				Y	
	123		this helps				Y	
	124		proteins of				Y	
	125		if people				Y	
	126		reason why				Y	
06	127		before				Y	
	128		to cherish				Y	
	129		owning				Y	
	130		correctly				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	131		changing				Y	
	132		changing to the good					N
	133		changing to the better					N
	134		changing to the				Y	
	135		can sometimes				Y	
	136		sometimes can				Y	
	137		particular				Y	
	138		particularly				Y	
	139		difficult				Y	
	140		get				Y	
	141		easily teach				Y 2 EX	
	142		teach easily					N
08	143		which				Y=110	=22
	144		point				Y	
	145		moreover				Y	
	146		it also				Y	
	147		with many				Y	
	148		will never				Y	
	149		will never be				Y	
	150		festival;					N
	151		festival ;				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	152		to inaugurate				Y	
09	153		attribute				Y	
	154		distribution				Y	
	155		concentration				Y	
	156		anthropogenic				Y	
	157		interrelate				Y	
	158		run-off				Y	
	159		glacial				Y	
	160		habitation				Y	
10	161		withdrawal behavior				Y 2 EX	
	162		withdrawal				Y	
	163		mouth of institutional					N
	164		institutional				Y	
	165		mouths				Y	
	166		publicly appeared				=130	N =25
	167		publicly				Y	
	168		from at least				Y	
	169		survey				Y	
	170		statistics				Y	
	171		survey				Y	
	172		their child				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	173		entertaining stuffs					N
	174		was diagnosed				Y	
	175		so				Y	
	176		So				Y	
	177		good university				Y	
	178		university				Y	
	179		not all the students					N
	180		not all				Y	
	181		worlds				Y	
14	182		studies state					N
	183		studies stated				Y 1 EX	
	184		studies stated that					N
	185		study stated					N
	186		a study stated					N
	187		studies shown				Y=145 1 EX	=31
	188		studies show that				Y	
	189		how they are going to get					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
	190		how they are going				Y	
	191		how they are				Y	
	192		How they are + ing				Y	
	193		the stuff that is needed					N
	194		the stuff needed					N
	195		stuff needed				Y 1 EX	
	196		stuff which needed					N
	197		prepare… needed					N
	198		prepare + needed					N
	199		that will be needed				Y	
	200		First is				Y	
	201		It is better to have				Y =153	=37
	202		when arrive at the party					N
	203		when arrive at					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
	204		when arriving at				Y	
	205		when arrive					N
	206		when arriving				Y	
	207		when you arrive at				Y	
	208		not a very good condition					N
	209		not very good condition					N
	210		carry wrong message					N
	211		carry wrong messages					N
	212		carries wrong message					N
	213		send wrong message					N
	214		found + in + condition			=157	Y 1 EX	=46
	215		found in + condition				Y 1 EX	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	216		not + very good condition				Y 1 EX	
	217		carry + wrong message					N
	218		send + wrong message				Y 2 EX	
15	219		desire				Y	
	220		focuss					N
	221		focusses				Y	
	222		effect				Y	
	223		amount				Y	
	224		effect				Y	
	225		whole				Y	
16	226		traditionally				Y	
	227		so instead				Y	
	228		in order to				Y	
	229		that to				Y	
17	230		stated				Y	
	231		causes \ that \ cause					N
	232		causes				Y	
	233		effect				Y	
	234		effect_V					N
	235		effect_VV				=173	N=51

Subject	No.	Min	Search term	Type	Search	Type	Results	
	236		affect_VV					N
	237		affect				Y	
	238		lead to				Y	
	239		center				Y	
	240		finding				Y	
	241		work				Y	
	242		in search				Y	
	243		claim				Y	
	244		fashionable				Y	
	245		price				Y	
	246		estimate				Y	
19	247		cause affecting				Y	
	248		by which absorption					N
	249		by which				Y	
	250		observe*				Y	
	251		anaesthetic				Y	
	252		another				Y	
	253		besides				Y	
20	254		products which				Y	
	255		products which could				Y 4 EX	
	256		them treat their					N
	257		treat their				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	258		women use cosmetics				=192	N =55
	259		women use				Y	
	260		treat their skin					N
	261		treat their				Y	
	262		skin				Y	
	263		its danger				Y	
	264		continuously improved					N
	265		continuously improve				Y 1 EX	
	266		used as				Y	
	267		compounds				Y	
	268							
	269		as compounds					N
	270		they use is				Y 5 EX	
	271		make up				Y	
	272		make-up				Y	
	273		who work all day				Y 4 EX	
	274		who worked all day					N
	275		pepper with				Y 3 EX	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	276		pepper				Y	
	277		many contaminations				=205	N=60
	278		many contamination					N
	279		contamination				Y	
	280		many occupations				Y	
	281		very necessary				Y	
	282		large quantity				Y	
	283		baneful				Y	
	284		baneful for					N
	285		innocuous				Y	
21	286		different personalities				Y	
	287		different personality				Y	
	288		pattern of relationship					N
	289		relationship				Y	
	290		get marry					N
	291		get				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	292		get married				Y	
	293		what mentioned					N
	294		what have been mentioned				=216	N =66
	295		what has been mentioned				Y 1 EX	
	296		what was mentioned					N
	297		what was mention					N
	298		are faced				Y	
	299		they face				Y	
	300		conflicts				Y	
	301		prodigious characteristic					N
	302		characteristic				Y	
	303		decent characteristic					N
	304		receive deposition					N
	305		get deposition					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
	306		deposition				Y	
	307		characteristic				Y	
	308		characteristic				Y	
	309		feel exposed				Y	
	310		no rules				Y=226	=72
	311		close				Y	
	312		are close				Y	
	313		being demanding				Y 2 EX	
	314		global communication				Y	
	315		punishment are soft					N
	316		punishments are soft					N
	317		punishment is soft					N
	318		punishment				Y	
	319		punishments are heavy					N
	320		soft punishments					N
	321		heavy punishment				Y 4 EX	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	322		light punishment				Y 2 EX	
	323		talents				Y	
22	324		mankind				Y	
	325		mankind + work					N
	326		mankind + in					N
	327		mankind				Y=236	=79
	328		concentration				Y	
	329		concentration of				Y	
	330		vehicle				Y	
	331		road				Y	
	332		refer				Y	
	333		accelerate				Y	
	334		trend				Y	
	335		drought				Y	
	336		decline				Y	
	337		sea ice				Y	
	338		glacier				Y	
	339		rise				Y	
	340		danger				Y	
23	341		serve	VERB	serve+us	VERB+PRON	Y	
	342		provide	VERB	provide+ PREP	VERB+PREP	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	343		massacre	VERB	to+massacre	TO+VERB	Y	
	344		couldn't have walked	MODAL+NOT+AUX+VERB	could+not+have+walk+ed	MODAL+NOT+AUX+VERB+-ed		N
	345		could + n't + have + walked	MODAL+(+)+NOT+(+)+AUX+(+)+VERB	could+not+have+walk+ed	MODAL+NOT+AUX+VERB+-ed	=252	N =81
	346		could n't have walked				Y 2 EX	
	347		only				Y	
	348		just				Y	
	349		free				Y	
	350		instruction				Y	
	351		classroom				Y	
	352		nowadays				Y	
	353		compare				Y	
24	354		navy	NOUN	navy+for	NOUN+PREP	Y	
	355		journey	NOUN	journey+in	NOUN+PREP	Y	
	356		mean	NOUN	mean+of	NOUN+PREP	Y	
	357		means	NOUN	means+of	NOUN+PREP	Y	
	358		sport	NOUN	sport+for	NOUN+PREP	Y	
	359		accordingly	ADV	accordingly+to	ADV+PREP	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	360		section	NOUN	section+of	NOUN+PREP	Y	
	361		unable	ADJ	unable+to	ADJ+PREP	Y	
	362		function	NOUN	function+in	NOUN+PREP	Y	
	363		function in harness	NOUN+ PREP+ NOUN	Does it exist?	NOUN+PREP + NOUN		N
	364		function in	NOUN+ PREP	function in+type of word	NOUN+PREP + What type of word?	Y	
	365		harness	NOUN	Part of speech	NOUN	Y	
	366		giving	VERB	give+what type of word?	VERB+what type of word	Y =272	=82
	367		furred	ADJ	part of speech	ADJ	Y	
	368		well furred appearance	ADV+ADJ+NOUN	Does it exist?	ADV+ADJ+ NOUN		N
	369		well furred	ADV+ADJ	Does it exist?	ADV+ADJ	Y 1 EX	
	370		medium in	ADJ+PREP	medium+in	ADJ+PREP	Y	
	371		training	NOUN	part of speech	NOUN	Y	
	372		does well	VERB+ ADV	part of speech	VERB+ADV	Y	
	373		get bored	VERB+ ADJ	NO INFO	VERB+ADJ	Y	
	374		stimulation	NOUN	stimulation+ on	NOUN+PREP	Y	
	375		excel	VERB	excel+in	VERB+PREP	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	376		know	VERB	know+to	VERB+PREP	Y	
	377		caring	VERB	care+for	VERB+PREP	Y	
	378		combination	NOUN	combination+of	NOUN+PREP	Y	
	379		use for	VERB+PREP	Use for+gerund	VERB+PREP+GERUND	Y	
26	380		industrial revolution	ADJ+NOUN	the+ industrial revolution	ART+ADJ+NOUN	Y	
	381		at the present time	PREP+ART+ADJ+NOUN	How it is used?	PREP+ART+ADJ+ NOUN	Y	
	382		electric	ADJ	electric+ equipment	ADJ+NOUN	Y =287	=83
	383		equipment	NOUN	electric+ equipment	ADJ+NOUN	Y	
	384		lead	VERB	lead+to	VERB+PREP	Y	
	385		besides	ADV	How it is used?	ADV	Y	
	386		in addition	PREP+NOUN	How it is used?	PREP+NOUN	Y	
	387		apart from that	PREP+PRON	How it is used?	PREP+PRON	Y	
	388		urge	VERB	How it is used?	VERB	Y	
	389		species	NOUN	Is it a sing or plu noun?	NOUN	Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	390		most active	ADV+ADJ	Can we use them together/ acceptable?	ADV+ADJ	Y	
	391		fever	NOUN	a+fever	ART+NOUN	Y	
	392		pain	NOUN	a+pain	ART+NOUN	Y	
	393		stomachache	NOUN	spelling	NOUN		N
	394		result	NOUN	result+PREP	NOUN+PREP	Y	
	395		confront	VERB	confront+ what type of word	VERB+what type of word	Y	
	396		confronting	VERB	Does it exist?	VERB	Y	
27	397		communicative basic					N
	398		basic				Y	
	399		communicative				Y =302	=85
	400		communicatived					N
			communicated basic					N
			communicated				Y	
			basic of communicat*					N
			basic of communicat+					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			basic_of_communicat+					N
			basic of commnicated					N
			Basic communication				Y	
			act a movement					N
			movement				Y	
	410		non verbal communication				Y 6 EX	
			non-verbal communication				Y =307	=92
			confer				Y	
			congratulation				Y	
			second meet					N
			second meeting				Y	
			people meet others					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			people meet other people					N
			meet other people				Y	
			meet others				Y	
S28	420		sustain				Y	
			center				Y	
			lack				Y	
			lacks				Y	
			population				Y	
			population				Y	
			populations				Y	
			die				Y	
			suffer				Y	
			keep up with				Y	
	430		extremists				Y	
			extremists into					N
			extremists				Y=324	=96
			effective				Y	
			less				Y	
			less tightly				Y	
			beggars				Y	
			prostitutions					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			prostitution				Y	
			manslaughter				Y	
	440		manslaughters				Y	
			lacks of				Y 1 EX	
			lack of				Y	
			so crowded				Y	
Sub 29			immigration				Y	
			migration				Y	
			span				Y	
			spans				Y	
			achievement				Y	
			enormously				Y	
	450		countryside				Y	
			urban				Y	
			institution				Y	
			role				Y	
			city				Y	
			overpopulated				Y =346	=97
			time go by				Y 1 EX	
			evident				Y	
			Today				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			sewage				Y	
	460		plenty				Y	
S30			born				Y	
			born of				Y	
			well known				Y	
			succeed				Y	
			succeed with				Y	
			call for				Y	
			satisfy				Y	
			satisfy with				Y 1 EX	
			satisfied with	แทรก			Y	
	470		debate				Y	
			response to				Y	
			response for				Y	
			just as				Y	
			such as				Y	
			won over				Y	
S31			lead to				Y	
			refer to				Y	
			likewise				Y	
			as well				Y	
	480		serious + n				=370	N=98
			serious problem				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			violent problem					N
			solemn problem					N
			problem				Y	
			adj + problem					N
			problem				Y	
			hard problem				Y 2 EX	
			big problem				Y	
S32			alternative to				Y	
	490		Its symptoms				Y	
			the loss of calcium					N
			the loss of				Y	
			calcium loss				Y 1 EX	
			we lost				Y	
			low in				Y	
			die from				Y	
			memory system				Y	
			important role				Y=384	=102

Subject	No.	Min	Search term	Type	Search	Type	Results	
			important role as				Y	
S33	500		define				Y	
			wellknown _A*				Y	
			wellknown				Y	
			well-known				Y	
			worldwide well-known					N
			pass to				Y	
			Japanese				Y	
			festival				Y	
			are opened				Y	
			medicine bottle				Y	
	510		WW II				Y 1 EX	
			world war 2				Y	
S34			make you scare of marriage					N
			make + you + scare					N
			make + scare					N
			make+scare					N
			scare				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			no one never be lonely				=397	N=108
			no one never				Y 3 EX	
			human beings				Y	
	520		must				Y	
			lonely time				Y 1 EX	
			negative attitudes				Y	
			work harder				Y	
			automatically encourage				Y 1 EX	
			encourage				Y	
			hard work				Y	
			doing hard work					N
			doing + hard work					N
			hard work				Y	
	530		succeed faster					N
			succeed + faster					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			succeed				Y	
			twenty four seven					N
			twenty four + seven				Y=4096 EX	=113
			24 7				Y 1 EX	
			twenty-four				Y	
			late night				Y	
			it should be noted				Y	
			nothing comes for free					N
	540		comes for free					N
			comes + free				Y 1 EX	
			evoke				Y	
			motivate				Y	
			environmental surroundings					N
			environmental surrounding					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			environmental + surrounding					N
			surrounding				Y=417	=118
			surroundings				Y	
			more comfortably				Y	
	550		recover fast					N
			recover				Y	
			deal with				Y	
			all the time				Y	
S35			so + called				Y	
			approximately				Y	
			integrate + oxygen					N
			integrate				Y	
			integrate				Y	
			effect				Y	
	560		rather				Y	
			whereabouts				Y	
			resource				Y	
			let + out				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			disappear				Y	
			the + number + of					N ค้างเลย เปลี่ยนใจ
			number + of				Y	
			amount + of				Y =434	=121
			amount				Y	
			comparable				Y	
S36	570		affect				Y	
			safety				Y	
			for example				Y	
			carry off				Y	
			survival				Y	
			steal				Y	
			the number of				Y	
			grow up				Y	
			grow				Y	
			harvest				Y	
	580		lacking				Y	
			risk				Y	
S37			demand				Y	
			resulted				Y	
			difficult				Y	
			result of				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			incessant				Y	
			thickly				Y	
			thickly populated				Y 6 EX	
			identify for				Y	
S38	590		usual				Y	
			when				Y	
			also				Y	
			for instance				Y	
			smiling				Y	
			hugging				Y=462	=121
			touching				Y	
			kissing				Y	
			each other				Y	
			the other way				Y	
	600		basic way				Y	
			people				Y	
S39			severely worried					N
			worried				Y	
			world of globalization					N
			globalization				Y	
			[world of globalization]					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			world + globalization					N
			aforesaid				Y	
			reason				Y	
	610		aforesaid				Y	
			reason				Y	
			aforesaid				Y	
			above-mentioned				Y	
			important				Y	
			rote				Y=478	=125
			understand				Y	
			determinant				Y	
			priority				Y	
			dare				Y	
	620		utilize				Y	
			practicing				Y	
			practice				Y	
			time				Y	
			major				Y	
			importance				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
S40			…..	Computer freezed so I could only hear the voice not the evidence of search except for ‘life’				
			….					
			…..					
			…..					
	630		…..					
			…….					
			….				=488	=125
			…..					
			life					
			….					
S41			keep in touch				Y	
			in reality				Y	
			encounter				Y	
			hang out				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	640		keep in mind				Y	
			obstacle				Y	
			health status				Y	
			harm				Y	
			prolonged				Y	
S42			has been developed for a long time					N
			developed for a long time					N
			for a long time				Y 2 EX	
			be+legendary					N
			be + legendary				Y 1 EX	
	650		could + be + legendary				=499	N =129
			legendary				Y	
			flash back				Y 4 EX	
			flashback				Y	
			flashback_V*					N

Subject	No.	Min	Search term	Type	Search	Type	Results	
			flash + back_V*					N
			flashback_ N*				Y	
			cut back to				Y	
			flash back				Y 4 EX	
			resemble				Y	
	660		resemble + in				Y	
			resemble				Y	
			realistic + movie					N
			realistic+movie					N
			realistic				Y	
			taught				Y	
S43			are dissolving				Y 1 EX	
			huge network				Y	
			their living				Y	
			a double-edged sword characteristic				=513	N =134

Subject	No.	Min	Search term	Type	Search	Type	Results	
	670		double-edged sword				Y	
			double-edged sword characteristic					N
			as the following				Y	
			as followings					N
			response information					N
			responsed information					N
			responsing information					N
			respond information					N
			response/information					N
			information				Y	
	680		entertainment				Y	
			entertainments				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
			entertainments such as			=519	Y 1 EX	=141
			relaxed time				Y 3 EX	
			violent thoughts				Y 1 EX	
			their daily lives				Y	
			attitude towards				Y	
			too much time				Y	
			get strained					N
			get strain					N
	690		get strained					N
			strained				Y	
			get tension				Y 1 EX	
			get stress					N
			stress				Y	
			cause stress				Y	
S45			heard many times				Y 6 EX	
			suited to				Y	
			because				Y	
			any problem				Y	

Subject	No.	Min	Search term	Type	Search	Type	Results	
	700		sustained feeling					N
			a sustained feeling					N
			a sustain feeling				=532	N=148
			feeling				Y	
			result in				Y	
			reduce feeling					N
			decrease feeling					N
			feeling				Y	
			high amounts				Y	
			leads to				Y	
	710		without				Y	
			Without				Y	
			without diseases					N
			without disease				Y 1 EX	
	714		your health is				Y =541	=151

Appendix 10

A list of strings with no search results

Student No.	**Strings**		
S3	everything include	also alike	agile exercise
	agile + exercise	may faint	
S4	Someone do n't	no hundred percent guarantee	no percent guarantee
	lots of danger	lot s danger	lot danger
	balance nutrients	stink food	nutritional source
	eat suitably	eat suitable	eat suit
	smell come		
S5	prevent by		
S6	changing to the good	changing to the better	teach easily
S8	festival;		
S10	mouth of institutional	publicly appeared	entertaining stuffs
	not all the students		

Student No.	Strings		
S14	studies state	studies stated that	study stated
	a study stated	how they are going to get	the stuff that is needed
	the stuff needed	stuff which needed	prepared… needed
	prepare + needed	when arrive at the party	when arrive at
	when arrive	not a very good condition	not very good condition
	carry wrong message	carry wrong messages	carries wrong message
	send wrong message	carry + wrong message	
S15	focuss		
S17	cause \ that \ cause	effect_V	effect_VV
	affect_VV		
S19	by which absorption		
S20	them treat their	women use cosmetics	treat their skin
	continuously improved	as compounds	who worked all day
	many conditions	many condition	baneful for
S21	pattern of relation ship	get marry	what mentioned
	what have been mentioned	what was mentioned	what was mention
	prodigious characteristic	decent characteristic	receive deposition
	get deposition	punishment are soft	punishments are soft
	punishment is soft	punishments are heavy	soft punishments

Student No.	Strings		
S22	mankind + work	mankind + in	
S23	couldn't have walked	could + n't + have + walked	
S24	function in harness	well furred appearance	
S26	stomachache		
S27	communicative basic	communicatived	communicated basic
	basic of communicat*	basic of communicat+	basic_of_communicat+
	basic of communicated	act a movement	second meet
	people meet others	people meet other people	
S28	extremists into	prostitutions	
S31	serious + n	violent problem	solemn problem
	adj + problem		
S32	the loss of calcium		
S33	worldwide well-known		
S34	make you scare of marriage	make + you + scare	make + scare
	make+scare	no one never be lonely	doing hard work
	doing + hard work	succeed faster	succeed + faster
	twenty four seven	nothing comes for free	comes for free
	environmental surroundings	environmental surrounding	environmental + surrounding
	recover fast		

Student No.	Strings		
S35	integrate + oxygen	the number + of ค้างเปลี่ยนใจ	
S39	severely worried	world of globalization	[world of globalization]
	world + globalization		
S42	has been developed for a long time	developed for a long time	be+legendary
	could + be + legendary	flashback_V*	flash + back_V*
	realistic + movie	realistic+movie	
S43	a double-edged sword characteristic	double-edged sword characteristic	as followings
	response information	responsed information	responsing information
	respond information	response/information	get strained
	get strain	get strained	get stress
S45	sustained feeling	a sustained feeling	a sustain feeling
	reduce feeling	decrease feeling	without diseases

Appendix 11

Translation of the students' think-aloud protocols

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
Sub 20	besides	หนูไม่แน่ใจคำนี้อีกคำ ใช้ขึ้นต้นประโยค แปลว่า นอกจากนี้ ใช้ได้ไหมเอ่ย ใช้ได้ นี่ไง ถูก มันใช้เหมือนกับ *moreover, in addition* ทั้งหลาย แปลว่า นอกจากนี้	This is another word I'm not sure of. Can I use it at the beginning of the sentence to mean 'in addition'? Yes, it's correct. It is used here like *moreover* and *in addition*. It means *in addition to*.	The student wanted to know if she could use *besides* at the beginning of the sentence, and she was successful.
Sub 24	unable	*unable* ใช้กับ *to* หรือเปล่า *unable to* ใช้ถูกต้องค่ะ เพราะฉะนั้น ตรงประโยคนี้ไม่ต้องเปลี่ยนอะไรทั้งสิ้น	Is *unable* followed by *to*? Yes, *unable to* is correct, so I don't need to change anything here.	The student wanted to know if *unble* is followed by *to*. She was successful.
Sub 24	use for	*use for taking, use for* ตามด้วยอะไร ตามด้วย gerund ได้ไหม ใช้ได้ เพราะใน corpus มี *use for brewing* เยี่ยม เพราะฉะนั้นอันนี้ก็ใช้ได้	*Use for taking.* What comes after *use for*? Can it be followed by a gerund? Yes, correct. Here I see *use for brewing* in the corpus. So it's correct to say *use for taking*.	The student wanted to know if *use for* is followed by a gerund. She was successful.
Sub 26	lead	ต่อมาก็เป็น *lead* การใช้ *lead* นี่ *lead* ตามด้วย *to* หรือเปล่า *lead us to discount, lead to* มันมี *lead to* กับ *lead* แล้วตามด้วย pronoun มี *us* มี *you* ในประโยคนี้เขาบอกในอุณหภูมิที่สูงขึ้นมันทำให้ปะการังฟอกขาวแล้วก็แนวปะการังตาย แล้วเวลาเขาใช้เขาใช้ยังไง (อ่านตัวอย่าง) มีอะไรที่ชัดเจนกว่านี้ไหมนี่ (อ่าน *The back dating of all VAT not charged on packed lunched*	Next, *lead*, the use of *lead*. Is *lead* followed by *to? Lead us to discount, lead to. Lead* is used in two ways: *lead to* and *lead* followed by pronouns like *us* and *you*. In this sentence, I want to say that the rising temperature causes coral bleaching and coral death. How can I say that? (looking at examples) Are there any clearer examples than these? *The back dating of all VAT not charged on packed lunches could lead to hefty bills for*	The student wanted to know if *lead* is followed by *to*. She was successful.

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
		could lead to hefty bills for hoteliers) อัน นี้แปลว่าไรนี่ (ค้น *hefty* ใน Longdo dictionary) อ๋อ เป็น adjective (ดู tag) *could lead to* อะไรสักอย่าง *lead to* เป็น verb, *lead to cancer* อ่า ไหนดูประโยคนี่สิ *In certain cases its cells undergo changes, which in time can lead to cancer, lead to* น่าจะส่งผลและนำไปสู่ นำไปสู่อันนี้ นี่ไงใช้ถูกแล้ว *lead to* ตามด้วยคำนาม โอเค ถูกต้อง	*hoteliers.* What does this mean? (looking up the word *hefty* in a Longdo online dictionary and looking at the tag in the corpus) Ah, *hefty* is an adjective. Could lead to something. *Lead to* is a verb, lead to cancer. Ah, let me look at this sentence: *In certain cases its cells undergo changes, which in time can lead to cancer. Lead to* probably means to give a result and cause something. It leads to this. It is used correctly. *Lead to* is followed by a noun. Ok. Correct.	
Sub 26	urge	แล้วอันต่อไปนี้ก็คือ *urge* กระตุ้น เขาใช้อย่างไร *urge to* อ๋อ เขาใช้สองแบบ *urge* ที่ตามด้วย *to* ก็คือต้องตามด้วย verb (แต่สังเกตที่ตามด้วย to urge เป็นนาม เช่น why this urge to add, the urge to create, I have this urge to show you, to + verb, urge to see, urge to go - (ดูตัวอย่าง tag urge ที่เป็น verb) แต่ถ้าเกิดว่าเป็น *urge* จะตามด้วยคำนามก็ได้นะ (จริงๆ urge ตรงนี้เป็น verb) *I would urge readers to think through the implications of all*	The next word is *urge*, meaning to strongly persuade someone to do something. How is it used? Urge to, ah, it is used in two ways. *Urge* followed by *to* is followed by a verb. *Urge* can also be followed by a noun. *I would urge readers to think through the implications of all home-made safety…* Umm, are there any other ways in which *urge* is used? Here it is: *that we urge the support of Labour's leaders and members*. It can also be followed by a noun, like this. It also needs to be separated by *to* like *urge readers to think*, so I should say *it urges bacteria to*	The student wanted to find the pattern of *urge*. She was partially successful as she noticed that *urge* is used in two ways – *urge + to*, and *urge + noun + to*, but did not distinguish the difference between the two patterns, where *urge* is used as a noun *(urge + to)* and as a verb (*urge + noun + to).*

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
		- *home-made safety* เอิ่ม มี *urge* อย่างอื่นไหม นี่ไง *that we urge the support of Labour's leaders and members* เอาตามด้วยคำนามก็ได้นี่ นี่ไง ต้องมี *to* คั่นด้วย อันนี้ *urge readers to think* อันนี้ก็ต้องเป็น *it urges bacteria to grow* นี่ไง อันนี้ด้วย *an urge to thrust aside the irresolute* … ไหนมีตัวอย่างอื่นไหม *urge them* นี่ไง *to urge them to put the future of their country* แสดงว่าตรงนี้ต้องใส่ *to* ด้วย อ๊ะ *urge something to* นี่ไง อันนี้ด้วย *So I urge anyone with a garden to visit Sainsbury's* อืม ใส่ *to* ด้วยนะจ๊ะ	*grow.* This is another example: *an urge to thrust aside the irresolute…* Are there any more examples? Here I see *urge them, to urge them to put the future of their country.* This means that I need to use *to* as well. Ah, here it is, *urge something to.* This is another example: *So I urge anyone with a garden to visit Sainsbury's…* Umm, it must be followed by *to.*	
Sub 26	Most active	*Active* เป็น *adjective, most active* เขาใช้กันไหม อืมใช้ได้ (อ่านตัวอย่าง) *one of the sector's most active lobbyists* นี่ไง *Tigers are most active at night.* โอเค ตรงนี้ก็เป็นอันนี้ได้ *mosquito is most active, active* เป็น	*Active* is an adjective. Is it correct to say *most active*? Umm, it's correct. *One of the sector's most active* lobbyists. Here it is. *Tigers are most active at night.* Ok, it is correct to say *mosquito is most active... Active* is an adjective. Ok, correct.	The student wanted to find out if she could say *most active*. She was successful.

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
		adjective (ดู tag ที่ Tigers are most active …) โอเค ตรงนี้ถูกแล้ว		
Sub 32	die from	หนูจะหาคำว่า *die from* จริงๆถูกทำให้ตายปะ ไม่ใช่ตายเอง (ดูตัวอย่าง) อืม แสดงว่าถูกแล้วนะคะ *die from, more than 15,000 women in Britain die from breast cancer and about 20,000 from cancer.* แสดงว่า *die from* ถูกแล้วนะคะ นี่ก็อีกอันที่ทำให้ถูก ไม่ได้ถูกทำให้ตายด้วยโรคหัวใจ	I want to search *die from.* Actually, does it mean to die of illness, not to die a natural death? Umm, *die from* is correct. *More than 15,000 women in Britain die from breast cancer and about 20,000 from cancer.* This means *die from* is correct.	The student wanted to check if she could say *die from*. She was successful.
Sub 34	must	อันต่อไป *As I mention above, everyone must has a lonely time.* อ๊ะ *must* ต้องตามด้วย *has* หรือ *have* แอบลังเลนิดหน่อย งั้นเรามาดูสิ *must* เช็ค *must* หน่อย (ตัวอย่าง must be must, must have, must not) เมื่อกี้เจอ *must have* ก็น่าจะเป็น *must have must be* กลับเป็นรูปเดิม เพราะฉะนั้น *has* ไม่ได้ ต้องกลับเป็นรูปเดิม *must have, everyone must have a lonely time.*	Next, '*As I mention above, everyone must has a lonely time*.' Ahh, is *must* followed by *has* or *have*? I'm a bit uncertain about it. Let's check how to use *must*. I have just seen *must have*, so it should be *must have*. In *must be, be* is in the basic form, so *has* is incorrect. It must be in the basic form: *must have*. *Everyone must have a lonely time.*	The student wanted to know if *must* is followed by *has* or *have*. She was successful.

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
Sub 34	hard work	*hard work* เหมือนเดิมนะ จะดู verb ที่ใช้กับ *hard work* หน่อยนะ *hard work* กำลังจะไปดู verb ที่ใช้กับ *hard work* (It's hard work) อ้ย *It's, can be, are hard work* ไม่มี *do hard work* เลยหรอวะ อ้า มีจริงด้วย *if you don't do hard work you'll get a flat face! do hard work* นะคะ โอเค *do hard work* มีจริงๆด้วย เพราะฉะนั้น *doing hard work* ก็ใช้ได้	Again, I want to check *hard work*. I want to check the verbs that go with *hard work*. Oops, *it's, can be, are hard work*. There are no examples of *do hard work*? Ahh, here it is. *If you don't do hard work you'll get a flat face!* It's ok to say *do hard work. Do hard work* occurs in the corpus, so *doing hard work* is acceptable.	The student wanted to know if she could use the verb *do* with *hard work*. She was successful.
Subject 36	The number of	ต่อไปมาดูการใช้ *the number of* ว่าใช้ยังไง เอ่อ อินเตอร์เน็ตโหลดช้าอยู่นะคะ ก็คือในประโยคนี้ *there is a potential for an increase in the number of cheese related-death as they eat …, … potential for the increase in the number of extremely hot days* ก็ *in the number of* ส่วนมากจะตามด้วย noun ที่มีการขยายด้วย adjective ข้างหน้า ก็น่าจะใช้ถูกแล้วนะคะสำหรับ *the number of extremely hot days*	Now I want to check how to use *the number of*. Err, the download speed is slow. In this sentence *'there is a potential for an increase in the number of cheese related-death as they eat …, … potential for the increase in the number of extremely hot days, in the number of* is usually followed by a noun preceded by an adjective. I think '*the number of extremely hot days'* is correct.	The student wanted to check how to use *the number of*. She was successful.

Subject	Search	Transcription of think-aloud protocols	Translation of think-aloud protocols	My interpretation
Sub 41	Keep in touch	ไม่แน่ใจว่า *keep in touch* ใช้กับ preposition *with* หรือเปล่า (ตัวอย่าง) *keep in touch and call them, keep in touch with her* เนี่ยใช้กับ *with* ได้ *keep in touch with the fish, with his brother* โอเค แสดงว่าอันนี้ใช้กับ *with* ถูกต้อง	I'm not sure if *keep in touch* is used with preposition *with* or not. Keep in touch and call them, keep in touch with her. It can be used with *with*. Keep in touch with the fish, with his brother. Ok, it is correct to use *with*.	The student wanted to check if *keep in touch* is followed by with. She was successful.

List of References

Anthony, L. (2007). *AntConc* (Version 3.2.1) [Computer program]. Retrieved from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Aston, G. (2001). Learning with corpora: An overview. In G. Aston (Ed.), *Learning with Corpora* (pp. 7-45). Houston: Athelstan.

Aston, G., Bernardini, S., & Stewart, D. (Eds.) (2004). *Corpora and Language Learners*. Amsterdam: John Benjamins.

Barnbrook, G. (1996). *Language and Computers: a practical introduction to the computer analysisof language*. Edinburgh Textbooks in Empirical Linguistics, Edinburgh: Edinburgh University Press.

Bennett, G. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Ann Arbor MI: University of Michigan Press.

Bernardini, S. (2000). *Competence, Capacity, Corpora: A Study in Corpus-aided Language Learning.* Bologna: Cooperative Libraria Universitaria Editrice Bologna.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard, & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 225-234). Frankfurt: Peter Lang.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann, & G. Marco (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 165-182). Amsterdam: Rodopi.

Bernardini, S. (2004). Corpora in the Classroom: An Overview of Some Reflections on Future Developments. In J. M. Sinclaire (Ed.), *How to Use Language Corpora in Language Teaching* (pp. 15-36). Amsterdam: John Benjamins.

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.* Harlow: Pearson Education. [LGSWE]

Bloch, J. (2007). *Technologies in the second language composition classroom*. Ann Arbor, MI: University of Michigan Press.

Boulton, A. (2008a). DDL: Reaching the parts other teaching can't reach?. In G. A. Frankenburg (Ed.), *Proceedings of the 8th Teaching and Language Corpora Conference, 2008* (pp. 38-44). Lisbon, Portugal: AssociaÇão de Estudos e de InvestigaÇão Cientufíca do ISLA-Lisboa.

Boulton, A. (2008b). Evaluating corpus use in language learning: State of play and future directions. Paper presented at *AACL 2008 (American Association for Corpus Linguistics)*. Provo, UT: Brigham Young University.

Boulton, A. (2008c). 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. (2009a). Data-driven learning: reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, 35(1), 81–106.

Boulton, A. (2009b). Testing the limits of data-driven-learning: language proficiency and training. *ReCALL*, 21(1), 37-51.

Boulton, A. (2010). Data-driven learning: taking the computer out of the equation. *Language Learning*, 60(3), 534–572.

Boulton, A. (2011). Data Driven Learning: the Perpetual Enigma. In S. Roszkowski, & B. Lewandowska-Tomaszczyk (Eds.), *Explorations across Languages and Corpora* (pp.563-580). Frankfurt: Peter Lang.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Braun, S. (2005). From Pedagogically Relevant Corpora to Authentic Language Learning Contents. *ReCALL*, 17(1), 47-64.

Burns, A. (2001). Analysing Spoken Discourse: Implications for TESOL. In A. Burns, & C. Coffin (Eds.), *Analysing English in a Global Context*: *A Reader* (pp. 123-148). London: Routledge.

Burns, A., Joyce, H., & Gollin, S. (2001). *'I See What You Mean': Using Spoken Discourse in the Classroom.* Sydney: National Center for English Language Teaching and Research.

Carter, R. A. (1998). Order of Reality: CANCODE, communication and culture. *ELT Journal* 52: 43-56.

Carter, R., & McCarthy M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Chambers, A., & O'Sullivan, I. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, 16(1), 158–172.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the Foreign Language Classroom* (pp. 3-16). Amsterdam: Rodopi.

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative research: An Introduction to Think-aloud. *Brook Education*, 12(2), 68-82.

Charles, M. (2012). Proper Vocabulary and Juicy Collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93-102.

Cobb, T. (1999). Applying constructivism: a test for the learner-as-scientist. *Educational Technology Research and Development*, 47(3), 15–31.

Cresswell, A. (2007). Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda, & S. Juan (Eds.), *Corpora in the foreign language classroom* (pp. 267–287). Amsterdam: Rodopi.

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-261). Cambridge: Cambridge University Press.

Ellis, R. (2005). *Instructed second language acquisition: A literature review*. Wellington: New Zealand Ministry of Education.

Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Ed.), *The power of CALL* (pp. 97-113). Houston: Athelstan.

Flowerdew, L. (2009). Applying Corpus Linguistics to Pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, 14(3), 397-417.

Gabrielatos, G. (2005). Corpora and Language Teaching: Just a fling or wedding bells?, Teaching English as a Second or Foreign Language. *TESL-EJ*, 8(4), A-1. Retrieved from http://www.tesl-ej.org/ej32/a1.html

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors?. *System*, 32, 301–319.

Gavioli, L., & Aston, G. (2001). Enriching reality: language corpora in language pedagogy. *ELT Journal*, 55(3): 238-246.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, 58(4), 363-374.

Gray M., & Wardle, H. (2013). Observing gambling behaviour using think aloud and video technology: a methodological review. *NatCen Social Research*. Retrieved from www.natcen.ac.uk

Granath, S. (2009). Who Benefits from Learning How to Use Corpora?. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 47-65). Amsterdam: John Benjamins.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. *Proceedings of the 2006 Conference on Human Factors in Computing Systems*, CHI: 1253-1262.

Hornby, A., & Crowther, J. (1999). *Oxford advanced learner's dictionary* (5th ed.) Oxford: Oxford University Press.

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

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

Johns, T. (1988). Whence and Whither Classroom Concordancing?. In T. Bongarerts, P. de Haan, S. Lobbe, & H. Wekker (Eds.), *Computer applications in language learning Dordrecht* (pp. 9-27). Holland: Foris.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, 4, 1-16.

Johns, T. (1997). Contexts: the background, development and trialling of concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 100-115). Harlow: Addison Wesley Longman.

Keck, C. M. (2004). Book review: corpus linguistics and language teaching research: bridging the gap. *Language Teaching Research*, 8(1), 83–109.

Kettemann, B. (1995). Concordancing in English Language Teaching, *TELL and CALL* 4, 4-15.

Krieger, D. (2003). Corpus linguistics: What it is and how it can be applied to teaching. *The Internet TESL Journal*, 9(3). Retrieved from http://iteslj.org/Articles/Krieger-Corpus.html

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, 5(3), 77–90.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, 14(1), 28–44.

Levy, M. (1990). Concordances and their integration into a word-processing environment for language learners. *System*, 18(2), 177-188.

Lee, D. (2005). Bookmarks for Corpus-based Linguists. Retrieved from http://devoted.to/corpora

Lee, S. (2011). Challenges of Using Corpora in Language Teaching and Learning. *Linguistic Research*, 28(1), 159-178.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25, 56–75.

Leńko-Szymańska, A., & Boulton, A. (Eds.) (2015). *Multiple Affordances of Language Corpora for Data-driven Learning.* Amsterdam: John Benjamins.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. De Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam: John Benjamins.

Long, M., & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-41). Cambridge, England: Cambridge University Press.

Longman (1995). *Longman Dictionary of Contemporary English* (3rd ed.). Harlow: Longman.

Lyster, R. (1998). Recasts, repetition, and ambiguity in L2 classroom discourse. *Studies in Second Language Acquisition, 20*, 51-81. Retrieved from http://dx.doi.org/10.1017/S027226319800103X

McCarthy, M. J. McCarten, J., & Sandiford, H. (2005a). *Touchstone. Student's Book 1*. Cambridge: Cambridge University Press.

McCarthy, M. J. McCarten, J., & Sandiford, H. (2005b). *Touchstone. Student's Book 2*. Cambridge: Cambridge University Press.

McCarthy, M. J. McCarten, J., & Sandiford, H. (2006a). *Touchstone. Student's Book 3*. Cambridge: Cambridge University Press.

McCarthy, M. J. McCarten, J., & Sandiford, H. (2006b). *Touchstone. Student's Book 4*. Cambridge: Cambridge University Press.

McCarthy, M., & O'Dell, F. (2004). *English Phrasal Verbs in Use.* Cambridge: Cambridge University Press.

McCarthy, M. J., & O'Keeffe, A. (2004). Research in the teaching of speaking. *Annual Review of Applied Linguistics*, 24, 26-43.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An advanced resource book***.** New York: Routledge.

McEnery, T., & Xiao, R. (2010). What Corpora can Offer in Language Teaching and Learning. In E. Hinkel. (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 364-380). Vol. 2, London & New York: Routledge.

Miceli, T., & Kennedy, C. (2002). An apprenticeship with the CWIC corpus: a tool for learner writers in Italian. In: C. Kennedy, (Ed.), 83-94.

Milton, J. (2006). Resource-rich web-based feedback: Helping learners become independent writers. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing* (pp. 123-139). Cambridge: Cambridge University Press.

Mishan, F. (2004). Authentic corpora for language learning: a problem and its resolution. *ELT Journal*, 58(3), 219-227.

Mull, J. (2013). The Learner as Researcher: Student concordancing and error correction. *Studies in Self-Access Learning Journal*, 4(1), 43-55.

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

Olson, G, J., Duffy, S. A., & Mack , L. R (1984). Thinking-out-loud as a method for studying real time comprehension process. In D. E. Kieras, & M. A. Just (Eds), New methods in reading comprehension research (pp. 253-286). Hillsdale,, NJ: Erlbaum

O'Sullivan, I., & Chambers, A. (2006). Learners' Writing Skills in French: Corpus Consultation and Learner Education. *Journal of Second Language Writing,* 15(1), 49-68.

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.

Phoocharoensil, S. (2012). Language Corpora for EFL Teachers: An exploration of English grammar through concordance lines, *Procedia-Social and Behavioral Sciences,* 64, 507-514.

Procter, P. (1995). *Cambridge international dictionary of English*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

Reppen, R. (2010). *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Richards, J., & Rodgers, T. (2014). *Approaches and Methods in Language Teaching***.** 3rd ed. Cambridge: Cambridge University Press.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling, & K. Merja (Eds.), *Corpus Linguistics. An International Handbook* (pp. 112-130) (volume 1). Berlin: Mouton de Gruyter.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 31, 205–225.

Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, 11, 129-158.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 11, 11-26.

Schmidt, R. (2001). "Attention." In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. London: HarperCollins.

Sinclair, J. (2003). *Reading concordances: An introduction*. London: Pearson.

Sinclair, J. (Ed.) (2004). *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

Sripicharn, P. (2003). Evaluating classroom concordancing: The use of concordance-Based materials by a group of Thai students. *Thammasat Review,* 8, 203-236.

Sripicharn, P. (2010). How can we prepare learners for using language corpora?. In M. McCarthy, & A. O'Keeffe (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp.371-384). London: Routledge.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. (2004). On very frequent phrases in English: Distributions, functions and structures. *Plenary address given at ICAME 25, Verona, Italy, 19–23 May.*

Sun, Y. C. (2007). Learner perceptions of a concordancing tool for academic writing. *Computer Assisted Language Learning*, 20(4), 323–343.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64-81). Cambridge: Cambridge University Press.

Taljard, E. (2012). Corpus-based language teaching: An African language perspective. *Southern African Linguistics and Applied Language Studies*, 30(3), 377-393.

Thornbury, S., & Slade, D. (2006). *Conversation: From Description to Pedagogy*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamin.

Tribble, C., & Jones, G. (1997). *Concordances in the Classroom. A Resource Book for Teachers.* Houston: Athelstan.

Turnbull, J., & Burston, J. (1998). Towards Independent Concordance Work for Students: Lessons from a case study. *On-Call*, 12(2), 10-21.

Watson Todd, R. (2011). Induction from Self-selected Concordances and Self-Correction. *System*, 29(1), 91-102.

Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.

Willis, D. (2003). *Rules, Patterns and Words: Grammar and Lexis in English Language Teaching*. Cambridge: Cambridge University Press.

Yoon, H. (2008). More than a linguistic reference: the influence of corpus technology on L2 academic writing. *Language Learning & Technology*, 12(2), 31–48.

Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10(3): 130-139.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257–283.

Yoshida, M. (2008). Think-Aloud Protocols and Type of Reading Task: The Issue of Reactivity in L2 Reading Research. In M. Bowles, R. Foote, S. Perpinan, & R. Bhatt (Eds.), *Selected Proceedings of the 2007 Second Language Research Forum* (pp. 199-209). Somerville, MA: Cascadilla Proceedings Project.