論文名稱：以語料庫為依據之學術英文字彙之研究

校所組別：國立交通大學英語教學研究所

畢業時間：九十五學年度第二學期

指導教授：郭志華教授

研究生：林美宏

中文摘要

由於英文在學術界的優勢地位以及高等教育學生人數日益增加等因素，學術英文寫作比以往更受重視。在眾多學術英文文類裡，因為期刊論文代表了主要學術研究成果，並且具有提昇學術地位的功能，使其一直被廣為研究。過去對期刊論文的研究，從不同的層面來探討此一文類，像是段落架構、修辭功能以及語言特色等。在 Swales 發展出 CARS 模式後，期刊論文之序論(Introduction)更成為期刊論文裡最被廣為深究的一個章節。

另一方面，字彙學習在近年來由於電腦語料庫相關技術的發展，重新開拓了不同的研究視野。有研究致力於建構相關字彙表，提供學習者明確的字彙學習目標。另有些研究，擴大對個別字彙的研究，延伸探討搭配語(collocation)或字詞組成(lexical bundles)。甚至更有研究探討字彙在不同言談情境(discourse contexts)的使用情形。現今，大多數的字彙研究均採用語料庫為依據之分析方法，並兼以自然語言分析工具協助，探究真實及大量的語料中之字彙使用。然而，在文類分析的範疇裡，很少有研究致力於探討特殊字彙所具有的文類特色，也就是字彙使用和文類的修辭功能有何關係。

本研究因此致力於探討期刊論文裡序論的字彙使用與文類修辭功能間的連結。我們以語料庫結合文類分析為研究方法，探究言步(moves)或是修辭功能如何透過字彙呈現。我們建構了一個以六十篇資訊工程領域期刊論文所組成的專業領域語料庫，然後用自行發展之標註系統標註所有期刊論文的言步，接著以自己

研發或是既有的自然語言分析工具量化分析語料庫中期刊論文的字彙。我們利用高頻字彙表分析語料庫裡一般英文字彙(GSL)、學術英文字彙(AWL)以及科技領域字彙(Technical Vocabulary)所佔的比例。結果顯示，科技領域字彙在資訊工程領域期刊論文裡佔有很大的比例。字彙頻率累計表(word frequency profiles)更顯示少部分字彙雖重複性很高，在語料庫裡所有出現的不同字彙中所佔比例卻很低，而低頻率字彙反而佔所有不同字彙一半以上的比例，這顯示某些低頻字也應為期刊論文寫作者的學習目標。我們更因此針對學習目的，建構了能夠涵蓋95%資訊工程領域期刊論文內容的字彙表。另一方面，為了探究能夠顯示言步功能的字彙，我們進而辨別期刊論文裡序論的修辭功能或是言步，並依據每一言步的出現頻率和分佈，將其分為主要及次要言步(major and optional moves)，同時也分析主要言步裡的常見言步組合(common move patterns)。為了瞭解言步如何透過字彙來呈現，我們把研究層面從字彙擴展到字詞組成，因為我們認為在文類裡應有一些能代表其修辭功能之字詞組成。我們分別在序論以及每個言步的語料庫裡探究字詞組成，並將所找到的字詞組成，以其功能分為兩類：一為能表現某一言步修辭功能之字詞組成，一為表現普遍學術語用功能之字詞組成。最後，我們探討如何將研究成果應用在學習學術英文字彙上。

ABSTRACT

English for Academic Purposes (EAP) has been attracting more attention than it was because of the predominant role of English in the research world and the increasing number of students in higher education. Research articles (RAs), among all the genres in EAP, have been widely studied as a result of their wide distribution and promotional nature. Studies of RAs have examined various aspects of this genre, especially the textual organization, rhetorical functions, and linguistic features. The examination of RA Introduction, in particular, becomes the most studied section, following the seminal work of Swales' CARS model.

On the other hand, vocabulary learning has regained momentum in recent years. Some studies focused on providing learners with specific vocabulary learning goals through developing wordlists of different purposes. Some further extended the study of vocabulary to word combinations such as collocations or lexical bundles. Still others investigated how words are used in various discourse contexts. Most vocabulary studies nowadays are based on the analysis of target corpora. The corpus-based approach exploits authentic and large amount of language use data, often using NLP tools to facilitate efficient analysis. However, in the field of genre analysis, little research has been devoted to the generic nature of specialized vocabulary; in other words, relating vocabulary use to the rhetorical functions of a genre.

This study, therefore, aims at exploring vocabulary use in RAs, particular in the Introduction section, in relation to its rhetorical functions. A corpus-based, genre-informed approach is used to examine how rhetorical functions or moves are realized through move-signaling words. We construct a specialized corpus, consisting of 60 RAs in the field of computer science (CS). All the RAs are coded with a set of self-developed coding scheme. Then, the text samples are analyzed quantitatively with the help of readily-available or self-developed NLP tools. To explore the nature of words used in the RAs in this particular field, we compile the frequency list of the corpus and analyze the coverage of the GSL(28.20%), AWL(12.75%), and technical words (as generally represented by off-list words) (59.05%) in the list. As shown from these figures, technical vocabulary accounts for a great deal in the CS corpus, suggesting the vocabulary learning goal of learners in CS could be directed towards words other than GSL or AWL. Word frequency profiles further reveal that a very small number of word-forms have very high occurrence rate while low frequency words account for more than half of the vocabulary of the corpus. It can then be inferred that the low-frequency words form a very wide range of vocabulary repertoire RA writers need to use. As a result, we further develop a CS wordlist for pedagogical purposes. It consists of 1402 word families and covers 95% of the vocabulary (types) in the corpus. Next, our focus is directed towards identifying rhetorical functions or moves in RA Introductions in order to further investigate move-signaling words. The major and optional moves are identified based on frequency and range. We then analyze common move patterns for each of the major moves, including 3-move and 4-move patterns. To explore how the moves are realized through vocabulary, we extend our examination from words to word combinations (or lexical bundles) since each register has its own set of lexical bundles which can represent its typical rhetorical functions. Lexical bundles in the Introduction as well as each major move are found. It is observed that there are two types of meaningful bundles. One is the bundles that can signal the rhetorical functions of a specific move while another type of bundles reflects general academic discourse functions, categorized in this study as general bundles. General bundles are further categorized into stances bundles, discourse organizers and referential bundles based on the discourse functions they perform in texts. Among them, referential bundles are found most frequently used. Pedagogical applications and implications such as the use of concordancing tools in the learning of academic vocabulary are finally discussed on the basis of research results.

ACKNOWLEDGEMENTS

This thesis is the result of collective efforts of a number of important and valued people who have directly or indirectly assisted and supported me during my graduate studies. To these people, I would like to express my deepest gratitude and appreciation.

First and foremost, I would like to thank my advisor and mentor, Professor Chih-Hua Kuo, for all her support and guidance during my years as a graduate student at the TESOL Institute of NCTU. Prof. Kuo accomplished the remarkable task of expanding my limited intellectual capabilities to explore domains of knowledge that were completely new to me when I first started my graduate school. Her guidance not only had profound impact on my thinking but was the key to my development as a researcher. I also appreciate the guidance received from my committee members, Dr. Hao-Jan Chen and Dr. Stephanie Weijung Cheng, for insightful comments on my thesis. A special thank you goes to all professors and colleagues in the Language Teaching and Research Center of NCTU, for providing inspiration, encouragement, and warm atmosphere throughout my stay at NCTU.

I would like to thank all of my friends from the various facets of my life who kept me smiling throughout this process – Lawrence, Amber, Leslie, Ruth, Cathy, Jenny, Sandra, Clarence, Eric, Ryan, Hsiun, Jill, Igent, Kelly, Joyce, and others far too numerous to mention here. I was so blessed with many amazing people supporting me and seeing me through some of the difficult parts of this journey.

Lastly and most importantly, I want to express my sincere appreciation and love to my family— my parents and Chia-Ju, for their love, understanding, support, and constant encouragement. I especially want to thank my mom who was a part of every hill and valley. Thank you for listening and loving, and all the “little things” you did to help me achieve this accomplishment.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER ONE
INTRODUCTION

Background

With the growing role of English as the predominant language in disseminating academic knowledge, the purposes of English language pedagogy and research have been extended, especially in the higher education, to familiarizing learners with the language use conventions in academia or even discipline-specific professional communication. Non-native speakers (NNS), in particular, need to acquire the social and linguistic demands of specific academic disciplines in order to survive in the competitive academia. English for Academic Purposes (EAP), thus, has thrived in response to these demands. EAP refers to language research and instruction that focus on the specific communicative needs and practice of particular groups in academic contexts (Hyland & Hamp-Lyons, 2002). As a result, EAP instruction emphasizes on equipping students with effective communication skills in order to actively participate in academic discourse community.

Genre analysis, an approach taken to analyze both the social functions and linguistic features of text, was proposed by Swales (1990). According to Swales, a genre is composed of a class of communicative events, in which some sets of shared communicative purposes are recognized by the expert members of the discourse community and thus constrains the structure, style as well as content of the discourse. With great emphasis on accomplishing social purposes, genre analysis has made itself distinguished from traditional textual analysis that fails to take contextual features of a text into consideration. Since EAP instruction focuses on the specific communicative needs of members in academic contexts, genre analysis has become a seminal approach in EAP research. One of the most influential genre-based work was conducted by Swales (1981), analyzing the information structure or “moves” of research article (RA) Introduction. The Create a Research Space (CARS) model proposed by Swales consists of three basic moves and captures the characteristics of RA introductions adequately.

The wide distribution and promotional nature of RA has made itself a key genre in academic discourse community. A number of researchers, thus, have been devoted to investigating characteristics of RA as an academic genre (Bhatia, 1993; Bhatia, 2004; Swales, 1990; Swales, 2004). Among existing studies, the examination of textual properties and rhetorical structures of RA is the most worth noting. Drawing upon Swales’ “move structure” analysis, many studies have investigated the rhetorical structure of different sections of RA and in a wide variety of disciplines (Brett, 1994; Crookes, 1986; Hopkins & Dudley-Evans, 1988; Swales, 1981; Swales & Najjar, 1987, Williams, 1999). The Introduction of RA, among all the other sections, is probably the most studied section in RA. This is because RA Introduction, for one thing, enables researchers to demonstrate the significance as well as relevance of the current research to academic realm. For another, the complicated nature of RA Introduction has been causing problems to both native and non-native academic writers. Since Swales’ (1981, 1990) seminal work on the schematic pattern of RA Introduction, numerous studies have been conducted to examine the realizations of the CARS model in different genres, even across disciplines (Bunton, 2002; Crookes, 1986; Kwan, 2006; Samraj, 2002). In addition to the macro-level structure analysis, the micro-level textual features such as tense, voice, modals, and collocational patterns have also been examined (Charles, 2003; Gledhill, 2000; Tarone et al., 1981; Salager-Meyer, 1990). Results obtained from these studies have revealed that RA is in fact a highly conventionalized genre.

Although not explicitly stated, the above genre-based studies have one thing in common; that is, they all adopted corpus-based methodologies in the investigation of genres in EAP, RA in particular. Biber, Conrad, and Reppen (1994) indicated that the large collection of authentic data has served as the main strength of the corpus-based approach to linguistic research. The large amount of corpus data, on the one hand, leads to findings that are statistically significant. On the other hand, results based on naturally-occurring data are closer to real-world language use. Despite the advantages, corpus-based methodologies still received criticisms for not taking contextual features of the text into consideration (Widdowson, 1998, 2002; Hunston, 2002). However, L. Flowerdew (2005) argued for an integration of corpus-based and genre-based approaches to text analysis in EAP/ ESP to level against criticisms toward corpus-based approach. She indicated that "corpus-based methodologies have been informed by genre principles of text analysis, while at the same time it has been shown that genre theories can profit from corpus-based methodologies." (pp. 329-330). The attraction of a combined approach of genre analysis and corpus analysis lies in the potential for a corpus to reveal recurrent patterns in a particular genre (Gledhill, 2000). For example, Gledhill (2000) investigated the discourse functions of collocation in research article introductions and found that collocations of high frequency words in medical research abstracts and articles are useful indicators of the genre. In this study, such a combined approach, therefore, is taken in an attempt to identify vocabulary which plays a role in realizing specific rhetorical functions of RA.

On the other hand, the role of vocabulary in academic writing has attracted much attention recently. For example, Coxhead and Nation (2001) suggested that once students have control over the 2000 high-frequency words, the vocabulary learning goals of EAP students could be directed to the learning of academic vocabulary or specialized vocabulary.

Vocabulary learning has gradually regained its force in second language learning. A large number of studies have been devoted to vocabulary learning and teaching in the past two decades (Schmitt, 2000). As indicated in Bogaards and Laufer (2004), recurrent themes in vocabulary research in L2 include the construct of vocabulary knowledge, the relationship between vocabulary knowledge and language proficiency, the role of word frequency in vocabulary learning, explicit versus implicit learning, and testing vocabulary knowledge. In the 70s and 80s, inferring word meanings from context has been the major trend in vocabulary teaching. The emphasis of vocabulary teaching, nevertheless, has gradually switched to explicit instruction in that a number of potential problems resulted from implicit teaching of vocabulary were detected (Sökmen, 1997). Knowing what to teach and how to teach efficiently, thus, becomes an issue in explicit teaching of vocabulary. Beglar and Hunt (2005) showed that the use of wordlists plays an important role in speeding up lexical acquisition. Nation and Kyongho (1995) suggested that the top 2000 high frequency words of English is not only the best choice for learners of general purposes but for learners of academic purposes. Moreover, a number of studies have shown that the 2000 most frequent words of English are able to provide around 80% coverage of academic text.

Many studies have indicated that academic vocabulary causes a great deal of difficulty for learners (Cohen, et al., 1988; Coxhead, 2000). However, two main problems arise with respect to developing vocabulary that EAP learners need most. The first problem is to know the kinds of words that frequently occur in the types of texts EAP learners aim at. The second problem is to know how these words are actually used in the context of specific types of texts, or genres. For the first problem, a number of wordlists, such as General Service List (GSL), University Word List (UWL), and Academic Word List (AWL), have been constructed, largely based on frequency and range analysis of words in a large corpus. They provide a systematic approach to vocabulary development (Coxhead, 2000). Nevertheless, as Coxhead (2000), the creator of AWL, indicated, the construction of these lists does not imply that language learning should rely on decontextualized methods. Therefore, the second problem of vocabulary learning in EAP is to relate target vocabulary to its context, i.e., the generic environment where it occurs as well as to examine how some special vocabulary may play a role in information structuring of a genre or contribute to conventionalized, recurrent lexico-grammatical patterns related to generic structure, as indicated by Swales (1990).

Rationale of the Study

Although many genre studies have investigated the rhetorical functions and organization of RA and produced fruitful results in identifying the information structures, represented as moves, following Swales' seminal work (1990) on RA, little information is available about how these moves are realized lexically and how the selection of words to be taught to EAP learners can be related to the generic distinctiveness of vocabulary. In addition, few attempts have been made to construct wordlists of specific genres or disciplines. On the other hand, despite the popularity of the wordlists mentioned earlier, such as GSL, UWL, or AWL, they are completely based on frequency and range as selection measures and criticized for its inability to extract low-frequency words, which often have high information content (Richards, 1970).

Purpose of the Study

In this study, we intend to explore vocabulary use in RA, with a particular focus on the Introduction section, in relation to its rhetorical structure, or moves. A genre-based, corpus-informed approach is taken to analyze and identify both the moves and the vocabulary, using both self-developed and freely accessible computer software. A corpus of research articles in the field of computer science (CS) will be compiled and both quantitative and qualitative analyses of target vocabulary will be conducted so that words can be examined in their generic environment. Specific research questions are posited as follows:

1. What are the high-frequency words that can characterize CS research articles?
2. What are the major moves in the Introduction section of CS research articles?
3. How are these moves realized lexically through move-signaling words?
4. What are the meaningful lexical bundles of these move-signaling words?

Definition of Terms

1. Type: “the number of types refers to the total number of the different word forms, so that a word which is repeated many times is counted only once.” (Read, 2000).
2. Token: “the number of tokens is the same as the total number of word forms, which means that individual words occurring more than once in the text are counted each time they are used.” (Read, 2000).
3. Off-list Words: The off-list words here refer to words that occur neither in the GSL nor in the AWL. In the current study, the majority of words in this category should be technical vocabulary, although numbers and symbols may be included.
4. BNC Corpus (Written): The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The written part constitutes 90% of the whole BNC corpus, extracting from many kinds of text including newspapers, journals, academic books, popular fiction, published and unpublished

letters, and university essays.(http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=intro).

CHAPTER TWO

LITERATURE REVIEW

The advent of computer corpora has ushered in an unprecedented era in which much linguistic research considered impossible before has become feasible. The corpus-based approach is characteristic of (1) the use of a large amount of authentic data, (2) data-driven, probabilistic computational model, (3) automatic or semi-automatic text analysis, and (4) language use in context (Kuo, et al., 2006). It, thus, promises a horizon of new possibilities for linguistic descriptions and provides new insights and materials for language pedagogy (Aston, 2001; Flowerdew, 2002). Recently, the corpus-based approach has been proposed for EAP (Flowerdew, 2002). Specific writing conventions of EAP are uncovered through the compilation of a specialized corpus which contains authentic EAP data. It can be further combined with genre analysis to inform the design of EAP syllabus and learning materials with research-identified information structure and linguistic features of the target genre and a cornucopia of examples from the corpus.

As indicated in Chapter 1, we intend to relate vocabulary use to the information structure of RA, in particular, move structures and rhetorical functions, taking a corpus-based approach. This chapter, therefore, gives an extensive review of important studies in the research areas involved in this specific research topic. Firstly, we will offer a general introduction to EAP and gradually narrow the focus down to the specific genre of RA. Genre analysis and text analysis studies on this genre are discussed. Findings and results related to the present study are highlighted. In the next part, we examine various perspectives of vocabulary acquisition. Different approaches to vocabulary learning that have been proposed are introduced. Concepts such as academic vocabulary, technical vocabulary, frequency analysis and word lists are presented as well. Finally, we move to corpus linguistics, explicating its origin, development, and applications, particularly in the analysis of vocabulary.

EAP

English for Academic Purposes (EAP), is one of the main branches of English for Specific Purposes (ESP) (Hyland & Hamp-Lyons, 2002). It attends to specific communication needs of professionals in various academic contexts. The field of EAP has developed rapidly ever since 1980s as a result of English as the lingua franca in the academic world. This irresistible trend has forced both native and non-native English speakers in higher education to acquaint themselves with the English language use conventions shared within the specific discourse community. In fact, teachers of EAP have indicated that teaching those who are using English in particular academic and cultural contexts is different from teaching those who are using English for general purposes only. To equip students with the communication skills to participate in the academic milieu, a better understanding of the social and linguistic demands in EAP is needed.

Investigation into different aspects of EAP has been going on over the past 25 years (Flowerdew & Peacock, 2001; Swales, 2001). In the early phase of EAP research, register was the spotlight in which syntactic and lexical features of text were the focus of interest (Biber, 1962; Swales, 1988). However, early register analysis has been criticized for its descriptive rather than explanatory nature and its insufficiency in identifying the underlying functions these surface forms might serve in specific types of text. Later research, thus, began to concentrate on how a particular rhetorical function could be realized through surface linguistic forms (Hyland, 1997). Nevertheless, these studies still failed to address how a given rhetorical function was expressed, especially in specific discourse contexts. In other words, the relationship among form, function, and genre was not explored. Not until early 1990s, particularly after the publication of Swales' canonical work of genre analysis (Swales, 1990), did research start to focus on how specific syntactic and rhetorical structures were used in a specific type of text or genre. Aside from the steady change of research focus in EAP delineated above, the language skills needed to meet the demands of EAP were also investigated.

Writing academically is an important but formidable task for many graduate or Ph.D. students. This apprehension is likely to result from the fact that writing for academic purposes is different from writing for general purposes. The purpose of academic writing is to transmit knowledge and share valuable research findings in a way that could be acknowledged by members of the same discipline or discourse community. In order to communicate effectively, therefore, students need to be aware of the expectations and writing conventions of the discourse community. NNS students, however, have difficulty understanding how to express ideas clearly, how to organize arguments coherently, and how to arrange information appropriately. This difficulty, as explained by Paltridge (2002) and Swales (2004), is likely to be the result of the novice writers' ignorance of the characteristics of specific academic genres. In other words, novice writers are often unfamiliar with the nature and conventions of academic writing. EAP instruction, thus, should aim at raising learners' awareness of important academic genres, acquainting them with the writing conventions of these genres, and finally socializing them into their disciplinary communities.

Genre Analysis

Since the importance of genre knowledge in helping learners to master academic discourse has been widely recognized, a system of analysis that takes into account information content, rhetorical functions as well as interactional features of a genre thus is needed within EAP context. Hopkins and Dudley-Evans (1988) indicated the need for an analysis system that can offer a pedagogically-informed description of academic discourse. However, most of the proposed models at that time were rarely concerned with pedagogical applications. Not until Swales' analysis of article Introductions (1981) did we find an analysis system that successfully incorporated a functional perspective in analyzing texts. Genre analysis, proposed by Swales (1981), was a new approach to analyzing texts. It started to attract interest in language description work in EAP for it provided explanations of how certain texts were linked to social contexts of writing as well as writers' communicative purposes. Swales (1990) presents a comprehensive and detailed definition of genre in his book, *Genre analysis: English in academic and research settings*:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constraints choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community. (Swales, 1990, p.58)

The original version of Swales' analysis of article introductions consisted of a four-move pattern, following the single progression from the first move to the fourth. However, the linear description of the four-move pattern used by Swales received criticisms for the negligence of cyclical patterning of the moves (Crooks, 1984; Hopkins & Dudley-Evans, 1988). Swales (1990), thus, modified the four-move pattern into the three-move, "Create a Research Space (CARS)" model, which is shown in Table 2.1.

Table 2.1 *The CARS Model for Article Introductions*

Move 1	Establishing a Territory
Step 1	Claiming centrality, and/ or
Step 2	Making topic generalization(s), and/ or
Step 3	Reviewing items of previous research
Move 2	Establishing a Niche
Step 1A	Counter-claiming, or
Step 1B	Indicating a gap, or
Step 1C	Question-raising, or
Step 1D	Continuing a tradition, or
Move 3	Occupying a Niche
Step 1A	Outlining purposes, or
Step 1B	Announcing present research
Step 2	Announcing principal findings
Step 3	Indicating RA structure

The development of genre analysis in 1990s was a response to the quest for incorporating discourse context into the description of language use. This approach brings traditionally descriptive linguistic analyses into a more explanatory one in which specific use of language in institutionalized settings is explained. Bhatia (1993) also captures the essential features of a genre: a set of communicative purposes and conventionalized construct recognized by the experts of the discourse community. Thus, the study of genre is not simply to investigate the text itself but also to explore how the generic features of text are related to the communicative purposes underlying the genre and shared by the expert members of a particular discourse community.

Swales' CARS model has been widely adopted for the analysis of RA introductions in other subject areas (Crookes, 1984; Cooper, 1985), different sections of RA (Hopkins & Dudley-Evans, 1988; Kwan, 2006; Ruiying & Allison, 2003) and various genres within the academic discourse (Bunton, 2002). Not all genres, however, have been paid equal attention in the eyes of their disciplinary practitioners (Swales, 2004). RA, among all these genres, is the one that has received the most attention. In the next section, the importance of the genre—RA is discussed.

Research Articles

A number of researchers have given a great deal of the available space to the study of RA in their books (Bhatia, 1993; Bhatia, 2004; Swales, 1990; Swales, 2004; Hyland, 2000). In fact, the phenomenon that RA stands out as a field of interest could be justified in a number of ways. To begin with, the high frequency and wide distribution of RA makes itself a key genre in the academic discourse community. In addition, since RA has become an index of research achievement in academia, its promotional nature leads to a more critical role it plays. However, there seems to be a gap between doing research and writing research. Although it is often thought that writing research is simply a reflection of the reality, that is, reporting the investigation procedure and results as they are, the research article, as a matter of fact, must present such information in convincing propositions and arguments in order to position itself in academia, the knowledge-manufacturing industry. To this end, both EAP research and pedagogy are concerned with what the expert members of the academic discourse community expect from RA, in other words, the distinguished writing conventions of this particular genre.

Recently, different perspectives have been proposed for the analysis of RA in order to shed light on the implicit consensus about its form and style developed over time. The literature that explored different aspects of RA was quite extensive. Prior studies using genre-based approach in the examination of RA, however, could be summarized into two groups based on their different focus. The first group was concerned with the textual properties of RA while the second was on the rhetorical structures. With respect to textual properties of RA, linguistic features such as uses of tense, hedging, modality, and reporting verbs were investigated (Hyland, 1996; Salager-Meyer, 1992; Thompson & Ye, 1991). Tarone et al. (1981; 1998), for instance, examined the use of passive in astrophysics articles. Charles (2003) also studied the use of nouns in the construction of stance in material science. The most interesting aspect of these studies is that although they seem to focus on surface linguistic features of RA, they, in fact, attempt to discover the roles these linguistic forms play in the realization of particular communicative purposes in RA.

On the other hand, a great deal of research has focused on the structure or patterns of rhetorical, informational or conceptual organization of RA in that rhetorical consideration has had a pervasive role in the construction of RA (Swales, 1990). Since Swales' seminal work on the move structure of RA introductions, many studies have been conducted, attempting to apply Swales' model to the analysis of introductions of other genres or of RAs in other disciplines (Bunton, 2002; Crookes, 1986; Posteguillo, 1999; Samraj, 2002). Crookes (1986), for instance, indicated the cyclical nature of introductions. Samraj (2002) conducted an analysis of RA introductions from two fields, Wildlife Behavior and Conservation Biology, using Swales' CARS model. Results revealed that disciplinary variations exist. Thus, a model with greater flexibility is needed. In addition to the prolific studies on the introduction section, a number of studies have examined the macrostructure of other sections of RA. Lim (2006), for instance, investigated lexical and syntactic structures in the method sections of management research articles. There have also been several studies on the results section (Brett, 1994; Nwogu, 1997; Posteguillo, 1999). Brett (1994) reported that results section included not only the statement of new findings but the interpretation as well as comment on the new findings. Following Brett's categories, Posteguillo's (1999) analysis of results section in computer science also supported Brett's findings. Nwogu (1997), on the other hand, used a different scheme of analysis for the results section. Hopkins and Dudley-Evans (1988) also conducted a detailed move analysis of the discussion section of both theses and research articles, attempting to offer a pedagogically useful framework for both teachers and learners. Ruiying and Allison (2003) not only examined the rhetorical structures of RA from results to conclusions but compared their findings with those of Hopkins and Dudley-Evans' (1988).

However, the picture we have of research articles is far from complete. More recently, efforts have been made on linking distinctive linguistic features of RA with its rhetorical structure, exploring how the discourse features may be realized by linguistic forms. For example, Gledhill (2000) explored the rhetorical function of collocation in research article introduction, indicating that recurrent lexical-grammatical patterns may be characteristic of a particular discourse community. The identification of the idiomatic features particular to a genre is of help for understanding the conventions. Of the various levels of linguistic features, lexis is the one that this thesis research intends to explore in depth. We, in particular, endeavor to investigate whether a particular communicative purpose could be realized through vocabulary. In the next section, therefore, we will discuss different aspects of vocabulary with the hope to shed light on the relation between rhetorical functions and academic vocabulary.

Vocabulary Learning

Looking back to the history of development in the teaching of vocabulary, we can note that the teaching and learning of vocabulary have always been secondary to those of grammar. In the past few decades, second language teachers generally held the view that acquisition of vocabulary does not start until the syntactic structures of a language has been mastered (Carter & McCarthy, 1988). However, it has been realized that overemphasis on the functional aspects of language will not ensure the acquisition of an adequate vocabulary (Carter & McCarthy). Not until the 20th century has systematic work been devoted to vocabulary acquisition (Schmitt, 2000).

Before we start to explore different approaches to vocabulary learning, we need to consider what it means to know a new word. Richards (1976) outlined a series of assumptions about lexical competence in which different aspects of word knowledge were covered. Later on, Nation (1990) incorporated Richards' assumptions but distinguished his study from Richards' by categorizing vocabulary knowledge into receptive and productive vocabulary. He proposed that the ability to comprehend a word while we see it or hear it is different from the ability to produce it in that producing language forms by speaking or writing requires higher level of knowledge (Nation, 1990). Thus, research findings have suggested that a person's receptive vocabulary far outnumbers productive vocabulary (Read, 2000; Nation, 2001). From the literature mentioned, we may conclude that the nature of word knowledge is so complex that its acquisition involves a great deal more than just memorizing the core meaning of a word.

Current popular practice of vocabulary learning includes both explicit instruction on selected vocabulary and incidental learning of vocabulary. Since the early 1980s, research has focused on exploring how native speakers (NS) of English acquire vocabulary (Read, 2000; Schmitt, 2000). Results revealed that most NS acquire words incidentally. In other words, a large proportion of their vocabulary is not taught but acquired through listening and interacting with other people. Incidental learning seems to be the dominant way of acquiring vocabulary in L1 (Read, 2000; Schmitt, 2000). This perspective of vocabulary acquisition, thus, greatly influenced vocabulary teaching in second language pedagogy. Incidental learning of vocabulary, such as guessing meanings from context, using monolingual dictionaries, and avoiding explicit instruction on wordlists has been advocated by some researchers. L2 learners are also encouraged to read extensively and infer meanings of unknown words from contextual clues. However, more and more research has indicated that a number of problems may occur if we solely rely on implicit instruction to facilitate second language vocabulary acquisition (Coady, 1993; Haynes, 1993; Sökmen, 1997). Sökmen(1997), for instance, argued that the acquisition of vocabulary through inferring from context might be inefficient for L2 learners who need to learn a great deal of words within a limited amount of time. Also, low-level learners are often frustrated in that their insufficient word knowledge often results in incorrect guessing. Most importantly, guessing from context does not guarantee the retention of vocabulary; that is, we are left unknown whether learners will remember the words they acquire implicitly when they encounter them again on other occasions. Our intention here is not to mitigate the possible potential of implicit learning of vocabulary. Rather, it is suggested that inferring word meanings from context should not be the only method for vocabulary learning and other approaches could be more effective and efficient.

In contrast to implicit learning of vocabulary, current research suggests incorporating explicit vocabulary learning in L2 classrooms. In foreign-language learning environments, explicit instruction or systematic learning of words is able to provide good vocabulary development for learners who have only limited exposure to the target language outside of classroom (Schumit, 2000). This view corresponds to the interaction input hypothesis proposed by Michael Long, indicating that mere exposure to input does not necessarily lead to acquisition. Rather, modified and learner-oriented enhanced input helps along the way and eventually leads to acquisition. Here, the notion of enhanced input implies that efficient explicit instruction not only raises learners' awareness but helps the learning of salient target vocabulary. Pedagogical themes related to explicit instruction of vocabulary, such as integrating new words with the old, providing a number of encounters with words, and promoting a deep level of processing, etc., have been widely discussed (Sökmen, 1997).

Despite possible effectiveness of explicit vocabulary instruction, questions such as how much vocabulary a second language learner needs and whether some words are more useful than others are often asked by teachers and learners. Thus, lexical research concerning effective ways to make vocabulary learning easier and to systematize the selection of vocabulary has been blooming in the last few decades. This research also came to be known as the Vocabulary Control Movement (Schmitt, 2000). One of the two approaches proposed in this vocabulary movement is to use systematic criteria to select the most useful words for language learning. For second language learners, it may not be a feasible goal to build a vocabulary size comparable to native speakers'. In fact, some studies have shown that a much smaller number of words is needed to provide basic comprehension (Hirsh & Nation, 1992). Thus, the criteria for selecting the most useful words to set learning goals become crucial. Frequency has long been the most commonly used criterion in the development of wordlists since the early 19th century. It has been found that a small number of words in English occur frequently. Thus, if learners have access to these high frequency words, they will know a large percentage of running words in texts (Nation & Waring, 1997). However, frequency alone is not effective enough to construct a useful wordlist in that wordlists of different text types based on frequency counts can be very different. In addition, it is observed that some high frequency words with low information content are not what learners need (Carter & McCarthy, 1988). Thus, the range of occurrence of a particular word across different texts has also been incorporated in deciding what to include in a wordlist (Carter & McCarthy, 1988; Nation & Waring, 1997). Recently, other criteria such as representativeness, word families, idioms and set expressions have been considered as well. The dogmatic attitude towards wordlist has been changed and the resulting lists based on these well-established criteria will finally be of great help for pedagogical purposes.

Published in 1953, West's General Service List (GSL) has been the most well-known general vocabulary wordlist. The list which contains around 2000 words was developed from a corpus of 5 million words. Although a variety of criteria were implemented by West to select these words, there has been criticism for its size, age, and the corpus used to develop the list (Carter & McCarthy, 1988). Nevertheless, its high coverage of different text types makes it not only a good source for learning key words, but also a useful guidance for teachers to decide which 2000 words should be taught first.

Although the effectiveness of vocabulary teaching based on wordlists is advocated, wordlists should not be regarded as the only approach to vocabulary learning. In any well-structured vocabulary program, it is suggested to properly strike the balance between explicit and incidental vocabulary teaching in that these two approaches have their own strengths and weaknesses, and thus are likely to complement each other. Other vocabulary learning approaches such as word association or vocabulary learning strategies have also been proposed in vocabulary research.

Pragmatic knowledge of vocabulary use is another aspect of vocabulary learning which is both critical and challenging to second language learners. Within Bachman and Palmer's framework (1996), pragmatic knowledge refers to the language knowledge of recognizing the communicative goals shared between interlocutors within a specific language setting. Many disciplines, for instance, have their own expected style of discourse not only in syntax but also in word choice constraints. This concept corresponds to the register variation proposed by Hallidays (1978). Register variation refers to the influence that certain language situations have on the appropriateness of language use. Many foreign language learners, however, are unaware that language use varies with the expectations of interlocutors and the contexts. Laufer (1997), for instance, indicated that although the ability to select the best word for each situation is crucial to maintain communication, learners seem to lack the competence of recognizing the register restriction of some words.

Academic writing is a genre in which writers have specific communicative purposes to achieve. To effectively fulfill the specific purposes, academic writers need to have a good command of register in language use as shared by the members of their discourse community. Since academic vocabulary serves as the most basic element for achieving communicative goals, better mastery of it can help achieve effective communication. In the next section, therefore, we will firstly delineate the classification of vocabulary made by prior works, and then narrow the focus down to academic vocabulary, exploring its nature, importance, and pedagogical implications.

Academic Vocabulary

To facilitate the teaching and learning of vocabulary, Coxhead and Nation (2001) divided the vocabulary of English into four groups, namely, high frequency words, low frequency words, academic vocabulary, and technical vocabulary. Research has shown that only a small number of the words of English occur very frequently. Thus, it is suggested that a vocabulary size of 2000 words is able to have 80 percent coverage of general written texts (Coxhead & Nation, 2001; Nation & Waring, 1997). When learners have mastered the 2000 words for basic comprehension in general texts, it is wise to direct vocabulary learning to more specialized areas (Coxhead & Nation, 2001; Nation, 2001). As a result, academic vocabulary, the shared vocabulary of several fields, stands out as a good vocabulary learning goal for students with academic purposes.

There are several reasons why academic vocabulary is considered a useful learning goal for learners already having a good control of the first 2000 words of GSL. First, academic vocabulary consists of words common to academic texts but uncommon in non-academic texts (Coxhead, 2000). Prior studies have indicated that most academic vocabulary is closely related to the concepts of science and technology, and thus the classification of academic vocabulary often corresponds to empirical research activities as well as processes (Martin, 1976; Meyer, 1990). An understanding of the distinguished nature of academic vocabulary equips learners with the ability to report and evaluate academic activities efficiently. In addition, academic vocabulary is generally not as well known as technical vocabulary or GSL (Nation, 2001, p.190). For EAP learners, academic vocabulary is reported to be problematic in that they occur with lower frequency than GSL (Worthington & Nation, 1996). Thus, many EAP learners use words acquired from general texts in academic writing, neglecting the formal and information-dense nature of academic texts. Also, it has been noted that academic vocabulary accounts for a substantial number of words in academic texts. The University Word List (UWL) developed by Xue and Nation (1984), for instance, was created through integrating four existing word lists. This wordlist contains 836 word families and provides around 8.5% text coverage in academic texts (Hwang & Nation, 1989). This combined list of academic vocabulary consists of words not in GSL but occur frequently over a range of academic texts. However, the inconsistent selection principles resulted from amalgamating the four different prior studies make the UWL inherit many of the weaknesses of its sources. Thus, in the late 1990s, the UWL has been replaced by the Academic Word List (AWL) (Coxhead, 2000). It is based on a corpus of 3,500,000 running words consisting of four academic disciplines and covering 28 subject areas. With selecting criteria more rigorous than those of the UWL, the resulting list comprises 570 word families and has a slightly better coverage of academic texts (10.0 %) than the UWL, although it contains fewer words. The coverage of the UWL and the AWL is quite substantial since the third 1000 words of GSL would only cover around 4.3 % of the same corpus (Nation, 2001). Finally, the development of academic wordlists is of help for teachers to set vocabulary learning goals, design teaching materials, and help students learn useful vocabulary in a more efficient way (Coxhead & Nation, 2001; Read, 2000).

Lexical Bundles

The knowledge of vocabulary specific to a register/genre is necessary for writers to be considered as a member of that discourse community. However, the knowledge restricted to individual words may not be enough in that words are always used in context. As a result, to explore how words connect with the larger discourse, the examination of vocabulary needs to go beyond the level of single words (Schmitt, 2000).

Lexical bundles are multi-word sequences that statistically co-occur in a given register (Biber & Barbieri, 2007; Cortes, 2004). Instead of being accidental, lexical bundles occur repeatedly in a register, serving important discourse function in texts. In addition to the frequently recurred nature, lexical bundles do not have idiomatic meaning since different bundles serve different discourse functions unique to a particular register. Although most of lexical bundles do not have complete structural units, they usually occur at the beginning of a clause or phrase, bridging two phrases and providing a setting for new information. Thus, they are regarded as important elements in the construction of discourse (Biber, 2007). The proper use of lexical bundles is also regarded as a marker of proficient language use within a register (Cortes, 2004).

In the past, the identification of lexical bundles is based on intuition. The impressionistic view of selecting bundles often leads to the ignorant of some unnoticed bundles out of idiosyncrasy. Recent studies, on the other hand, employ corpus-based approach in recognizing lexical bundles. The empirical selection of lexical bundles is not only based on frequency but facilitated by software programs such as N-Gram Phrase Extractor, kfNgram, or Wordsmith Tools. The frequency cut-off used to identify lexical bundles is, in some way, arbitrary. For example, Biber et al. (1999) considered word combinations recurred over 10 times per millions words as lexical bundles. Cortes (2004), however, employed a frequency cut-off 20 times per millions words as criteria in the comparison of lexical bundles used by students and published authors in two different fields. Biber, Conrad, and Cortes (2004) set a relatively high frequency cut-off of 40 times per million words in recognizing lexical bundles. The concept that lies beyond these numbers is that high frequency is a reflection of the status of lexical bundles.

Since lexical bundles are simply determined by frequency, proper interpretation of them is needed for them to be meaningful in the discourse as a whole. To date, lexical bundles have been analyzed from two perspectives, one considers the structural characteristics of lexical bundles and the other examines the discourse functions they perform (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Cortes, 2004). Some studies have developed a taxonomy of structural types of lexical bundles in that they seem to have strong grammatical correlates (Biber, Conrad, & Cortes, 2004; Cortes, 2004). Others analyzed functions of lexical bundles through examining the context they occur in a given register. Three primary functions have been adopted in analyzing discourse functions of lexical bundles: (1) stance expressions, (2) discourse organizers, and (3) referential expressions (Biber, 2003). According to Biber (2007),

> Stance bundles express attitudes or assessments of certainty that frame some other proposition. Discourse organizers reflect relationships between prior and coming discourse. Referential bundles make direct reference to physical or abstract entities.

Biber, Conrad, and Cortes (2004), for instance, investigated discourse functions of lexical bundles in two university registers – classroom teaching and textbooks and compared them with lexical bundles in conversation and academic prose, using the taxonomy of Biber's (2003). Results indicated that three of the discourse functions are very common in classroom teaching than in any other registers. Also, the most common discourse function in academic prose was referential bundles while the least discourse function employed in this register was discourse organizers. As a result, we may infer that each register has their own set of lexical bundles, representing the typical communicative purposes of that register. An understanding of lexical bundles used in a specific register thus is of great help for us to connect these frequently occurred word combinations in a larger discourse.

Since the above wordlists are all developed using a corpus-based approach, we will explore why such an approach has been widely adopted in recent years and how it has been used to investigate different aspects of language, particularly lexicography.

Corpus Linguistics

Since the above studies on wordlists and lexical bundles are all developed using a corpus-based approach, we will explore why such an approach has been widely adopted in recent years and how it has been used to investigate different aspects of language, particularly lexicography.

Corpus linguistics refers to the study of language on the basis of a large collection of written texts (Kennedy, 1998). In the past, the study of language usually relied on only a small amount of, or elicited, language data. The advent of corpus-based research analysis has brought a new horizon. The collection of a large amount of naturally occurring data of corpus linguistics not only serves as the main strength of it but distinguishes itself from those traditional approaches to linguistic analysis (Biber, Conrad, & Reppen, 1994). Important features of corpus-based linguistic research have been discussed (Kuo, 2002). To begin with, corpus linguistics can be used for a wide range of linguistic research, investigating topics from lower-level word analysis to higher-level discourse structures. In addition, the data-oriented and quantitative nature of corpus-based analysis leads to findings that are statistically significant. Moreover, the data in a corpus is usually real world language data. Thus, corpus-based approach is empirical in nature. Aside from the above features, the development of corpora is also beneficial for pedagogical purposes (Hunston, 2002). For language teaching, on the one hand, corpus-based studies offer information that may not be accessible to native speakers' intuition. For instance, why the word *utterly* often occurs before *different* rather than before *similar*? (Hunston, 2002, p. 137). Also, why some verbs are frequently used in present tense while others are often employed in past tense? On the other hand, students are also encouraged to explore corpora by themselves and thus take the responsibility of learning. Combining the use of corpora and a concordancer, in particular, students are able to investigate and observe the language through such a discovery learning. Studies have also shown that this data-driven learning method is not only beneficial for students but more successful than any other methods (Cobb & Horst, 2001).

The exploitation and utilization of corpus linguistics has changed with the advance of computers. In the early days, constructing and analyzing a corpus is a tedious and painstaking work in that all the work needs to be done manually (Kennedy, 1998). Nevertheless, with advances in computer technology, a revolutionized change has been brought to corpus linguistics (Kennedy, 1998). Computer corpora not only broaden the scope of analysis but also increase the speed and reliability of analysis which are not possible in the past. Moreover, computers make possible the creation of immensely large corpora from a variety of sources, thus solve the problem associated with representativeness. Despite all the advantages brought by computer technology in corpus-based research, we need to keep in mind that the emphasis of corpus-based research is not simply on its quantitative findings but also on qualitative interpretations based on language use data in context. After all, the value of a corpus lies in the insight that we gain from analyzing the large quantity of data.

The corpus-based approach has been used to explore a wide range of research topics in applied linguistics. The application areas include grammatical analysis, collocation analysis, lexicography, language variation analysis, and genre or text type analysis. With respect to lexical applications of corpus data, it has been used in the following areas: (1) providing frequency information (Meijs, 1996; (2) disambiguating meanings and functions of words; and (3) investigating the distribution and use of closely related words (Biber, Conrad, & Reppen, 1996; Meijs, 1996). Gledhill (2000), for instance, investigated the collocational patterns of a number of grammatical words from a corpus of 150 RAs. Kuo (2002) also examined the communicative values of lexicon and grammatical structures in discourse with a corpus of 36 scientific RAs. Cortes (2004) explored the lexical bundles employed by students in two different disciplines and compared students' employment of these word combinations with that of professional authors. The above studies have one thing in common, that is, they all adopted corpus-based approach to investigate the lexical aspects of language. From the literature reviewed, it is expected that corpus-based approach is feasible to explore a broad range of language aspects as well as provides information about social and textual factors that influence language use.

CHAPTER THREE

METHODOLOGY

To explore lexical realization of the rhetorical functions of RAs, genre analysis and corpus-based text analysis are integrated in this study. The former is conducted to investigate the information structure of RA Introduction in computer science in terms of moves, while the latter is aimed at identifying essential vocabulary in this specific genre and field, especially in relation to moves. In this chapter, we start with corpus compilation, followed by frequency analysis and wordlist construction. Next, we explain the development of a coding scheme for move analysis as well as the move tagging process. The construction of the subcorpora of moves and analysis of lexical bundles are finally presented. In explicating the various phases of the study, special attention is drawn to the proposed methodological solutions in response to the research questions indicated in the first chapter. In addition, the study is also characterized by the use of Natural Language Processing (NLP) tools to facilitate data analysis.

The Corpus

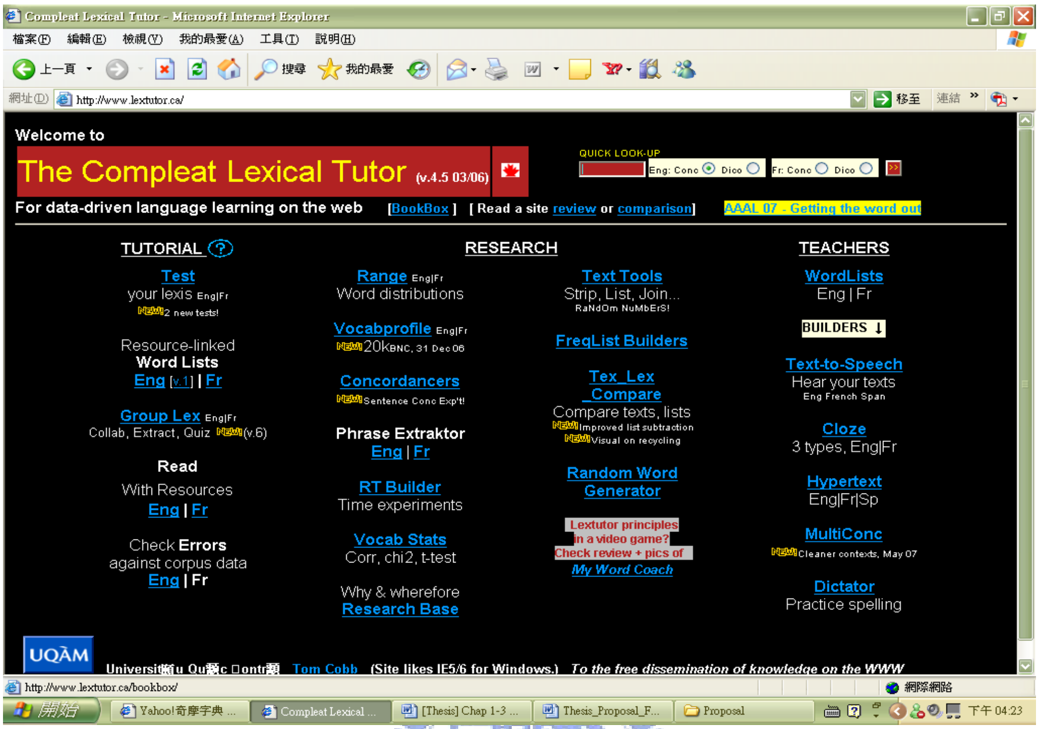
In the study, we compiled a discipline-specific corpus of RAs, focusing on the field of computer science (CS). Three major journals of CS, namely, *IEEE Transactions on Computers*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *Computational Linguistics*, were selected on the basis of the recommendation of the faculty members at the Department of Computer Engineering and Information Science in both Chiao Tung University and Tsing Hua University. Twenty RAs were randomly selected from each journal (See Appendix A for a complete list). These sample texts were taken from issues ranging from 1996 to 2005, approximately two RAs from each year. Therefore, the CS corpus consists of 60 RAs.

Statistic Analysis of the CS Corpus

Frequency List, Vocabulary Profile, and Word Frequency Profiles

After the compilation of the CS corpus, general data analysis was conducted, including basic statistics, such as tokens, types, type/token ratio, and token/type ratio, as well as the construction of a frequency list and its vocabulary profile. Sinclair (1991) suggested that the examination of frequency list is of help in revealing the composition of word-forms in a large corpus. The frequency list was first obtained using software AntConc 3.01. To further unveil the composition of the frequency list, Vocabulary Profilers from the website Compleat Lexical Tutor (www.lextutor.ca) was used to develop a vocabulary profile which can show the percentages of the GSL (West, 1953), the AWL (Coxhead, 2000), and off-list words in the list (See Figure 3.1). The Compleat Lexical Tutor, developed by Tom Cobb, contains a variety of tools for data-driven analyses on the web.

As indicated by Sinclair (1991), to know how words distribute in the corpus, we can construct word frequency profiles. That is, we calculated the coverage of the ten word-forms having the highest frequencies and word-forms having frequencies low than 10 times to learn the proportions they constitute in the corpus. Such word proficiency profiles also inform us of the relative rates of high-frequency words and low-frequency words so that we can have a better understanding of the range of vocabulary use in the corpus. We also screened out the 50 most frequent content words in the CS corpus and examined how they occur in a TESOL Corpus and the BNC Written. The comparison was done by a self-developed program – Freqlist, which is capable of comparing various wordlists. It is hoped that the whole picture of vocabulary use in the –RAs in computer science could be revealed through various data management techniques.



Welcome to

The Compleat Lexical Tutor (v.4.5 03/06)

QUICK LOOK-UP Eng: Conc ● Dico ○ Fr: Conc ○ Dico ○ >>

For data-driven language learning on the web [BookBox] [Read a site review or comparison] AAAL 07 - Getting the word out

TUTORIAL ?

Test
your lexis Eng|Fr
NEW!2 new tests!

Resource-linked
Word Lists
Eng [v.1] | Fr

Group Lex Eng|Fr
Collab, Extract, Quiz NEW!(v.6)

Read
With Resources
Eng | Fr

Check Errors
against corpus data
Eng | Fr

RESEARCH

Range Eng|Fr
Word distributions

Vocabprofile Eng|Fr
NEW!20kBNC, 31 Dec 06

Concordancers
NEW!Sentence Conc Exp't!

Phrase Extraktor
Eng | Fr

RT Builder
Time experiments

Vocab Stats
Corr, chi2, t-test

Why & wherefore
Research Base

Text Tools
Strip, List, Join...
RaNdOm NuMbErS!

FreqList Builders

Tex_Lex
_Compare
Compare texts, lists
NEW!Improved list subtraction
NEW!Visual on recycling

Random Word
Generator

Lextutor principles
in a video game?
Check review + pics of
My Word Coach

TEACHERS

WordLists
Eng | Fr

BUILDERS ↓

Text-to-Speech
Hear your texts
Eng French Span

Cloze
3 types, Eng|Fr

Hypertext
Eng|Fr|Sp

MultiConc
NEW!Cleaner contexts, May 07

Dictator
Practice spelling

UQÀM Universit[illegible]u Qu[illegible]c [illegible]ontr[illegible] Tom Cobb (Site likes IE5/6 for Windows.) *To the free dissemination of knowledge on the WWW*

Figure 3.1. The Compleat Lexical Tutor

Construction of CS Wordlist

Since the frequency list, which consists of all types (or different word-forms), is hugely long and not suitable for vocabulary learning, we are interested in producing a genre- and field-specific wordlist, a CS wordlist, for pedagogical purposes. An important issue in the development of word lists is the criteria for word selection, as different criteria may lead to different results (Coxhead, 2000). Studies (Hirsh & Nation, 1992; Laufer, 1989) have suggested that 95% coverage is sufficient to allow reasonable comprehension of any text in concern. Word-forms that had 95% coverage of the whole corpus were thus regarded as an appropriate criterion in the development of the CS wordlist. Since the corpus had 388,396 running words, 95% coverage meant 368,976 running words. Based on the frequency list, which contained both word forms and their frequencies in a descending order, we calculated accumulated frequencies of word forms until we reached the number of 368,976. The corresponding frequency of the word is 5 when the added frequency reached 368,976. Thus, it was determined that words that had a frequency of 5 or higher constituted 95% of the whole corpus. A second consideration in developing the CS wordlist was that it would not include simple and general words, such as those in the GSL, since RA writers usually had at least intermediate or high-intermediate proficiency level. Therefore, after using the threshold of a frequency of 5 to exclude low frequency words, we further compared these words with the words in GSL, using Freqlist again, in order to screen out words that are regarded as general, function or easy words. Nevertheless, if words in the GSL have special meanings or usages in the field of CS, they are retained. In other words, they should be regarded as field-specific vocabulary and therefore included in our wordlist. We then grouped words in the final list into word families. (See Appendix B)

Statistic Analysis of the RA Introduction

Basic Statistics of RA Introduction

Although the corpus compiled consists of complete RAs, we focused the more qualitative and in-depth analysis of move-signaling words and lexical bundles on a single section—the Introduction in this study, since the study was part of a large NSC research project which investigated various features of complete RAs. Therefore, we extracted the Introduction section of the 60 RAs in the CS corpus and compiled a smaller RA Introduction sub-corpus. With the same procedure, a frequency list specific to Introduction was constructed. With the frequency list, the coverage of the GSL, AWL and off-list words could be revealed, using the Vocabulary Profilers again.

The Coding Scheme and Move Analysis

Identification of major and optional moves

A coding scheme containing rhetorical moves in the Introduction was developed for move analysis of the RA introduction samples in the corpus. Based on research findings of important genre analysis of RA Introduction (Bhatia, 1993; Bunton, 2002; Dudley-Evans, 1996; Samraj, 2002; Swales, 1990), a prototypical scheme was first developed. Preliminary analysis was then conducted using the scheme. Modifications were made throughout the process of analysis in order to develop a feasible coding scheme of moves that reflects the real nature of the information structure of RA Introduction in CS. The final version of the coding scheme, thus, was realistic and empirically tested, accommodating possible variations as a result of the nature of CS RAs. The coding scheme is illustrated in Table 3.1.

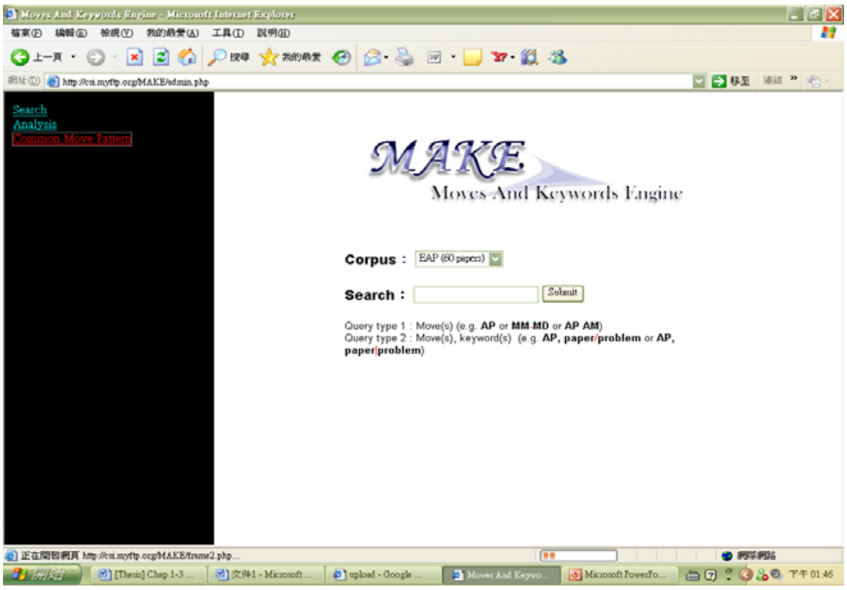
Table 3.1 *Coding Scheme for Research Article Introduction in CS*

Move	Rhetorical Function of Move
IL	literature review or reference to other studies
IM	methods or theories
IP	purposes or major tasks
IB	background information
IG	gap or missing information
IO	local or global organization
IV	values
IC	partial or complete conclusions, evaluation
IJ	justification or reasons
IF	reference to tables or figures
IR	results
IQ	research questions

Move analysis was carried out using the self-developed coding scheme. Four raters participated in the analysis, three at the TESOL Institute in Chiao Tung University (two faculty members and the researcher herself) and a Ph.D. student in

CS, who also served as a specialist informant for the study. Pilot analysis started with the clarification and elaboration of the defining criteria of each move. Three samples of RA Introduction were randomly selected from the corpus and hand-coded by the four raters to check inter-rater reliability. Two problems were encountered during the coding process. It was difficult, for one thing, to identify the boundaries of some moves in the samples. For another, it was found that certain text segments serve more than one rhetorical purpose and thus need to be coded with a combination of different moves. Face-to-face discussion of such issues was held to reach consensus. Move analysis was then conducted for all samples. Each week the four raters met once to discuss problems encountered in identifying moves. After move analysis was completed, the moves were tagged on all electronic text samples in the computer. The construction of the computerized corpus is to facilitate data analysis, using self-developed or ready-made natural language processing (NLP) tools.

MAKE (Moves And Keywords Engine) (See Figure 3.2) is one of such tools capable of extracting all occurrences of any combination of move(s) and/or keyword(s) from a corpus, the sample in which are tagged with moves. Thus, MAKE was used to extract all occurrences of each move from the corpus. The frequency and range of all the moves in 60 Introduction samples were then calculated. This revealed how commonly each move occurs; in other words, whether a move is a major or optional move in RA Introduction.



Moves And Keywords Engine - Microsoft Internet Explorer
http://exi.myftp.org/MAKE/admin.php
Search
Analysis
Common Move Pattern
MAKE
Moves And Keywords Engine
Submit

Corpus : EAP (60 papers)

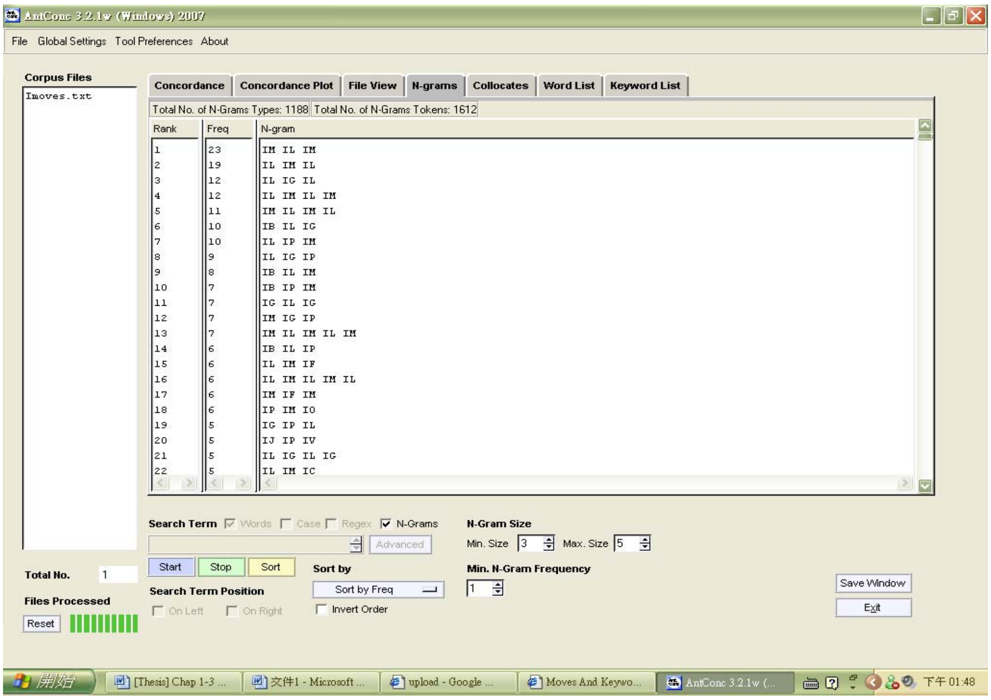
Search :

Query type 1 : Move(s) (e.g. **AP** or **MM-MD** or **AP AM**)
Query type 2 : Move(s), keyword(s) (e.g. **AP, paper/problem** or **AP, paper|problem**)

Figure 3.2. MAKE

Identification of common move patterns

In addition to identifying the major and optional moves in RA Introduction, we further examined how these moves were used in combination with other moves to realize the communicative purposes of the Introduction. As a result, we uploaded the Introduction corpus to AntConc, and used the function “clusters.” Among all the move combinations derived, a set of criteria, including both range and frequency, were used for selecting common move patterns. Moves with a range rate over 50%, that is, occurring in more than 30 RAs, were included in the consideration of common move patterns. A second criterion was frequency. We selected move patterns only with a frequency cut-off of 10 or higher. In addition, since there were too many 2-move patterns and no 5-move patterns with a frequency of 10 or higher, we finally selected only 3-move and 4-move patterns (See Figure 3.3 as an illustration).



Corpus Files: Imoves.txt

Total No. of N-Grams Types: 1188 Total No. of N-Grams Tokens: 1612

Rank	Freq	N-gram
1	23	IM IL IM
2	19	IL IM IL
3	12	IL IG IL
4	12	IL IM IL IM
5	11	IM IL IM IL
6	10	IB IL IG
7	10	IL IP IM
8	9	IL IG IP
9	8	IB IL IM
10	7	IB IP IM
11	7	IG IL IG
12	7	IM IG IP
13	7	IM IL IM IL IM
14	6	IB IL IP
15	6	IL IM IF
16	6	IL IM IL IM IL
17	6	IM IF IM
18	6	IP IM IO
19	5	IG IP IL
20	5	IJ IP IV
21	5	IL IG IL IG
22	5	IL IM IC

Search Term ☑ Words □ Case □ Regex ☑ N-Grams

N-Gram Size Min. Size 3 Max. Size 5

Min. N-Gram Frequency 1

Sort by: Sort by Freq □ Invert Order

Search Term Position □ On Left □ On Right

Total No. 1

Files Processed

Figure 3.3 Common move patterns extracted from AntConc

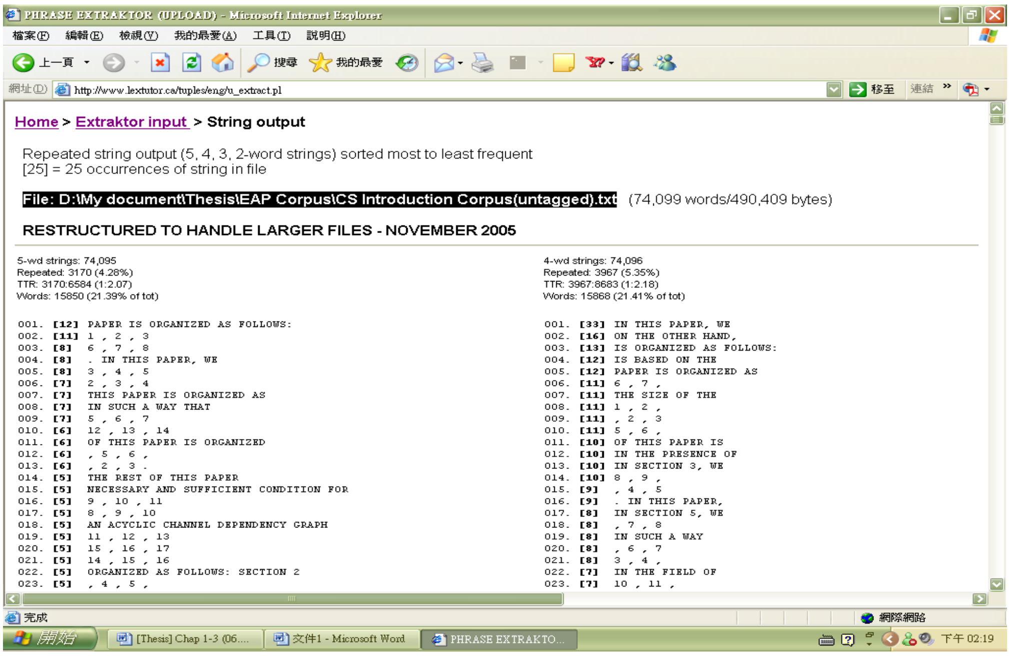
Lexical Bundles in Introduction and in Major Moves

To empirically explore the linguistic realization of moves, the text of all occurrences of each move, as extracted by MAKE from the corpus, was collected and compiled into a move corpus. Thus, we had 12 corpora of moves.

Frequency analysis of the corpus of a move also yielded a frequency list of words in that move. Based on the frequency list, we attempted to qualitatively examine whether high-frequency words of each move can be related to the rhetorical functions of the move.

Aside from investigating how moves were realized by words, we also extended our analysis from words to lexical bundles of words in both the Introduction and major moves. The idea of analyzing lexical bundles is based on the concept of phraseology. That is, recurrent lexical bundles can be regarded as formulaic expressions in a particular genre where conventionalized language use is expected. Thus, the intention was to find lexical bundles that recurred in the Introduction and in each move.

The Introduction corpus and the sub-corpus of each move were searched using an online tool N-Gram Phrase Extractor (http://www.lextutor.ca/typles/eng/) (See Figure 3.4). The major focus of the present study was on the three- and four-word bundles because of the limited size of the corpus. Through the process, it was hoped that meaningful lexical bundles of each move could be identified.


PHRASE EXTRAKTOR (UPLOAD) - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://www.lextutor.ca/tuples/eng/u_extract.pl
移至 連結
完成
網際網路
開始 [Thesis] Chap 1-3 (06.... 文件1 - Microsoft Word PHRASE EXTRAKTO...
下午 02:19

Home > Extraktor input > String output

Repeated string output (5, 4, 3, 2-word strings) sorted most to least frequent
[25] = 25 occurrences of string in file

File: D:\My document\Thesis\EAP Corpus\CS Introduction Corpus(untagged).txt (74,099 words/490,409 bytes)

RESTRUCTURED TO HANDLE LARGER FILES - NOVEMBER 2005

```
5-wd strings: 74,095
Repeated: 3170 (4.28%)
TTR: 3170:6584 (1:2.07)
Words: 15850 (21.39% of tot)

001. [12] PAPER IS ORGANIZED AS FOLLOWS:
002. [11] 1 , 2 , 3
003. [8]  6 , 7 , 8
004. [8]  . IN THIS PAPER, WE
005. [8]  3 , 4 , 5
006. [7]  2 , 3 , 4
007. [7]  THIS PAPER IS ORGANIZED AS
008. [7]  IN SUCH A WAY THAT
009. [7]  5 , 6 , 7
010. [6]  12 , 13 , 14
011. [6]  OF THIS PAPER IS ORGANIZED
012. [6]  , 5 , 6 ,
013. [6]  , 2 , 3 .
014. [5]  THE REST OF THIS PAPER
015. [5]  NECESSARY AND SUFFICIENT CONDITION FOR
016. [5]  9 , 10 , 11
017. [5]  8 , 9 , 10
018. [5]  AN ACYCLIC CHANNEL DEPENDENCY GRAPH
019. [5]  11 , 12 , 13
020. [5]  15 , 16 , 17
021. [5]  14 , 15 , 16
022. [5]  ORGANIZED AS FOLLOWS: SECTION 2
023. [5]  , 4 , 5 ,
```

```
4-wd strings: 74,096
Repeated: 3967 (5.35%)
TTR: 3967:8683 (1:2.18)
Words: 15868 (21.41% of tot)

001. [33] IN THIS PAPER, WE
002. [16] ON THE OTHER HAND,
003. [13] IS ORGANIZED AS FOLLOWS:
004. [12] IS BASED ON THE
005. [12] PAPER IS ORGANIZED AS
006. [11] 6 , 7 ,
007. [11] THE SIZE OF THE
008. [11] 1 , 2 ,
009. [11] , 2 , 3
010. [11] 5 , 6 ,
011. [10] OF THIS PAPER IS
012. [10] IN THE PRESENCE OF
013. [10] IN SECTION 3, WE
014. [10] 8 , 9 ,
015. [9]  , 4 , 5
016. [9]  . IN THIS PAPER,
017. [8]  IN SECTION 5, WE
018. [8]  , 7 , 8
019. [8]  IN SUCH A WAY
020. [8]  , 6 , 7
021. [8]  3 , 4 ,
022. [7]  IN THE FIELD OF
023. [7]  10 , 11 ,
```

Figure 3.4. N-Gram Phrase Extractor

CHAPTER FOUR

RESULTS

This chapter reports the main findings of the study. The results will be presented firstly at the level of the whole CS corpus, and then to the level of the CS Introduction subcorpus. In terms of the whole CS corpus, statistic analysis includes the composition of the CS corpus, coverage of the GSL,AWL and technical words in the CS corpus, word frequency profiles, and comparison of the top 50 content words of CS corpus with those of a TESOL corpus and the BNC Written. Then, the focus is narrowed down to the Introduction section. To shed light on the rhetorical functions of RA Introduction, major and optional moves as well as the common move patterns are analyzed and discussed. Since the study is aimed to relate vocabulary use with rhetorical functions of each move in RA Introduction, generic vocabulary and move-signaling bundles are identified based on both the Introduction corpus and the subcorpus of each move. Pedagogical implications are finally discussed.

Vocabulary Use in CS Research Articles

The first research question posed in this study is: what kinds of words characterize CS research articles? This section presents the results that provide an answer to this question. We firstly examined the composition of our specialized CS corpus, calculating its coverage of general vocabulary, academic vocabulary, and technical vocabulary. In addition, since current second language vocabulary learning research suggests, as discussed in Chapter 1, explicit instruction or systematic learning of words can provide good vocabulary development for learners who have only limited exposure to the target language outside of classroom, we developed a CS wordlist, targeting the needs of EFL students in one specific field. Finally, to unveil the distinguished nature of vocabulary in this scientific discipline, the fifty most frequent word forms in the CS corpus were compared with those in a TESOL corpus and with those in the Written part of the British National Corpus (that is, the BNC Written).

Composition and Vocabulary Profile

Table 4.1 shows the composition of the CS Corpus. As shown in the table, the corpus consists of 60 RAs in three major journals in the field of computer science. The total running words of the corpus is 375,978 although the number of running words in each journal varies slightly. The token/type ratio indicates that a word occurs, on average, 15.91 times in the corpus. On the other hand, the type/token ratio shows that there are 63 different word forms per 10^3 words.

Table 4.1 *Composition of the CS Corpus*

Basic Text Statistics	IEEE Transactions on Computers	IEEE Transactions on Pattern Analysis and Machine Intelligence	Computational Linguistics	Total
Number of Articles	20	20	20	60
Length of the Text in Word-forms (Tokens)	142,169	104,757	129,052	375,978
Number of Different Word-forms (Types)	7,615	7,145	8,875	23,635
Token/Type	18.67	14.66	14.54	15.91
Type/Token	0.054	0.068	0.069	0.063

The vocabulary profile analysis of the CS corpus in terms of the total number of types (i.e., different word forms) was carried out using the online Web Vocabulary Profiler provided by the well-known website Compleat Lexical Tutor. Table 4.2 demonstrates that the proportions of three types of vocabulary in the CS corpus. The first 1,000 (K1 words) and the second 1,000 (K2 words), combined as the first type and usually referred to as the GSL, account for 28.20% of the frequency list, which is actually the word types in the CS corpus. The AWL accounts for 12.75% and the off-list words 59.05%. The off-list words here refer to words that occur neither in the GSL nor in the AWL. In the current study, the majority of words in this category should be technical vocabulary, although numbers and symbols may be included. The proportions show the vocabulary register of both the genre of RA and the field of CS. The data also mean writers of RA in CS use a lot of field-specific vocabulary as well as academic vocabulary. This provides a rationale for the development of a CS wordlist.

Table 4.2 *Vocabulary Profile of the CS Corpus (in terms of types)*

K1 Words	20.63%
K2 Words	7.57 %
K1+K2	28.2 %
AWL Words	12.75 %
Off-List Words	59.05 %

To shed light on the nature of high-frequency words in the corpus, we further examined the top 100, 200, and 300 high-frequency word-forms. Table 4.3 reveals various proportions of K1 words, K2 words, AWL words and off-list words that constitute the first 100, 200, and 300 words in the corpus. As shown in the table, K1 and K2 in total account for 93.46% of the first 100 words in the corpus, while the AWL and Off-List words constitute less than 10%. For the top 200 words and 300 words, however, the coverage of K1 and K2 words gradually decreases while the proportions of the AWL and off-List words increase. In terms of the first 300 words, the phenomenon is even more obvious since the AWL and off-list words in total cover more than 10% of the whole corpus. The results not only reflect the nature of the corpus but imply that the vocabulary learning goals should be adjusted in accordance with the nature of texts.

Table 4.3 *The Composition of Top 100, 200, and 300 High-Frequency Words*

	top 100 words	top 200 words	top 300 words
K1 Words	90.95%	88.47%	84.64 %
K2 Words	2.51 %	1.75 %	2.67 %
K1+K2	93.46 %	90.22 %	87.31 %
AWL Words	4.02 %	6.77 %	8.85 %
Off-List Words	2.51 %	3.01 %	3.84 %

Word Frequency Profiles

As indicated in Sinclair (1991: 30), statistical information provided by word frequency profiles can serve as a guide to the way words are distributed in a text. Two word frequency profiles, therefore, were compiled in this study to show the distribution of words in the CS corpus. Table 4.4 shows the add-up percentages of the top ten high-frequency word-forms in the corpus in terms of both word-form count and vocabulary count. Here, the word-forms refer to the number of running words while vocabulary is the number of different word-forms in the corpus. As can be observed in the table, the top ten high-frequency word-forms account for only 0.04% of the vocabulary whereas they constitute 24.87% of the total running words in the corpus. On the other hand, we studied the word-forms from the other way round, focusing on words occurring 1 to 10 times in the corpus. As revealed in Table 4.5, the total number of these word forms is 12,108 and they constitute 51.23% of all vocabulary in the corpus; however, they account for only 8.31% of the total running words. The implication, thus, is that about half of the word-forms in the corpus recur less than 10 times, while a small number of word-forms recur very often and constitute a large proportion of the corpus. Combining these statistics with those observed in Table 4.4 and Table 4.5,we may infer that a small number of word-forms with very high occurrence rate are the GSL words. Although most of them are function words, they constitute nearly 1/4 of the total running words. On the other hand, more than half of the word forms (or types) have very low occurrence rates. Most of them might be the AWL or off-list words since the more we go down the word frequency list, the higher coverage the AWL and off-list words have.

Table 4.4 *Word Frequency Profile for the Whole Corpus (1)*

Word-Form Count	Number	Vocabulary Total	Percentage of Vocabulary	Word-Form Total (/375978)	Percentage of Text
26,876	1	1	0.004	26,876	7.15
14,394	1	2	0.008	41,270	10.9
8,848	1	3	0.013	50,118	13.33
8,585	1	4	0.017	58,703	15.61
8,284	1	5	0.021	66,987	17.81
7,712	1	6	0.025	74,699	19.87
6,934	1	7	0.030	81,633	21.71
4,308	1	8	0.034	85,941	22.86
4,047	1	9	0.038	89,988	23.93
3,513	1	10	0.042	93,501	24.87

Table 4.5 *Word Frequency Profile for the Whole Corpus (2)*

Word-Form Count	Number	Vocabulary Total	Percentage of Vocabulary (/23635)	Word-Form Total (/375978)	Percentage of Text
1	5,928	5,928	25.08	5,928	1.58
2	2,126	8,054	34.07	10,180	2.71
3	1,214	9,268	39.21	13,822	3.68
4	748	10,016	42.38	16,814	4.47
5	560	10,576	44.75	19,614	5.22
6	444	11,020	46.63	22,278	5.93
7	352	11,372	48.12	24,742	6.58
8	297	11,669	49.37	27,118	7.21
9	252	11,921	50.44	29,386	7.82
10	187	12,108	51.23	31,256	8.31

Comparison of the 50 Most Frequent Content Word Forms in CS Corpus with a TESOL Corpus and the BNC Written

To further shed light on the distinguished nature of words in CS, we took the 50 most frequent content words from the word frequency list, and compared them with those in a TESOL Corpus (Liou, et. al, 2005) and the BNC Written. The former is a corpus which consists of also journal articles but in a different field while the latter is a general corpus comprising various genres. Since the sizes of the three corpora are different, we thus compared them based on the ratio of the frequency of the word forms to the total tokens of each corpus. As can be seen in Table 4.6, a lot of high-frequency content word forms in the CS corpus are of scientific register, such as *data, algorithm, system, image,* etc. Furthermore, it could be found that most words that are frequently used in the CS corpus are rather infrequent in the TESOL Corpus and the BNC Written except very general words like *have, time, and performance.* The result corresponds to that of Mudraya (2006), suggesting that words frequently used in one discipline may be infrequent in other disciplines or for general purposes.

Pedagogically, it implies that field-specific words deserve more attention in ESP or EAP classrooms.

Table 4.6 *The 50 Most Frequent Content Word Forms in CS Corpus Compared against a TESOL Corpus and the BNC Written*

	Frequency in Corpus			% in Corpus		
Word	CS	TESOL	BNC W.	CS W.	TESOL	BNC
each	1,174	779	539	0.30	0.05	0.05
have	1,104	1,234	4,416	0.28	0.08	0.44
data	1,102	594	197	0.28	0.04	0.02
used	905	892	497	0.23	0.06	0.05
number	860	455	488	0.22	0.03	0.05
set	818	186	350	0.21	0.01	0.04
time	784	920	1,509	0.20	0.06	0.15
using	715	394	257	0.18	0.03	0.03
algorithm	705	2	0*	0.18	0.00013	0*
based	699	556	199	0.18	0.04	0.02
results	676	696	159	0.17	0.05	0.02
Fig	630	110	178	0.16	0.01	0.02
system	627	144	476	0.16	0.01	0.05
performance	592	218	1,065	0.15	0.01	0.11
process	584	305	231	0.15	0.02	0.02
use	572	1,349	303	0.15	0.09	0.03
word(words)	566	710	249	0.15	0.05	0.02
different	556	485	482	0.14	0.03	005
image	549	5	81	0.14	0.00033	0.01
node	539	2	0*	0.14	0.00013	0*

* The frequency and range of the two words are zero because the frequency of them could not be found from the available source (Leech, Rayson,& Wilson, 2001).

CS Wordlist

It has been shown that technical vocabulary accounts for a great deal of types in the CS corpus, suggesting the vocabulary learning goal of learners in CS could be directed towards specialized vocabulary other than GSL or AWL. To date, many wordlists, such as the two well-known GSL and AWL, have been developed on the basis of frequency as well as coverage. However, seldom did researchers take the specific needs of learners into consideration. We, thus, are interested in developing a field-specific wordlist.

The CS wordlist was developed based on the procedure described in Chap 3. In total, it consists of 1402 word families. Following the procedure, words in the GSL were not included in the list; nevertheless, a few GSL words such as *performance, framework, mapping, tree, block, branch,* etc. were retained since they have technical meanings that are different from their definitions in general texts. For example, the word *tree* in general English refers to a woody plant having a single main stem while it is used to represent a type of data structure in which each element is attached to one or more elements directly beneath it in the field of CS. Since these words might cause problems for learners, we think they should be included in our specialized wordlist.

The Introduction Section of Research Articles

The Introduction Section of RAs serves as an overview of the study in concern. It introduces the research topic for readers, reviews existing research, gives the rationale and purpose of the study, and indicates the significance or application of the results. In this section, we first present basic statistics of the Introduction section of RAs in our corpus. Again, we also analyze the proportions of K1 words, K2 words, AWL words, and off-list words in this subcorpus of RA Introduction. Then, the major and optional moves in the Introduction are identified, using the self-developed coding scheme. An analysis of move combinations (or move patterns) was conducted to reveal the common move patterns employed in RA Introduction. Finally, we extended our exploration from the level of individual words to lexical bundles since each register has its own bundles that reflect the discourse function of it. To explore how rhetorical functions of RA Introduction were realized through vocabulary, lexical bundles were thus examined.

Basic Statistics of the Introduction Sub-corpus

Table 4.7 shows the coverage of K1 words and K2 words, the AWL, and the off-list words in our Introduction sub-corpus. As shown in the table, the first 1000 words and the second 1000 words in total cover 36.82% of the corpus. On the other hand, the AWL and the off-list words account for 19.22% and 43.97% of the sub-corpus, respectively. It is worth noting that AWL words cover 19.22% of the CS Introduction; this percentage is much higher than that of the whole CS corpus, and of course the 10% of general academic text indicated by Coxhead (2000). However, off-list words, which can be generally referred to as technical vocabulary, have a lower percentage in the Introduction sub-corpus than that in the whole CS corpus. This is probably because the Introduction section focuses on more general descriptions and discussions of the research topic and research questions without going into specific research procedures, data collection and analysis, and results which the other sections are aimed at. A larger amount of general academic vocabulary, rather than technical (or specialized) vocabulary, therefore, is used in the Introduction section.

Table 4.7 *Coverage of the texts by the various types of vocabulary in Introduction*

Word level	Types	Tokens	Percent
K1 Words(1-1000)	736	2,076	27.76%
K2 Words(1001-2000)	328	625	9.06%
K1+K2			36.82%
AWL Words	474	1,259	19.22%
Off-List Words	1,376	3,129	43.97%

Major and Optional Moves in Introduction

As indicated in Chap 3, move analysis was conducted, using a self-developed coding scheme. Then the moves were tagged on all electronic text samples in the corpus. The moves represent specific rhetorical functions in RA Introduction in the field of computer science. Since not all the moves occur in the Introduction section of all RAs, we want to identify which moves are obligatory while which are optional for pedagogical purposes. Both frequency and range are considered in identifying major and optional moves because moves with high frequency alone might result from the idiosyncrasy of a single writer. The consideration of frequency along with the distribution of each move (range) in the 60 RAs provide a solid foundation on the determination of major and optional moves. With the freeware AntConc 3.0, both the frequency and the range of each move were then calculated.

As shown in Table 4.8, 714 moves are identified. Among them, IL (literature review), IM (methods), IP (purposes), IB (background information), IG (gap), and IO (organization) rank 1 to 6 in terms of frequency. These six moves not only have higher frequencies than the other six moves, but they also have a distribution rate of more than 50% (in fact, 65%), or a range of more than 30 (in fact, 39) out of 60 RAs. As a result, they are categorized as major moves in the corpus. On the other hand, the other six moves with fewer occurrences and distribution rates less than 50% are categorized as optional moves, namely, IV (values), IC (conclusion), IJ (justification), IF (tables or figures), IR (results), and IQ (research questions). This categorization suggests that a number of rhetorical functions are essential and thus occur frequently in RA Introduction; thus pedagogically, RA writers should pay more attention to them. The optional moves do not occur as frequently as the major moves, but they represent rhetorical functions that may still occur in specific RAs. Therefore, RA writers can be informed of these possible rhetorical functions specific to the field of CS.

Table 4.8 *Frequency and Range of Major Moves and Optional Moves*

	Moves	Total occurrences of each move		Range of each move		
		Number	Ranking	Range (N=60)	%	Ranking
Major Moves	IL	144	1	54	90.0	4
	IM	143	2	57	95.0	1
	IP	87	3	55	91.7	3
	IB	81	4	56	93.3	2
	IG	71	5	39	65.0	6
	IO	58	6	47	78.3	5
Optional Moves	IV	36	7	27	45.0	7
	IC	25	8	16	26.7	9
	IJ	25	8	19	31.7	8
	IF	22	10	16	26.7	9
	IR	16	11	13	21.7	11
	IQ	6	12	4	6.7	12
	Total	714				

Analysis of Common Move Patterns in Introduction

In addition to the major and optional moves, we further examined how these moves are organized in text. In other words, the common move patterns were investigated. As indicated in Chap.3, with the help of AntConc, both the clusters of moves and their distribution frequency were identified. Although the results of these move combinations seem not salient in terms of frequency as a result of the small size of our corpus, the move patterns show possible combinations and sequences of moves in the Introduction, hence, they can provide useful information for pedagogical use. The procedure for selecting common move patterns is as follows: (1) we selected individual moves with a range rate over 50%, that is, occurring in more than 30 RAs (major moves) (2) 3-move and 4-move patterns of the above moves with a frequency of ten or higher were selected. On the basis of the criteria, Table 4.9 shows the common move patterns identified in the Introductions of all RAs in our corpus.

Table 4.9 *Common Move Patterns of Introduction*

Common Move Patterns	Frequency
4-move patterns	
IL-IM-IL-IM	12
IM-IL-IM-IL	11
3-move patterns	
IL-IM-IL	19
IL-IG-IL	12
IL-IP-IM	10
IM-IL-IM	23
IB-IL-IG	10

Two 4-move patterns were found: IL-IM-IL-IM and IM-IL-IM-IL. It can be found that in Table 4.9, IL can be followed by IM, IG, and IP. The move combination of IL-IM appears to occur frequently in the Introduction of RAs in CS. It is not only the highest frequently used move combination in the 3-move patterns, but demonstrates cyclical nature in 4-move patterns. This pattern reflects the specific rhetorical conventions in the Introduction section in CS RAs in which the Introduction of a research method is combined with reference to studies adopting this research method. Since there may be different methods proposed by different studies cited, the IM-IL or IL-IM often occur in cycles. This explains why in 3-move patterns, we also found IM-IL-IM and IL-IM-IL. Following is an example of IL-IM-IL.

> //IL// Metadiscourse is ubiquitous in scientific writing. Hyland(1998) found a metadiscourse phrase on average after every 15 words in running text. //IM// A large proportion of scientific metadiscourse is conventionalized, particularly in the experimental sciences, and particularly in the methodology or result section

> (e.g. we present original work….., or An ANOVA analysis revealed a marginal interaction/ a main effect of…) //IL// Swales(1990) lists many such fixed phrases as co-occurring with the moves of his CARS model. They are useful indicators of overall importance (Pollock and Zamora, 1975)….

In total, five 3-move patterns were identified. Except the two related to the cycle of IM-IL, they are IL-IG-IL, IL-IP-IM, and IB-IL-IG. It can be observed that the move sequence IL followed by IG serves as another frequently used move pattern in the Introduction. As we examined occurrences of this pattern in the concordancer, it usually opens with a negative sentence connector as a signal, indicating the insufficiency or weakness of previous research. In so doing, the author not only indicates a gap in previous research but establishes a niche for the current study. The rhetorical function of this move pattern seems to be common and applicable to the Introduction of any field since it is one of the rhetorical functions proposed in Swales' CARS model. An example of this move pattern is given below:

> //IL// See Ni and McKinley [25] for a detailed explanation of wormhole routing. //IG// The primary drawback to wormhole routing is the contention that can occur, even with moderate traffic, which leads to higher message latency. Whenever a message is unable to proceed due to contention, the header and data flits are not removes from the network. Instead, the message holds all the channels it currently occupies. Although each channel is released after the entire message traverses that channel, a long message that occupied several channel buffers can block many messages during transmission. These blocked messages can in turn block other messages, which further increase the message latency. //IL// A cost-effective method of reducing message latency, proposed by Dally

[7], is to allow multiple virtual channels to share the same physical channel.

As can be seen in Table 4.9, the move pattern of IL-IG could also be preceded by the move IB. This pattern also well reflects one of the writing conventions in the Introduction section where background knowledge about the research topic to be studied is provided, followed by a review of studies already conducted by other researchers and finally a niche is established by indicating a gap. The segment of IB-IL-IG is thus formed.

IL can also be followed by IP, outlining the purpose of a study after reviewing the literature for readers. Following the statement of purpose, then, research methods to be used are provided. From the occurrences of each move and common move patterns, it is noted that IM seems to be a significant move specific to the Introduction of RAs in CS, since according to Swales (1990), research method is not an indispensable move in the Introduction.

Lexical Bundles in Introduction

In this section, we examined the lexical bundles frequently used in the RA Introduction. Lexical bundles are pre-fabricated or fixed expressions that can be found in nearly all registers. The use of lexical bundles unique to particular registers not only signifies competent language use within a register but demonstrates the familiarity of the conventions of that register (Cortes, 2004). Since lexical bundles are able to characterize the nature of specific text types, we, thus, examined the five-word, four-word and three-word lexical bundles, respectively, in our CS Introduction sub-corpus, focusing on the discourse functions that these bundles may perform in the Introduction section.

The identification of lexical bundles was mainly data-driven; that is, a computer program capable of retrieving the bundles and counting their frequencies was employed. However, not all bundles identified by the computer are meaningful. Therefore, we need to decide on a cut-off point, in this case, a frequency number used as a criterion for selecting useful lexical bundles, as suggested by other studies on lexical bundles. For example, Biber et al. (1999) used a frequency cut-off of 40 times per million words as a criterion in the selection of bundles. We intended to adopt a more rigorous criterion. Therefore, a frequency of 3 per ten thousand words was applied to our selection of lexical bundles. A second criterion is the consideration of the nature and function of the bundles. Once the bundles were identified, two raters went through each bundle together and discussed the nature and function of each bundle. Thus, all the bundles selected had inter-rater reliability.

Since the purpose of this analysis was to investigate how lexical bundles were associated with the communicative purposes of the Introduction, the selected bundles were then categorized on the basis of the discourse functions they perform. It was found that both four-word and three-word bundles could be categorized into two groups – bundles that reflect specific rhetorical functions of RA Introduction and bundles that are used for general academic purposes, while all five-word bundles belong to only the first group. Bundles reflecting the rhetorical functions of RA Introduction were then linked to their corresponding moves. Thus, a linkage between bundles and moves in the Introduction can be shown (See Table 4.10, 4.11, 4.12).

Bundles in the second group were further categorized into (1) stance expressions, (2) discourse organizers, and (3) referential expressions based on the taxonomy proposed by previous studies (Bibier & Barbieri, 2007; Biber, Conrad & Cortes, 2004; Cortes, 2004). This taxonomy was firstly developed by Biber et al. (1999), but later adopted by many researchers in the study of discourse functions lexical bundles perform in context. The following sub-sections describe in detail the five-word, four-word and three-word bundles retrieved from our Introduction sub-corpus as well as their discourse functions.

Five-word lexical bundles in Introduction.

Table 4.10 shows the list of the selected five-word lexical bundles grouped on the basis of their rhetorical functions. As can be seen in the table, most of the five-word bundles mainly serve the rhetorical function of stating the purpose of a study or outlining the local organization of the Introduction section or global organization of the whole research article. Among the bundles characterizing IP, the most frequent expression is *in this paper, we present/propose/focus* with a variety of verbs to pinpoint the major purpose of a study. An example is shown as follows:

[4.1] /IP/ *In this paper, we propose* a deterministic matching method for verifying both isomorphism and subgraph isomorphism.

In addition to IP, lexical bundles characterizing the rhetorical function of IO also occur frequently. Among these bundles, two major patterns describing the organization of RA are identified. One is *the paper/this paper is organized as follows* and the other is *in section X/ in the next section, we present/describe*. The following are two of the examples:

[4.2] /IO/ *This paper is organized as follows.* Section 2 describes the fundamental of hidden Markov models. Section 3 details the steps of preprocessing, segmentation and feature extraction.

[4.3] /IO/ *In section 5, we present* experimental results for both performance

> and area using our modeling approach. *In section 6, we describe* related work and conclude in section 7.

Another five-word bundle that has a frequency higher than 3 is *the basic idea is to*. Since we were uncertain about the corresponding rhetorical function this bundle may perform, we examined the various contexts it occurs. It was found that four instances of this bundle occur in the same move – IM, serving as a general indication of the concept that lies behind the method to be used. The bundle also often occurs at the beginning of IM, directly followed by a literature review of the methods used in previous research or it follows a purpose statement to delineate how the purpose of a study could be carried out through a specific method. For example,

> [4.4] /IP/ We introduce two new measures of classification complexity called…/IM/ *The basic idea is to* complete these measures cumulatively by partitioning data space at various resolutions where each resolution is defined by the number of partitions per feature.

Table 4.10 *Five-Word Lexical Bundles in Introduction*

	Rhetorical Functions (Moves)	Bundles
Bundles that reflect communicative purposes of RA Introduction	IP	in this paper, we present (5) in this paper, we propose (5) in this paper, we focus (4)

	IO	paper is organized as follows (12) This paper is organized as (7) the paper is organized as (5) of this paper is organized (6) the rest of this paper (5) is organized as follows: section (5) in the next section, we (5) in section 5, we present (3) in section 4, we describe (3) are given in section 5 (3)
	IM	the basic idea is to (4)

Four-word lexical bundles in Introduction

With respect to four-word lexical bundles, it could be observed that the frequency of four-word bundles was higher than that of the five-word bundles. In addition, there were more four-word bundles than five-word bundles. The nature of the four-word lexical bundles was also different from that of the five-word bundles in that their rhetorical functions were not limited to specific generic move mapping; in fact, a number of four-word bundles identified are geared towards general academic purposes. These four-word bundles, as a result, were categorized into two groups in terms of the discourse functions they perform – one containing bundles that reflect the rhetorical functions of RA Introduction and the other containing bundles for general academic use.

Table 4.11 shows four-word bundles characterizing the rhetorical functions of IO, IP, IL/IB, and IF in the Introduction section. A variety of lexical bundles are used to indicate the organization of RA (that is, linked to the move of IO), implying that the realization of this rhetorical function is highly conventionalized. It can also be observed that the four-word bundles of IO and IP, as a matter of fact, are just part of the five-word bundles. Other four-word bundles can be related to the rhetorical functions or moves of IL, IB, or IF. For example, the bundle of *in the field of* serves to refer to a research field or a topic, as shown below. It is thus a bundle that can occur in either IL or IB.

> [4.5] //IL// Recent developments *in the field of* learning in structured domains (e.g., [7], [8]) offer new unexplored and promising research domains, some of which are reviewed in the following.

Another four-word bundle which performs the rhetorical function of reviewing literature is *have been proposed in*. This bundle is frequently preceded by a research topic, domain or method and followed by the cited literature. An example is given as follows:

> [4.6] //IL// Several methods *have been proposed in* the literature to perform edge linking, or edge aggregation [11], [36], [45].

Still another four-word bundle is *as shown in Fig*, representing the rhetorical function of IF, that is, referring to a table or figure. An example is given below:

> [4.7] //IM+IF// Our prototype implementation is geared toward two-dimensional meshes, *as shown in Fig. 1*, such topologies have been widely used as the interconnection network for a variety of commercial parallel machines.

In the example, it can be observed that this bundle occurs in a context that actually combines two moves – IM and IF.

As indicated earlier, the other category of the four-word bundles is lexical bundles used for general academic purposes. For these bundles, we further grouped them into subcategories on the basis of three discourse functions – stance bundles, discourse organizers and referential bundles. As shown in Table 4.11, among the 17 four-word lexical bundles for general academic purposes, three are stance bundles, two are discourse organizers, while twelve of them belong to referential bundles. It is worth noting that the result is consistent with that of Biber & Barbieri's (2007) in which referential bundles were found dominant particularly in academic writing. An example of one of the stance bundles *plays an important role* is given below, which performs the discourse function of claiming the centrality of a research topic:

> [4.8] As these applications grow in size and complexity, parallel processing *plays an important role* in satisfying the large computational demands.

Two other bundles *is based on the* and *on the basis of* are in the subgroup of referential bundles, frequently used to support a proposition or argument. For example,

> [4.9] Barzilay, McKeown, and Elhadad (1999) introduce the concept of information fusion, which *is based on the* identification of re-current descriptions of the same events in news articles.

> [4.10] The rhetorical status of a sentence is determined *on the basis of* the global context of the paper.

Table 4.11 *Four-Word Lexical Bundles in Introduction*

	Rhetorical Function (Move)	Bundles
Bundles that reflect rhetorical functions of RA Introduction	IO	is organized as follows: (13) this paper is organized (7) the paper is organized (6)
		in section 3, we (10) in section 4, we (7) in section 5, we (8) in the next section (5) in this section, we (4)
		are given in section (6) are presented in section (4)
	IP	in this paper, we (33)
	IL/IB	in the field of (7)
	IL	have been proposed in (5)
	IF	as shown in Fig. (4)
General bundles	Stance Bundles	can be used to (6) it is possible to (6) plays an important role (4)
	Discourse Organizers	on the other hand (16) in the presence of (10)
	Referential Bundles	is based on the (12) in the context of (7) on the basis of (6) a small number of (6) a wide range of (5) as a sequence of (5) can be viewed as (5) a wide variety of (4) at the expense of (4) can be found in (4) is an example of (4) the performance of the (4) for the purposes of (4)

Three-word lexical bundles in Introduction

Our examination of lexical bundles in RA Introduction started from the five-word bundles to the three-word bundles. The three-word lexical bundles were sorted and categorized with the same procedure. Table 4.12 illustrates the distribution of the three-word lexical bundles used in the Introduction section of RAs in CS. They are categorized again into two categories – bundles reflecting the rhetorical functions of RA and bundles for general academic purposes. Compared with five-word and four-word bundles, three-word bundles seem not so clearly related to the rhetorical functions specific to RA Introduction. Rather, most of them look like bundles for general academic purposes or bundles related to the genre of RA as a whole.

Some bundles by themselves do not clearly reveal the rhetorical functions they perform. To properly determine the discourse functions of these bundles, we went back to concordance listings to see how they were used in context. Take the bundles of *an overview of* and *is defined as* for instance. It seems that both bundles can be used for general purposes but they can also be related to specific moves. The examination of their discourse contexts in the corpus shows that the bundle of *an overview of* is mostly used for the rhetorical function of IO, that is, indicating the location where overview of certain concepts/ tasks/ systems is given. An example is given below:

> [4.11] //IO// The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 gives *an overview of* the CSR system. Section 4 introduces the image preprocessing technique.

The function of the bundle of *is defined as* in RA Introduction is also unclear at first. After examining the context, we found it occurs more often in IM to provide an explanation of a professional term related to research method.

Also, *in addition to* and *in contrast to* are frequently used 3-word bundles for general academic purposes. Generally, these two bundles are not only employed as a transition to maintain the smooth flow of a text but used to relate ideas that precede or follow to what is discussed in the context. Examples of the two bundles are shown below:

[4.12] *In addition to* determining the boundary-based complexity of data using a nearest neighbor approach, they also implicitly measure data compactness and distance between distribution under a unified work.

[4.13] *In contrast to* the primarily qualitative methodologies that characterized research in the 1980's, purely quantitative evaluation methods now pervade all aspects of the research and development process in many areas of NLP.

Table 4.12 *Three-Word Lexical Bundles in Introduction*

	Rhetorical Functions (Moves)	Bundles
Bundles that reflect communicative purposes of RA Introduction.	IP	in this paper (37) in the paper (9) have been proposed (13)
	IO	is organized as (15) in section 2/3/4/5 (40) the next section (5) the result of (9) an overview of (6)
	IF	shown in fig. (9) as shown/described in fig. (10)
	IL	in the literature (6)
	IM	is defined as (10)

		our approach is (6)
General bundles	Stance Bundles	be used to (15)
	Discourse Organizers	in addition to (11) in contrast to (10) as opposed to (9) as a result (7) in spite of (6)
	Referential Bundles	is based on (28) in terms of (24) a variety of (15) a sequence of (10) the rest of (9) can be found (6) can be viewed (6) is compared with (5)

Linking Vocabulary and Lexical Bundles with Moves

After we identified the 5-word, 4-word and 3-word lexical bundles used in the Introduction, we were also interested in examining how moves are realized through individual words, and if there was any move-signaling bundle or language use specifically in relation to the rhetorical functions of moves.

To identify words in relation to specific moves, a frequency list was first compiled for each small sub-corpus of move. Although there have been quite a number of genre analysis studies on RAs and specifically on the Introduction section, there has been little research on the generic nature of vocabulary. As argued in Chapter 1, we strongly suspect there are words and phrases which are conventionally and frequently used to perform certain rhetorical functions in a specific genre like RAs. If we can identify these words and phrases, they can be of great pedagogical value. Further, a link between lexis and discourse structure can be established. In the present study, we take a cut-off point of 3 times per ten thousand words in the selection of bundles, as indicated earlier. Since the size of the subcorpus of each move is not big, we examined only 4-word and 3-word lexical bundles. In the following subsections, move-related vocabulary as well as lexical bundles in specific moves are reported (meaningful bundles could not be found in a couple of moves). They are discussed on the basis of word frequency in the move in concern and possible move association of the bundles.

IL

According to Swales' CARS model, reviewing previous research is an obligatory step in Move 1. From the results of our move analysis, literature review is also one of the major moves in the Introduction section of RAs in CS.

Since the rhetorical function of literature review is to discuss previous studies and their findings, reporting verbs are often employed to show stances of the authors. Based on the word frequency list of IL move corpus, reporting verbs found in the top 200 words of IL were *proposed, described, considered, and obtained.* Since reporting verbs carry the writer's evaluation or degree of commitment towards the cited study (Thompson and Ye, 1991; Swales, 1990), we thus were interested to know how RA writers in the field of CS express their stances towards the cited work. We first categorized the reporting verbs into *evaluative*, *tentative*, and *neutral*. Evaluative reporting verbs carry the positive or negative evaluation of a writer towards the study he/she cited. Reporting verbs such as *reduce*, *improve*, or *inspire* are of this category. Tentative reporting verbs demonstrate the author's tentative attitude towards the cited work. In other words, the author makes plausible interpretation towards the cited work. Reporting verbs of this category include *suggest, imply,* and *speculate*, etc. Neutral reporting verbs, on the other hand, focus on reporting results and findings of the cited work without carrying commitment from authors, such as *present, propose,* or *report*, etc. Of the 141 occurrences of IL using reporting verbs in our corpus, 22 are evaluative, 3 are tentative, while 116 are neutral. The results suggest that while reviewing previous studies, writers in CS tend to report the cited studies in a general and non-evaluative way rather than providing subjective interpretation towards the cited work.

Observation of the frequency analysis of 4-word and 3-word bundles of IL did not reveal any significant word combinations closely associated with the rhetorical functions of IL. This suggests that although reporting verbs are often employed in IL, there might be no conventionalized fixed expressions to report or review previous studies.

IP

The indication of research purpose is a major move in Introduction. The rhetorical function of IP is to fill a created gap by announcing what the study intends to do. An examination of the top 200 words of IP did not reveal any salient vocabulary specifically related to the rhetorical function of this move.

However, results obtained from the lexical bundles of this move were fruitful. Table 4.13 shows the five-word and four-word lexical bundles with a frequency cut-off of 3. As revealed in the table, the five-word strings alone contain 455 words, accounting for 11.38% of the IP corpus. Given the small size of the move corpus, this coverage of the five-word strings is considered high, implying the highly conventionalized structures used in IP.

Table 4.14 reveals two commonly used bundles in this move; they are *In this paper, we+V+N* and *This paper V+N*. The frequency of the former is especially high. An examination of the verbs in these two patterns revealed the kinds of verbs RA writers often use to introduce or present their own study. These nouns usually represent the proposed research method, system, scheme, model, etc. of the study.

Therefore, the list of verbs in these two proposed patterns or structures can be pedagogically useful, serving as a reference list for student writers of RAs. They serve either to introduce methods/measures or to present a system, approach, model, or algorithm for the current problem.

Table 4.13 *Five-Word and Four-Word Lexical Bundles in IP*

5-wd strings		4-wd strings	
5-wd strings: 3,992		4-wd strings: 3,993	
Words: 455 (11.38% of tot)		Words: 472 (11.81% of tot)	
001. **[5]**	IN THIS PAPER, WE PROPOSE	001. **[23]**	IN THIS PAPER, WE
002. **[5]**	IN THIS PAPER, WE PRESENT	002. **[5]**	THIS PAPER, WE PRESENT
003. **[3]**	IN THIS PAPER, WE FOCUS	003. **[5]**	THIS PAPER, WE PROPOSE
004. **[3]**	THIS PAPER, WE PROPOSE A	004. **[3]**	THIS PAPER PRESENTS A
		005. **[3]**	PAPER, WE PROPOSE A
		006. **[3]**	OF A SET OF
		007. **[3]**	IN THIS WORK, WE
		008. **[3]**	THIS PAPER, WE FOCUS

Table 4.14 *Two Commonly Used Lexical Bundles in the IP*

Commonly used patterns	V	N
In this paper, we (20)	focus on (5)	sorting methods
	present (5)	……….MDS systems
	propose (4)	a single case study
	address (1)	a case study
	apply (1)	results
	build on (1)	a matching method
	employ (1)	mechanism
	introduce (1)	problem
	use (1)	scheme
		framework
		approach
		algorithm
		measures
		model
		method
		criterion

This paper (10)	describe (3) present (3) examine (1) is aimed at (1) propose (1) set out (1)	issues questions approach scheme

IB

At the onset of most introductions of RA, the first step is to establish a context to situate the current research in a wider field of research either by providing background information or by claiming the significance of the field or study (Swales, 1990; Weissberg & Buker, 1990). From the observation of the frequency list of IB move corpus, many of the words in the top 100 words are nouns such as *applications, systems, networks, approach, classification, and performance*. In fact, these words are closely related to the rhetorical functions of IB for they serve to introduce research topics. Moreover, they represent common topics in the research field of computer science. With the use of these topic-establishing words, research contexts are thus created.

An examination of the lexical bundles of IB reveals that many of the high frequency bundles are for general academic purposes instead of bundles specifically related to the rhetorical function of IB. It might be that an introduction of background information can cover a wide variety of vocabulary use and phraseology. As a result, bundles of general purposes are common in this move.

Although no move-signaling bundles were found in the move corpus of IB, we observed a grammatical structure frequently employed to signal the rhetorical function of IB. The structure of *S+ have/has been +V* is often employed in IB, reflecting the rhetorical functions of IB to introduce what has been done from the past to the present in a research field or topic.

Table 4.15 *the Subjects and Verbs Used in the S+ have/has been +V*

S		V
studies problems approaches projects tasks applications features strides machines rules researchers	have been (11)	devoted to overcome used replying on developed used made looking for
N		V
research effort time requirement of	has been (9)	devoted to studied

IG

In Swales' CARS model, indicating a gap is a preparation step for introducing one's own study. In other words, it is a move to establish a niche which is later to be filled by the author's own study. Ways such as describing an inadequate aspect of previous studies, pinpointing an unresolved conflict, and raising a new research question have been used to indicate what is missing from previous research or what can be extended. Although IG is one of the major moves in our corpus, all of its 5-word, 4-word and 3-word lexical bundles have low frequency. This probably resulted from the small size of the corpus and the various possible ways for gap statements. Thus, no specific lexical bundles closely related to the rhetorical functions of IG were found. We observed, however, concessive sentence-connectors such as *however*, *but*, or *although* have very high frequencies. As shown in Table 4.16, *however* seems to be the most preferred word which performs the function of a transition from previous research to the author’s own study.

Table 4.16 *The Frequency and Rank of However, But, and Although in IG*

	Frequency	Rank
however	29	16
but	18	21
although	5	112

Although words such as *but* and *although* are semantically similar to *however*, the use of one instead of another may lead to different syntactic structures of IG statements. Following are examples of the three words:

[4.14] *However*, most of these efforts do *not* study the influence on the energy consumption of the other system components and even fewer consider the integrated impact of the hardware and software optimizations. It is important to evaluate the influence of optimizations on the overall system energy savings and the power distribution across different components of the system. Such a study…..

[4.15] In all cases, significant benefits have been reported, *but* the absolute figures are not comparable due to very different architectural assumptions.

[4.16] It turns out that, *although* domain ontologies are recognized as crucial resources for the semantic web, in practice they are not available and when available, they are ready used outside specific research environments.

To shed light on the possible reasons for the different uses of the three words similar in meaning, we went back to the concordance listings to see how they were used in context within the IG move corpus. In the examples [4.14] and [4.15], adversative connectors – *however* and *but* are mainly used to directly indicate the insufficiency or limitations of previous research. On the other hand, the use of *although* in [4.16] seems to focus on comparison and contrast, indicating what has been accomplished in previous research but stressing what can still be modified, added, or extended. We also observed that the three words often co-occurred with negative expressions such as *not*, *few,* or *little*.

IO

To indicate the organization of research articles at the end of the Introduction section seems to be a convention of RAs in CS. From the top 200 word frequency list of IO, it was observed that verbs with high frequencies were those describing or reporting the content of the various sections following the Introduction section such as *describe, present, discuss and propose*. Other words closely related to the rhetorical function of IO were *section* and *organize*.

Although the size of the IO corpus is small, results of lexical bundles are, to our surprise, quite insightful. Table 4.17 shows part of the five-word, four-word and three-word lexical bundles of IO. As can be seen in Table 4.17, the frequencies of the five-word, four-word and three-word lexical bundles, compared with other moves, are high despite the small size of the IO corpus. The five-word lexical bundles alone cover 14.78% of the corpus. The result is unusual since the corpus has only 5532 running words in total. These results suggest the highly conventionalized nature of this move.

Table 4.17 *Five-word, Four-word and Three-word Lexical Bundles of IO*

5-wd strings 5-wd strings: 5,745 Words: 850 (14.78% of tot)	4-wd strings 4-wd strings: 5,746 Words: 892 (15.51% of tot)	3-wd strings 3-wd strings: 5,747 Words: 1053 (18.31% of tot)
001. **[13]** PAPER IS ORGANIZED AS FOLLOWS:	001. **[14]** IS ORGANIZED AS FOLLOWS:	001. **[17]** IS ORGANIZED AS
002. **[9]** THIS PAPER IS ORGANIZED AS	002. **[14]** PAPER IS ORGANIZED AS	002. **[15]** PAPER IS ORGANIZED
003. **[7]** OF THIS PAPER IS ORGANIZED	003. **[11]** IN SECTION 3, WE	003. **[14]** ORGANIZED AS FOLLOWS:
004. **[6]** REST OF THIS PAPER IS	004. **[9]** THIS PAPER IS ORGANIZED	004. **[12]** THIS PAPER IS
005. **[6]** THE REST OF THIS PAPER	005. **[9]** OF THIS PAPER IS	005. **[12]** IN SECTION 3,
006. **[5]** AS FOLLOWS: IN SECTION 2,	006. **[9]** IN SECTION 5, WE	006. **[11]** IN SECTION 2,
007. **[5]** IN THE NEXT SECTION, WE	007. **[8]** IN SECTION 4, WE	007. **[11]** SECTION 3, WE
008. **[5]** IS ORGANIZED AS FOLLOWS: IN	008. **[7]** IN SECTION 2, WE	008. **[10]** IN SECTION 4,
009. **[5]** IS ORGANIZED AS FOLLOWS: SECTION	009. **[7]** THE REST OF THIS	009. **[10]** IN SECTION 5,
010. **[5]** THE PAPER IS ORGANIZED AS	010. **[6]** THE PAPER IS ORGANIZED	010. **[9]** OF THIS PAPER

Pedagogical Implications

In light of the results found in our study, a number of pedagogical implications are provided below. First, the corpus-based approach to the study of language use provides solid foundation for the course design or material development, particularly in ESP or EAP setting. Traditionally, EAP courses are designed on the basis of instructors' experience or out of intuition because of the limited research support. An EAP course designed on the basis of findings using corpus-based approach to academic genres not only target learners' needs but enhance the effectiveness of EAP instruction. Also, corpus-based analysis results could be incorporated with Computer Assisted Language Learning (CALL). For instance, teaching materials or learning tasks could be uploaded to online platforms, providing students with an access to learning materials without being constrained by time or space.

Since the role of vocabulary arouses more attention than it is used to be, numerous ways have been suggested in the teaching and learning vocabulary. To make the results of the current study more insightful, we propose a research-based pedagogical application in the teaching of academic vocabulary. The effectiveness of explicit and incidental learning of vocabulary has been a controversial issue and discussed widely. We, however, believe the combination of them is able to make the best use of our research results. Since we suspect that the findings of the study will mostly be known by specialists in the our discourse community, teachers in EAP classroom, thus, play a crucial role in bridging the gap and make the valuable research results accessible to EAP learners. The teaching materials of explicit teaching may be designed on the basis of research results. Instructors, for instance, may provide the CS wordlist, present the major and optional moves, and introduce move-signaling words and lexical bundles associated with rhetorical functions in RAs. This awareness-raising offers students a guideline in terms of what are the essential elements that should be involved in writing research articles. The explicit approach may further be facilitated by incidental learning in which learning tasks based on the content of explicit teaching are provided, aiming to activate the passive knowledge into active one through practicing. Tasks concerning acquiring high frequency words, identifying moves, and using common move patterns with move-signaling words to realize rhetorical functions are of help in increasing the autonomy of knowledge taught explicitly. In addition, concordancing tools able to retrieve language data from corpora may be provided. The access to concordancers offer learners the opportunity to explore language features on their own, such as the collocation or lexical bundles of vocabulary, and thus is likely to acquire word knowledge inductively. It is hoped that with the combination of explicit and implicit teaching, the acquisition of vocabulary could be more effective.

CHAPTER FIVE

DISCUSSIONS AND CONCLUSIONS

The present study explores vocabulary use in RAs, the Introduction section in particular, in relation to its communicative purposes or moves, using a data-driven, corpus-based approach. In this chapter, we first discuss and summarize the major findings of the study. Then, pedagogical implications as well as possible applications of the results are discussed. We finally provide a few directions for future research.

Summary of the Study

The study takes a genre-based, corpus-informed approach to analyze the use of vocabulary in RAs in the field of computer science. The corpus consists of 60 RAs selected from 3 major journals in computer science. All the text samples were analyzed both quantitatively and qualitatively. What distinguishes our study from most genre analysis studies or vocabulary studies is that we attempt to connect the two research fields; in other words, we aim to explore the generic nature of vocabulary. Specifically, we investigate move-signaling words in RAs since they can play an essential role in the pedagogy of academic writing, research paper writing in particular. Moreover, we approach the research questions mainly from a data-driven, probabilistic perspective. The quantitative analysis is solidly based on statistical measures or facilitated by NLP tools.

Data analysis focuses both on the whole RA corpus and the RA Introduction sub-corpus. To explore the nature of vocabulary used in the genre of RAs in computer science, the corpus is analyzed from different perspectives. Analysis of the word frequency list of the corpus shows the coverage of the GSL (28.20%), AWL (12.75%), and technical words (as generally represented by off-list words) (59.05%) used in

RAs. in the list. This suggests that general-purpose words constitute only a little more than one-fourth of all vocabulary in this genre, while academic and technical vocabulary account for almost three-fourths. Particularly, words of technical nature play an essential role in writing RAs in computer science. The percentages thus reflect the vocabulary register of both the genre and the field.

A second quantitative analysis is an examination of the top 100, 200, and 300 high-frequency words. It is found that the percentages of academic and technical words increase consistently in the order of 100, 200, and 300 word lists. For example, a lot more content words with field-specific meanings occur in the top 300 high frequency word list, such as *channel*, *output*, *hardware* etc. However, if we look at the proportions of these different categories of vocabulary from a different perspective, namely the coverage of the total running words (tokens), the results are totally different. This is demonstrated by the two word frequency profiles also compiled in the study. They reveal that actually a very small number of word-forms (the GSL words, and mostly function words) have very high occurrence rate, constituting nearly 1/4 of the whole corpus in terms of running words. On the other hand, low frequency words (those occurring less than 10 times) account for more than half of the vocabulary (or types) of the corpus. This phenomenon poses an interesting question about vocabulary learning: should learners of academic writing learn high-frequency words or low-frequency words? Although low-frequency words do not recur frequently, they form the wide range of vocabulary repertoire RA writers need to use, even merely once or twice. The pedagogical implication of this finding is thus significant.

To learn how vocabulary use may reflect the field of research, a simple comparison of the 50 most frequent content word forms among the CS corpus, a TESOL corpus, and the BNC Written is made. The result reveals that words frequently used in the CS corpus are rather infrequent in the TESOL Corpus or the BNC Written. We, thus, may draw the conclusion that the genre as well as subject content of a corpus may influence the results of corpus-based vocabulary analysis. In addition, vocabulary register characterized by field and genre should be taken into account in selecting target words for vocabulary learning. The field-specific words deserve more attention in EAP classrooms since they play an important role in the comprehension and production of academic texts.

As indicated earlier in this section, this study intends to investigate move-signaling words in RAs. We, therefore, narrow the focus down to one single section of RAs -- the Introduction. Again, statistical analysis reveals that the AWL words constitute an even higher percentage of the total vocabulary in the Introduction sub-corpus than that in the whole CS corpus. However, the proportion of the technical vocabulary (the off-list words) drops might result from the nature of Introduction in which general words are more used frequently.

To connect individual words with the rhetorical functions of RA Introduction, or to find move-signaling words, move analysis is conducted. A self-developed coding scheme is used to identify all the moves in the text samples. The major and optional moves as well as 3-move and 4-move patterns representing the information structures of RA Introduction are further identified based on frequency and range. Results indicate that among the six major moves, the combination of IL with IM, or vise versa, seems to be very common in both the 3-move and 4-move patterns, accounting for 4 instances among the 7 selected common move patterns. The other three common move patterns are IL-IG-IL, IL-IP-IM, and IB-IL-IG. Although the frequencies of these move patterns are not significantly high because of the small size of the corpus, they are pedagogically helpful since they exemplify how major/optional moves are used in combination in the Introduction, providing learners with useful information in writing this section.

Lexical bundles refer to fixed expressions that can be found in a register or genre. As they lexical bundles consistently in a specific text type, they can reveal its important discourse functions. We thus examined lexical bundles in the Introduction subcorpus and move-signaling words used to realized the rhetorical functions in the subcorpus of each move. In the Introduction subcorpus, we examined the five-word, four-word, and three-word lexical bundles, categorizing them into bundles that reflect the rhetorical functions of RAs and general academic bundles. It is found that the majority of the former bundles characterize the rhetorical functions of IP and IO, such as *in this paper, we present* or *paper is organized as follows*, while bundles reflecting referential stance such as *on the basis of* or *can be viewed as,* among the latter bundles, are the most frequently employed category. This implies that IP and IO are moves that are highly conventionalized in terms of language use, the realization of which is fixed, providing significant pedagogical implications for both EAP teaching and learning.

We also investigated move-signaling words and lexical bundles of some of the major moves to shed light on how they are used to realize the rhetorical functions of them. Results firstly presented the move signaling words observed from the high-frequency wordlists. Then, lexical bundles characterizing the rhetorical functions with high frequencies were selected. It was found that the examination of high-frequency wordlists revealed words associated with the rhetorical functions such as the reporting verbs in IP or concessive sentence-connectors used in IG. Also, high frequently recurred lexical bundles of some moves are designated in the representation of the rhetorical functions of move. We may conclude that the examination of language use from subcorpus of each move helps reveal subtle linguistic features hard to be noticed by investigating only the whole corpus.

Implications of the Study

The quantitative analysis of the study was mainly based on the construction of a corpus. The word frequency lists, move/common move patterns, and lexical bundles of the study were all derived from the analysis of the corpus with NLP tools, setting a good example in terms of the use of corpora in vocabulary studies. Corpus-based results enable researchers, teachers and students to have an access to language use in real world instead of relying on intuition or made-up examples. *Frequency* serves as the most important information that relies a great deal on the use of corpus studies. An understanding of how frequently words occur and how words are covered by wordlists developed for different purposes is of help to know the characteristics of words to be studied. In addition, the comparison of word frequency lists of different genres or in different fields might result in information regarding the composition of word frequency list. This information not only reflects the characteristics of a genre but helps teachers set an appropriate learning goal that fits learners' needs. On the other hand, many studies in the past have been emphasizing the importance of the GSL, indicating its high coverage in texts is useful in comprehending texts of various types. Since the majority of English teachers are lack of the specialist knowledge of learner's technical area, specialized vocabulary such as technical vocabulary is often neglected. Although language teachers may not have knowledge of learners' specialized areas, what they can do is to provide learning materials specifically designed for students' field such as the construction of a wordlist for specific purposes. Since most students may have certain control over the GSL, the supplement of the AWL or specialized vocabulary may enhance their comprehension of specialized texts. Finally, the learning of vocabulary should not be constrained to individual word meaning. Rather, knowing how a word relates to its discourse function is important because words are meaningful when used in context. As a result, knowledge about words that co-occur with the words in concern or how words are used in context such as collocation or lexical bundles of a word is important since it is the essence of language knowledge and distinguishes native speakers from non-native speakers.

Limitations and Future Research

The results of this research show that the use of corpus-based approach is insightful in exploring the nature of vocabulary, linking the use of vocabulary with its corresponding rhetorical function in RAs. Because of time limitation, some aspects worthy of being investigated are not completed in this study. We, thus, provide a number of directions for future research. First, some of the results of our study are constrained or insignificant because of the small size of the corpus. To generalize the research results, it is suggested that a larger corpus is used for future investigation. Also, since our study only focuses on the Introduction section of RAs, it is believed that analyses of other sections of RAs will be insightful for an understanding of the genre of RAs as a whole. Finally, to identify the distinguished characteristics of a discipline, future research might be aimed at comparing findings obtained from different research fields. Further, the comparison of native speakers’ corpus with learner corpus is likely to bring valuable information concerning the needs and difficulties learners have, providing a solid foundation for curriculum design and materials development.

APPENDIXES

Appendix A
Sources

IEEE Transactions on Computers

Text 1a. Rexford, Jennifer, Hall, John, & Shin, Kang, G.., (1998). A router architecture for real-time communication in multicomputer networks. *IEEE Transactions on computers, 47, 10,* 1088-1101.

Text 2a. Fiore, Paul, D., (1999). Parallel multiplication using fast sorting networks. *IEEE Transactions on computers, 48, 6*, 640-645.

Text 3a. Kumar, Vijay, Prabhu, Nitin, Dunham, Magaret H., & Seydim, Ayse Yasemin, (2002). TCOT –a timeout-based mobile transaction commitment protocol. *IEEE Transactions on computers, 51, 10,* 1212-1218.

Text 4a. Park, Joonseok, Diniz, Pedro C., & Shayee, K. R. Shesha, (2004). Performance and area modeling of complete FPGA designs in the presence of loop transformations. *IEEE Transactions on computers, 53, 11*, 1420-1435.

Text 5a. Ofek, Yoram, Yener, Bulent, & Yung, Moti, (1997). Concurrent asynchronous broadcast on the metanet. *IEEE Transactions on computers, 46, 7*, 737-748.

Text 6a. Marcuello, Pedro, Gonzalez, Antonio, & Tubella, Jordi, (2004). Threaad partitioning and value prediction for exploiting speculative thread-level parallelism. *IEEE Transactions on computers, 53, 2*, 114-125.

Text 7a. Pineiro, Jose-Alejandro, Bruguera, Javier Diaz, (2002). High-speed double-precision computation of reciprocal, division, square root, and

inverse square root. *IEEE Transactions on computers, 51, 12,* 1377-1388.

Text 8a. Chisholm, G. H., & Wojcik, A. S., (1999). An application of formal analysis to software in a fault-tolerant environment. *IEEE Transactions on computers, 48, 10,* 1053-1064.

Text 9a. Schwiebert, Loren, (2001). Deaadlock –free oblivious wormhole routing with cyclic dependencies. *IEEE Transactions on computers, 50, 9,* 865-876.

Text 10a. Danysh, Albert, & Tan, Dimitri, (2005). Architecture and implementation of a vector/SIMD multiply-accumulate unit. *IEEE Transactions on computers, 54, 3,* 284-293.

Text 11a. Phipatanasuphorn, V., & Ramanathan, P., (2004). Vulnerability of sensor networks to unauthorized traversal and monitoring. *IEEE Transactions on Computers, 53, 3,* 364-369.

Text 12a. Pedregal-Martin, C., & Ramamritham, K., (2002). Support for recovery in mobile systems. *IEEE Transactions on Computers, 51, 10,* 1219-1224.

Text 13a. Radhakrishnan, R., Vijaykrishnan, N., John, L. K., Sivasubramaniam, A., Rubio, J., & Sabarinathan, J., (2001). Java runtime systems: characterization and architectural implications. *IEEE Transactions on Computers, 50, 2,* 131-146.

Text 14a. Abdelzaher, T. F., & Shin, K. G.., (2000). Period-based load partitioning and assignment for large real-time applications. *IEEE Transactions on Computers, 49, 1,* 81-87.

Text 15a. Mishra, P., & Srivastava, M., (1998). Effect of connection rerouting on application performance in mobile networks. *IEEE Transactions on*

Computers, 47, 4, 371-390.

Text 16a. Zuberi, K. M., & Shin, K. G.., (2000). Design and implementation of efficient message scheduling for controller area network. *IEEE Transactions on Computers, 49, 2,* 182-188.

Text 17a. Chanchio, K., & Sun, Xian-He, (2004). Communication state transfer for the mobility concurrent heterogeneous computing. *IEEE Transactions on Computers, 53, 10,* 1260-1273.

Text 18a. Vijaykrishnan, N., Kandemir, M., Irwin, M. J., Kim, H. S., Ye, W., & Duarte, D., (2003). Evaluating integrated hardware-software optimizations using a unified energy estimation framework. *IEEE Transactions on Computers, 52, 1,* 59-76.

Text 19a. Park, J., Diniz, P. C., & Shayee, K.R. S., (2004). Performance and area modeling of complete FPGA designs in the presence of loop transformations. *IEEE Transactions on Computers*, 53, 11, 1420-1435.

Text 20a. Sabbineni, H., & Chakrabarty, K., (2005). Location-aided flooding: an energy-efficient data dissemination protocol for wireless senior networks. *IEEE Transactions on Computers, 54, 1,* 36-46.

IEEE Transactions on Pattern Analysis and Machine Intelligence

Teax 1b. Chuang, J. H., Tsai, C. H., & Ko, M. C., (2000). Skeletonization of three-dimensional object using generalized potential field. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 11,* 1241-1251.

Text 2b. Senior, A., (2001). A combination fingerprint classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23, 10,* 1165-1174.

Text 3b. Beiden, S. V., Maloof, M. A., & Wagner, R. F., (2003). A general model for finite-sample effects in training and testing of competing classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, 12*, 1561-1569.

Text 4b. Cordella, L. P., Foggia P., Sansone, C., & Vento, M., (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26, 10,* 1367-1372.

Text 5b. Lam, L., & Suen, C. Y., (1995). An evaluation of parallel thinning algorithms for character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17, 9*, 914-919.

Text 6b. McCrowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, D., (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27, 3*, 305-317.

Text 7b. Liu, C. L., Koga, M., & Fujisawa, H., (2002). Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24, 11*, 1425-1437.

Text 8b. Borgefors, G.., Ramella, G.., & di Baja, G.. S., (2001). Hierarchical decomposition of multiscale skeletons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23, 11*, 1296-1312.

Text 9b. El-Yacoubi, A., Gilloux, M., & Suen, C.Y., (1999). An HMM-based approach for off-line uncontrained handwritten word modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21, 8*, 752-760.

Text 10b. Rocha, J., & Pavlidis, T., (1995). Character recognition without segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17, 9*, 903-909.

Text 11b. Ahmed, M., & Ward R., (2002). A rotation invariant rule-based thinning algorithm for character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, *12*, 1672-1678.

Text 12b. Singh, S., (2003). Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, *12*, 1534-1539.

Text 13b. Ho, Tin Kam, & Baird, Henry S., (1997). Large-scale simulation studies in image pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, *10*, 1067-1079.

Text 14b. Madhvanath, S., Kleinberg, E., & Govindaraju, V., (1999). Holistic verification of handwritten phrases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*, *12*, 1344-1356.

Text 15b. Watanabe, M., & Nayar, S. K., (1997). Telecentric optics for focus analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19,12*, 1360-1365.

Text 16b. Havaldar, P., & Medioni, G., (1998). Full volumetric descriptions from three intensity images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, *5*, 540-545.

Text 17b. Starner, Thad, Weaver, J., & Pentland, A., (1998). Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, *12*, 1371.

Text 18b. Jiang, Xiaoyi, (2000). An adaptive contour closure algorithm and its experimental evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22, 11*, 1252-1265.

Text 19b. Fredembach, C., Schroder, M., Susstrunk, S., (2004). Eigenregions for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26, 12*, 1645-1649.

Text 20b. Marinai, S., Gori, M., & Soda, G.., (2005). Artificial neural networks for document analysis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27, 1*, 23-35.

Computational Linguistics

Text 1c. Venkataraman, A., (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27, 3*, 351-372.

Text 2c. Ploux, S., & Ji, H., (2003) A model for matching semantic maps between languages (French/ English, English/ French). *Computational Linguistics*, *29, 2*, 155-178.

Text 3c. Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin M., (2004). Learning subjective language. *Computational Linguistics*, *30, 3*, 277-308.

Text 4c. Kibble, R., & Power, R., (2004). Optimizing referential coherence in text generation. *Computational Linguistics*, *30, 4*, 401-416.

Text 5c. Teahan, W. J., Wen, Yingying, McNab, R., & Witten, Ian H., (2000). A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, *26, 3*, 375-393.

Text 6c. Navigli, R., & Velardi, P., (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, *30, 2*, 151-179.

Text 7c. Silber, H. Gregory, & McCoyy, Kathleen F., (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, *28, 4*, 487-496.

Text 8c. Santamara, C., & Gonzalo, J., & Verdejo, F., (2003). Automatic association of web directories with word senses. *Computational Linguistics*, *29, 3*, 485-502.

Text 9c. Keller, F., & Lapatay, M., (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, *29, 3*, 459-484.

Text 10c. Fais, Laurel, (2004). Inferable centers, centering transitions, and the notion of coherence. *Computational Linguistics*, *30, 2,* 119-150.

Text 11c. Stamatatos, E., Fakotakis, N., & Kokkinakis, G.., (2001). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26, 4*, 471-495.

Text 12c. Pevzner, L., & Hearsty, Marti A., (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, *28, 1*, 19-36.

Text 13c. Li, Hang, & Li Cong, (2004). Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, *30, 1*, 1-22.

Text 14c. Mason, Zachary J., (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, *30, 1*, 23-44.

Text 15c. Branco, AntAonio, (2002). Binding Machines. *Computational Linguistics*, *28, 1*, 1-18.

Text 16c. Oflazer, Kemal, (2003). Dependency parsing with an extended finite-state approach. *Computational Linguistics*, *29, 4*, 515-544.

Text 17c. Marchand, Y., &Damper, R., (2000). A multistrtegy approach to improving pronunciation by analogy. *Computational Linguistics*, *26, 2*,

195-219.

Text 18c. Ke, J., Ogura, M., & Wang, William S.-Y., (2003). Optimization models of sound systems using genetic algorithms. *Computational Linguistics, 29, 1*, 1-18.

Text 19c. Kehler, A., Bear, J., & Appelt, D., (2001). The need for accurate alignment in natural language system evaluation. *Computational Linguistics*, *27, 2,* 231-248.

Text 20c. Teufel, S., & Moens, M., (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, *28, 4*, 409-445.

Appendix B
The CS Wordlist

abstract
access (accesses)
account (accounts, accounted)
accuracy (accurate, accurately)
achieve (achieves, achieved, achievement, achievements)
adapt (adapts, adapted, adaptive, adaptation)
adjacent
algorithm (algorithms)
align (aligns, aligned, alignment, alignments)
allow (allows, allowed)
analyze (analyzes, analyzed, analytical, analytically, analysis, analyses)
annotate (annotates, annotated, annotation, annotations, annotator, annotators)
approach (approaches)
appropriate (appropriately)
approximation (approximate, approximates)
architecture (architectures, architectural)
area
array
aspect (aspects)
assign (assigns, assigned, assignment, assignments)
associate (associates, associated, association, associations)
assume (assumes, assumed, assumption, assumptions)
author (authors)
automatic (automatically, automation)
available
axis
background
bandwidth
baseline
base-station
basis
bi-gram (bi-grams)
binary
bind (binding)

bit (bits)
block (blocks)
Branch (branches)
buffer (buffers)
bus
cache (caches)
calculate (calculates, calculated, calculation, calculations)
candidate (candidates)
capture (captures, captured)
category (categories)
cell (cells)
channel (channels)
characteristic (characteristics)
chip (chips)
classify (classifies, classified, classification, classifications, classifier, classifiers)
cluster (clusters, clustering)
code (codes)
cohere (coheres, cohesion, cohesive, coherence)
column (columns)
commit (commits, committed, commitment)
communicate (communicates, communicated, communication)
compile (compiles, compiled, compiler, compilers)
complex (complexity)
component (components)
compute (computes, computed, computing, computer, computation, computational)
concept (concepts, conceptual, conceptually)
configure (configures, configured, configuration, configurations)
consist (consists, consisting)
constant
constrain (constrains, constrained, constraint, constraints)
consume (consumes, consumed, consumption)
context
contour (contours)
contrast (contrasts, contrasted, contrastive, contrastively)
contribute (contributes, contributed, contribution, contributions)
convention (conventions, conventional)
core (cores)

corpus (corpora)
correlate (correlates, correlated, correlative, correlation)
correspond (corresponds, corresponded, corresponding, correspondence, correspondences)
create (creates, created, creation)
criterion (criteria)
critical (critically)
current (currently)
cycle (cycles)
data (datum)
database
data-path
deadlock
define (defines, defined, definition, definitions)
density
depend (depends, depended, dependency, dependencies, dependent, depending)
derive (derives, derived, derivation, derivations)
design (designs, designed)
destination (destinations)
detect (detects, detected, detection)
disambiguate (disambiguates, disambiguated, disambiguation)
discourse
distinct (distinction)
distribute (distributes, distributed, distribution, distributions)
document (documents)
domain (domains)
driven
dynamic (dynamically)
efficiency (efficient)
element (elements)
embed (embeds, embedded, embedment)
employ (employs, employed, employment)
energy
entire (entirely)
entity (entities)
environment
estimate (estimates, estimated, estimation)

evaluate (evaluates, evaluated, evaluation)
execute (executes, executed, execution)
existing
explicit (explicitly, explicitness)
extend (extends, extended, extension)
extract (extracts, extracted, extraction)
factor (factors)
feature (features)
feedback
figure (fig., figures)
finite
focus (foci)
following
framework
free
frequency (frequencies, frequent)
function (functions, functioned, functional)
generate (generates, generated, generation)
genre (genres)
given
global
goal (goals)
gram (grams)
graph (graphs)
grid (grids)
guarantee (guarantees)
handoff
handwritten
hardware
header (headers)
hence
heuristic (heuristically)
hierarchical
id
identify (identifies, identified, identifying, identical, identification)
image (images)
impact (impacts)

implement (implements, implemented, implementation)
index (indices)
individual (individually)
inferable (inferably)
initial (initially)
input
instance (instances)
instruct (instructs, instructed, instruction)
interface (interfaces)
intermediate
internal (internally)
interpret (interprets, interpreted, interpretive, interpretative, interpretation)
inverse (inversely, inversion)
issue (issues)
Iterate (iterates, iterated, iteration, iterations)
Java
known
label (labels, labeled)
latency
layout
len (lens)
lexicon (lexis, lexical)
linear
link (links)
logic (logical, logically)
loop (loops)
manual (manually)
map (mapped, mapping)
marker (markers)
mask (masks)
maximum
measure (measures, measured, measurement)
mechanism (mechanisms)
metaphor (metaphors)
method (methods, methodology)
metric (metrics)
migrate (migrates, migrated, migrating, migration)

minimum
mobile
mode (modes)
modify (modifies, modified, modification, modifications)
module (modules)
multiple
namely
negative (negatively)
neural
NLP
node (nodes)
normalize (normalizes, normalized, normal, normalization)
observed
obtain (obtains, obtained)
occur (occurs, occurred, occurring, occurrence, occurrences)
OCR
ontology
optimal (optimization, optimizations)
output
overall
overhead
packet (packets)
parallel (parallelism)
parameter (parameters)
participant (participants)
penalty
percentage (percentages, percent)
perceptual (perceptually)
perform (performs, performed, performance)
phase (phases)
phrase (phrases)
physical (physically)
pixel (pixels)
plane (planes)
policy (policies)
port (ports)
positive (positively)

potential
precision (precise, precisely)
predict (predicts, predicted, prediction, predictions, predictor, predictors)
preprocess (preprocesses, preprocessed, preprocessing)
present (presence)
previous (previously)
principle (principles)
prior
priority
procedure (procedures)
process (processes, processed, processing, processor, processors)
property (properties)
propose (proposes, proposed)
protocol (protocols)
provided
queue (queues)
random (randomly)
range (ranges, ranged)
ratio (ratios)
rebuild (rebuilds, rebuilt)
recall (recalls, recalled)
receiver (receivers)
reciprocal (reciprocally)
recognize (recognizes, recognized, recognition)
recover (recovers, recovered, recovery)
reduce (reduces, reduced, reduction)
region (regions)
register (registers)
reject (rejects, rejected, rejection)
relate (relates, related, relation, relations, relationship, relationships)
relative (relatively)
relevant (relevance)
rely (relies, relied, reliance)
remove (removes, removed, removal)
represent (represents, represented, representation)
require (requires, required, requirement, requirements, requiring)
rerouting

research
resolution (resolutions)
resource (resources)
respective (respectively)
reuse (reuses, reused)
rhetorical (rhetorically)
router (routing)
schedule (schedules, scheduled, scheduler, scheduling)
scheme (schemes)
score (scores, scored)
section (sections)
segment (segments, segmented, segmentation)
semantic (semantically)
sensor (sensors)
sequence (sequences)
shift (shifts, shifted, shifting)
significant (significantly, significance)
simulate (simulates, simulated, simulating, simulation)
skeleton (skeletons, skeletonization)
slot (slots)
smoothing (smooth)
software
source (sources)
specific (specifically)
specification (specify, specifies, specified)
speculate (speculates, speculated, speculative, speculatively, speculation)
statistical (statistically)
status
strategy (strategies)
stroke (strokes)
structure (structures, structural, structurally)
style (styles)
subjective (subjectively, subjectivity)
subset (subsets)
summary (summaries)
switch (switches, switched)
symbol (symbols)

syntactic (syntactically)
synthesis (syntheses)
table (tables)
tap (taps, tapped)
target (targets)
task (tasks)
technique (techniques)
technology
tele-centric
text (texts)
theory (theories)
thin (thinning)
thread (threads)
threshold
throughput
tile (tiling)
tone (tones)
topic (topics)
topology
transfer (transfers, transferred)
transform (transforms, transformed, transformation, transformations)
transition (transitions)
transmit (transmits, transmitted, transmission)
tree (trees)
typical (typically)
unique (uniquely, uniqueness)
unit (units)
unreachable (unreachably)
unrolling
unseen
utterance (utterances)
variable (variables)
variation (various, variance)
vector (vectors)
verify (verifies, verified, verification)
version (versions)
vertical (vertically)

via
virtual (virtually)
web (webs)
whereas
wireless
word-net (word-nets)
Note: 1. The CS wordlist here contains only 335 word families which cover 80% out of 388,396 running words of our corpus. The complete CS wordlist constitutes 1402 word families, accounting for 95% for the whole corpus. 2. A number of GSL words with different meanings in CS are retained, including *bit, block, branch, bus, depend, driven, figure, framework, free, frequency, given, map, observed, and recognize.*

REFERENCES

Aston, Guy (Ed.) (2001). *Learning with Corpora*. Houston: Athelstan.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Beglar, D., & Hunt, A. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*(1), 23-59.

Beglar, D., and Hunt, A. (2005). Six principles for teaching foreign language vocabulary: a commentary on Laufer, Meara, and Nation's "ten best ideas." *The Language Teacher*, 29, 7, 7-10.

Bhatia, V. (1993). *Analyzing genre: Language use in professional settings*. London& New York: Longman.

Bhatia, V. (2004). *World of written discourse: A genre-based view*. New York: Continuum.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes, Article in Press.

Biber, D., Conrad, S., & Cortes, V. (2004) If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in Applied Linguistics. *Applied Linguistics*, *15*, 169-189.

Biber, D., Conrad, S., & Reppen, R. (1996). Corpus-based investigations of language use. *Annual Review of Applied Linguistics*, *16*, 115-136.

Biber, D. & Cortes, V. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard and S. Oksefijell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181-189). Amsterdam: Rpdopi.

Bogaards, P. and Laufer, B. (eds.) (2004). Vocabulary in a second language.

Amsterdam/Philadelphia: John Benjamins.

Brett, P. (1994). A genre analysis of the Results section of sociology articles. *English for Specific Purposes, 13,* 47-60.

Bunton, D. (2002). Generic moves in Ph.D. thesis introductions. In J. Flowerdew (Ed.), *Academic discourse* (pp. 57-75). London: Pearson Education.

Carter, R., & McCarthy, M. (1988). *Vocabulary and language teaching*. London: Longman.

Charles, M. (2003). "This mystery…": a corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes, 2*(4), 313-326.

Coady, J. (1993). Research on ESL/EFL vocabulary acquisition: Putting it in context. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 3-23). Norwood, NJ: Ablex.

Cobb, T. & Horst, M. (2001). Reading academic English: carrying learners across the lexical threshold. In J. Flowerdew and M. Peacock (Eds.), *Research perspectives on English for Academic Purposes* (pp. 315-329). Cambridge: Cambridge University Press.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1988). Reading English for specialized purposes: Discourse analysis and the use of standard informants. In P. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 152-167). Cambridge: Cambridge University Press.

Conrad, S. (2002). Corpus linguistics approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75-95.

Cooper, C. (1985). *Aspects of article introductions in IEEE publications*. MSc. Dissertation, University of Aston, Birmungham.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397-423.

Coxhead, A. (2000). A new academic word list. *Tesol Quarterly, 34*(2), 213-238

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English of academic purposes. In J. Flowerdew & M. Peacock, (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge, England: Cambridge University Press.

Crookes, G. (1984). Towards a validated analysis of scientific text structure. *Applied Linguistics, 7*(1), 57-70.

Crookes, G. (1986). Towards a validated analysis of scientific text structure. *Applied Linguistics, 7*, 57-70.

Deutch, Y. (2003). Needs analysis for academic legal English courses in Israel: a model of setting priorities. *English for academic purposes, 2*, 125-146.

Flowerdew, J. (Ed.) (2002). *Academic Discourse*. London: Pearson Education.

Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock, (Eds.), *Research perspectives on English for academic purposes* (pp. 8-24). Cambridge, England: Cambridge University Press.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*, 321-332.

Gledhill, C. (2000). The discourse function of collocation in research article introduction. *English for Specific Purposes, 19*, 115-135.

Halliday, M. A. K. (1978). *Language as social semiotic*. London: Edward Arnold.

Haynes, M. (1993). Patterns and perils of guessing in second language reading. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and* *vocabulary learning* (pp. 46-65). Norwood, NJ: Ablex.

Hirsh, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8(2)*, 689-696.

Hopkins, A. & Dudley-Evans, T. (1988). A genre-based investigation of the discussion sections in articles and dissertations. *English for Specific Purposes, 7*, 113-121.

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

Hwang, K., & Nation, P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language, 6*(1), 323-335.

Hyland, K. (1996). Writing without conviction? Hedging in scientific research articles. *Applied Linguistics, 17*(4), 433-454.

Hyland, K. (1997). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (2000). *Disciplinary discourse: social interactions in academic writing*. Harlow, England: Pearson Education Limited.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: issues and directions. *Journal of English for Academic Purposes, 1*, 1-12.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman: New York.

Kuo, C. H. (1987, May). A Needs Analysis of University Undergraduates, Graduates, and Technical Professionals. Paper presented at the 4th National Conference on English Teaching and Learning, Taipei.

Kuo, C. H. (2002). Phraseology in scientific research articles. In *Selected papers from the Eleventh international symposium on English teaching* (pp. 405-411). Taipei: Crane.

Kuo, C. H., Chang, C. F., Lin, M. H., & Lin, B. H. (2006, September). *A Corpus-based Approach to EAP Materials Development*. Paper presented at EuroCALL 2006, Granada, Spain.

Kwan, B. S. C. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes, 25*, 30-55.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordam, (Eds.), *Special language: From humans thinking to thinking machines*. Clevedon: Multilingual Matters.

Laufer, B. (1997). What's in a word that makes it hard or easy? Some intralexical factors that affect the learning of words. In Schmitt and McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140-155). Cambridge, England: Cambridge University Press.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow: Pearson Education.

Lim, J. M. H. (2006). Method sections of management research articles: A pedagogically motivated qualitative study. *English for Specific Purposes, 25*, 282-309.

Liou, H. C., Chang, J. S., Kuo, C. H., Chen, H. J., & Chang, C. F. (2006). Web-based academic English course design and material development. *Selected papers from the fourteenth International Symposium and Book Fair on English Teaching. P 452-462*, Taipei, Taiwan.

Martin, A. V. (1976). Teaching academic vocabulary to foreign graduate students.

TESOL Quarterly, 19, 91-98.

Meijs, W. (1996). Linguistic corpora and lexicography. *Annual Review of Applied Linguistics, 16,* 99-114.

Meyer, P. G. (1990). *Non-technical vocabulary in technical language,* paper delivered at AILA congress in Thessalonika.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25,* 235-256.

Nation, P. (1990). *Teaching and learning vocabulary*. New York: Heinle and Heinle.

Nation, P. (2001). *Learning vocabulary in another language*. United Kingdom: Cambridge.

Nation, P., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35-41.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy, (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge, England: Cambridge University Press.

Nwogu, K. N. (1997). The medical research paper: structure and functions. *English for Specific Purposes, 16*(2), 119-138.

Paltridge, B. (2001). Linguistic research and EAP pedagogy. In J. Flowerdew & M. Peacock, (Eds.), *Research perspectives on English for academic purposes* (pp. 55-70). Cambridge, England: Cambridge University Press.

Paltridge, B. (2002). Thesis and dissertation writing: an examination of published advice and actual practice. *English for Academic Purposes, 21*, 125-143.

Posteguillo, S. (1999). The semantic structure of computer science research articles, *English for Specific Purposes, 18*(2), 139-160.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, J. C. (1970). A psycholinguistic measure of vocabulary selection. *IRAL*, 8, 2, 87-102.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77-89.

Ruiying, Y., & Allison, D. (2003). Research articles in applied linguistics: moving from results to conclusions. *English for Specific Purposes, 22*, 365-385.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes, 11*(2), 93-113.

Samaraj, B. (2002). Introductions in research articles: variations across disciplines. *English for Specific Purposes, 21*, 1-17.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

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

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy, (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237-257). Cambridge, England: Cambridge University Press.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, *25*(2), 34-50.

Swales, J. M. & Najjar, H. (1987). The writing of research article introductions. *Written Communication, 4*, 175-192.

Swales, J. M. (1974). Notes on the function of attributive en-participles in scientific discourse. *Papers for Special University Purposes No.1, ELSU*, University of Khatoum.

Swales, J. M. (1981). *Aspects of article introductions*. Birmingham, UK: The University of Aston, Language Studies Unit.

Swales, J. M. (1988). 20 years of TESOL Quarterly. *TESOL Quarterly, 22*, 151-163.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2001). EAP-related linguistic research: an intellectual history. In J. Flowerdew & M. Peacock, (Eds.), *Research perspectives on English for academic purposes* (pp. 42-54). Cambridge, England: Cambridge University Press.

Swales, J. M. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.

Tarone, E. S., S. Dwyer, S. Gillette, & Icke, V. (1981). On the use of the passive in two astrophysics journal papers. *English for Specific Purposes, 1*, 123-140.

Tarone, E., Dwyer, S., Gillette, S., & Icke,V. (1998). On the use of the passive and active voice in astrophysics journal papers: With extensions to other languages and other fields. *English for Specific Purposes*, 17, 1, 113-132.

Thompson, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics, 12*(4), 365-382.

Thurstun, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes, 17*(3), 267-280.

Trimble, L. (1985). *English for science and technology*. Cambridge: Cambridge University Press.

West, M. (1953). *A general service list of English words*. London: Longman.

Weissberg, R., & Buker, S. (2005).*Writing up research: Experimental research report for students of English*. Taiwan: Pearson Education Taiwan.

Widdowson, H. G. (1998). Context, community and authentic language. *TESOL Quarterly, 32*(4), 705-716.

Widdowson, H. G. (2000). Corpora and language teaching tomorrow. In Keynote lecture delivered at 5th teaching and language corpora conference, Bertinoro, Italy, 29 July.

Williams, I. A. (1999). Results sections of medical research articles: Analysis of rhetorical categories for pedagogical purposes. *English for Specific Purposes, 18,* 347-366.

Worthington, D., & Nation, P. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal*, *27*(2), 1-11.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication, 3,* 215-229.