An Investigation of Lexical Bundles in Electrical Engineering Introductory Textbooks and ESP Textbooks

By

Lin Chen

A Thesis
Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
For the Degree of

Master of Arts

School of Linguistics and Applied Language Studies
Carleton University
Ottawa, Ontario
September 8, 2008

© 2008, Lin Chen
NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Canada
Abstract

This corpus-based research examines whether there is a gap in language use between ESP textbooks and Electrical Engineering introductory textbooks. Two corpora were set up, the Electrical Engineering introductory textbook corpus (EEITC) and the English for Specific Purposes textbook corpus (ESPTC). Lexical bundles and their pragmatic functions were identified in the EEITC; the majority are referential and stance bundles, which construct the body of knowledge and indicate the reader/author relationships in introductory EE textbooks. The comparison of lexical bundles in the two corpora suggests that there is a gap: compared with the EEITC, the ESPTC lacks five subcategories of lexical bundles and misrepresents target language use. The two main implications of this research are: 1) lexical bundles with their pragmatic functions should be incorporated into the EAP/ESP pedagogies, and 2) lexical bundles can become one way to judge the authenticity of EAP/ESP pedagogical materials.
Acknowledgements

It is my great pleasure to express my sincere gratitude and appreciation to everyone who made this thesis possible. First and foremost, I would like to thank Dr. Ellen Cray for her generous help. I could go nowhere without her endless encouragement and timely feedback throughout the research and writing process. She coached me for the presentation of this research at the 39th annual conference of CAAL (Canadian Association of Applied Linguistics) and directed me to the completion of my thesis. No words could express my special thanks to Ellen.

My supervisor, Dr. Ellen Cray, and I are greatly indebted to Dr. David Wood for his generous help and invaluable suggestions. He sacrificed his time to discuss key issues of my research and offered cogent comments. I appreciate David’s trust in me as his research assistant. The insights drawn from this work experience helped me shape my thesis. In addition, David, an amazing and incredible professor with a great sense of humour, has helped me overcome my initial fear of writing and presenting.

I would like to thank my external examiner, Dr. Tom Cobb, for his invaluable feedback and generous encouragement, which will motivate me to continue my work in the future.

I would also like to thank my friend Diane Fraser, who constantly offers me support and encouragement. Special thanks to my husband, Xinjie Wang, who has been supporting me all these years and has offered me a shoulder to lean on.
# Table of Contents

Abstract................................................................................................................................. i

Acknowledgements ............................................................................................................... ii

List of Tables......................................................................................................................... vi

List of Figures........................................................................................................................ vii

List of Appendices................................................................................................................ viii

Chapter 1 Introduction.......................................................................................................... 1

1.1 Research questions........................................................................................................... 9

1.2 About this research ......................................................................................................... 10

Chapter 2 Corpus and Formulaic Language......................................................................... 12

2.1 General corpus vs. Specialized corpus........................................................................... 13

2.2 Representativeness: general vs. specialized corpora....................................................... 17

2.3 Formulaic language and specialized corpora................................................................. 22

2.4 Formulaic language........................................................................................................ 24

2.4.1 Forms of formulaic language.................................................................................... 26

2.4.2 Compositionality of formulaic language.................................................................. 27

2.4.3 Pragmatic functions of formulaic language.............................................................. 28

2.4.3.1 An introduction to formulaic language in spoken and written discourse........... 29

2.4.3.2 Formulaic language in conversational and transactional spoken discourse........ 30

Chapter 3 Lexical bundles.................................................................................................... 34

3.1 Lexical bundles are not phraseological units................................................................. 34

3.2 Research on lexical bundles........................................................................................... 36

3.3 Features of lexical bundles............................................................................................. 37

3.4 Grammatical structures of lexical bundles.................................................................... 39

3.5 Pragmatic specialization of lexical bundles.................................................................... 41
3.5.1 Stance expressions ......................................................... 44
3.5.2 Discourse organisers ..................................................... 45
3.5.3 Referential bundles ..................................................... 46
3.6 Discussion of lexical bundles ........................................... 48
  3.6.1 Frequency and formulaicity ........................................ 48
  3.6.2 Written text, context, and lexical bundles ..................... 50

Chapter 4 Methodology .............................................................. 57
  4.1 Setting up the EEITC ....................................................... 57
    4.1.1 Selection of representative introductory textbooks of Electrical Engineering ........................................... 58
    4.1.2 The compilation of the Electrical Engineering Introductory Textbook Corpus (EEITC) ....................................................... 61
  4.2 Setting up the comparison corpus: the ESP Textbook Corpus (ESPTC) ................................................................. 62
  4.3 The identification criteria of lexical bundles ....................... 63
  4.4 Using WordSmith Tools 4.0 for identification of lexical bundles and textual analysis ......................................................... 65
  4.5 The analytical framework: the functional taxonomy by Biber et al. (2004a) ................................................................. 67

Chapter 5 Results and discussion ................................................... 70
  5.1 What does text do in discipline textbooks? ............................ 70
  5.2 Structural characteristics of lexical bundles in the EEITC ............ 74
    5.2.1 Phrasal lexical bundles ........................................... 76
    5.2.2 Clausal lexical bundles ........................................... 78
    5.2.3 Discussion of structural characteristics of lexical bundles ....... 80
  5.3 Functional characteristics of lexical bundles in the EEITC .......... 82
    5.3.1 Referential bundles in the EEITC ................................... 84
      5.3.1.1 Referential identification/focus bundles in the EEITC ....... 84
      5.3.1.2 Attribute specification bundles in the EEITC ............... 86
      5.3.1.3 Place/text/multi-functional referential bundles in the EEITC .... 88
    5.3.2 Stance bundles in the EEITC ....................................... 90
5.3.2.1 Epistemic stance bundles in the EEITC ........................................... 90
5.3.2.2 Attitudinal/modality stance bundles in the EEITC ......................... 93
5.3.3 Discourse organizing bundles in the EEITC ........................................ 95
5.4 Lexical bundles in the ESP textbooks .................................................... 95
  5.4.1 Stance bundles in the ESPTC ............................................................ 98
  5.4.2 Discourse organizers in the ESPTC .................................................. 98
  5.4.3 Referential bundles in the ESPTC .................................................... 98
5.5 The comparison of lexical bundles in the ESPTC and the EEITC .......... 100
  5.5.1 The comparison of referential bundles in the EEITC and the ESPTC .. 104
  5.5.2 The comparison of stance bundles in the EEITC and the ESPTC ........ 108
  5.5.3 The comparison of discourse organizers in the EEITC and the ESPTC .. 111
5.6 ESP textbook treatment of target lexical bundles in the EEITC .......... 112
  5.7 Critique of the functional taxonomy .................................................. 117
  5.8 Summary ............................................................................................... 122

Chapter 6 The teaching and learning implications of lexical bundles ........ 126

Chapter 7 Conclusion .................................................................................... 134

References ..................................................................................................... 144
List of Tables

Table 2.1 Situational parameters listed as hierarchical sampling strata .........................18
Table 4.1 Textbooks used in introductory courses of electrical engineering at three Ontario universities ..................................................................................................................59
Table 5.1 Structural types and proportional distribution of lexical bundle in the EEITC .................................................................................................................................75
Table 5.2 Structural classification of lexical bundles in the EEITC ........................................75
Table 5.3 Functional classification of 4-word lexical bundles in the Electrical Engineering Introductory Textbook Corpus (EEITC) ........................................................................83
Table 5.4 Functional classification of 4-word lexical bundles in the ESP Textbook Corpus (ESPTC) ..................................................................................................................97
Table 5.5 Proportional function distribution of lexical bundles in the EEITC and ESPTC ...............................................................................................................................102
Table 5.6 The treatment of lexical bundles in *Oxford English for electronics* ..............113
Table 5.7 The treatment of lexical bundles in *Oxford English for electrical and mechanical engineering* .........................................................................................................114
List of Figures

Figure 3.1 Functional taxonomy of lexical bundles across four registers .................. 43
Figure 4.1 The frequency list of four-word clusters in the EEITC .......................... 66
Figure 4.2 The Concord window of the current in the ........................................ 67
Figure 5.1 Functional distributions of lexical bundles in the EEITC and the ESPTC .... 101
Figure 5.2 Proportional distributions of function subcategories of lexical bundles in the EEITC and the ESPTC ........................................................................................................ 103
List of Appendices

Appendix A: Questionnaire for Evaluation of introductory textbooks in Electrical Engineering.................................................................138

Appendix B: Breakdown of the Electrical Engineering Introductory Textbook Corpus (EEITC) and the ESP Textbook Corpus (ESPTC) ......................................................... 141

Appendix C: Encoding the functional taxonomy (Biber et al., 2004a).........................................................142
Chapter 1 Introduction

The majority of theoretical linguists seem to favor Noam Chomsky's notion of language structure and base their theoretical claims on introspection examples (Simpson & Swales, 2001). Rationalists like Chomskyans believe that language consists of competence (internalized grammar) and performance (language in use) and argue that language is essentially based on an internalized grammar, represented with generative rules. Wray (2002) has pointed out that the Chomskyan camp avoids "any engagement with what people actually say, or which grammatically possible constructions of their language they might find more difficult to encode and/or decode than others" (p. 13). In a word, theoretical linguists assume that "[...] language is understood in terms of what can be said or written (competence-based), derived from introspective and elicitation techniques" (Flowerdew, 2004, p. 13).

In contrast, applied linguists have greater interest in how language is used in everyday life. The development of corpus linguistics has made it possible for us to observe language through authentic samples of language use. A corpus is "a collection of texts assumed to be representative of a given language dialect, or other subset of a language, to be used for linguistic analysis" (Francis, 1982, p. 7), and thus offers evidence of language use. The availability of corpus data allows linguists to focus on what language users actually do and say. For example, a study by Tao (2001) provided corpus-based data as evidence against findings based on linguists' introspection. He investigated the use of remember in natural spoken language represented by corpora consisting of spoken discourse with interactive levels ranging from casual conversations among family members to formal occasions of religious sermons, business meetings and
lectures at universities. While Van Valin and Wilkins (1993) depended on intuitions to reach the conclusion that the verb *remember* is used as "a third person reporting past events", as indicated in the example as "John remembered that he left the faucet on", Tao examined the corpora data and discovered that *remember* actually co-occurs with first-/second- pronouns for purposes of stating speakers' epistemic stance, or introducing/elaborating (new) topic(s) in conversation and discussion. Obviously the real use of *remember* shares no similarities with what Van Vanlin and Wilkins suggested.

If we agree that generative grammar was the leading way that language users encode/decode language, it would be natural to predict that "one may produce an utterance that is grammatical but unnatural or highly marked" (Pawley and Syder, 1983, p. 12). However, it has been observed that native-speaking adults rarely make grammatical yet non-native like utterances. The generative rules can not be used to explain why such cases seem to never have occurred and why native-speaking adults prefer to associate words with certain words in certain ways (Wray, 2002). Wray explained that "in English you *reach* a decision, *carry* the responsibility and *reap* the benefit" (p. 73). This preference for certain association patterns is one of the puzzles that have not been fully explained by a competence-focused theory.

Hymes (1971) expanded the coverage of competence by pointing out that it covers four distinct kinds of knowledge, including "the knowledge of what is formally possible given by the constraints of the language system, of what is feasible, of what is appropriate, and of what is performed " (quoted in Partington, 1998, p. 18). Similarly, Widdowson (1979) argued that "[...] knowledge of a language does not mean only knowledge of rules which will generate an infinite number of sentences, but a knowledge
of the rules which regulate the use of sentences making appropriate sense " (p. 12).

Adopting pragmatic competence into the model of language, applied linguists have a better understanding of the role context plays in shaping language in a particular situation, and of how writers'/speakers' awareness of the context affects the way they use language. Following Leech (1983), Nattinger and DeCarrico (1992) positioned pragmatic competence between the two poles of competence (grammatical competence) and performance. The notion of pragmatics by Nattinger and DeCarrico echoed the two categories from Hyme's division of competence (quoted in Partington, 1998), the knowledge of feasibility and of appropriateness. Nattinger and DeCarrico specified that pragmatic competence is more closely related with pragmalinguistics, namely, "the more linguistic end of pragmatics - where we consider the particular resources which a given language provide for conveying particular illocutions "(Leech, 1983, p. 10). In contrast to pragmalinguistics, socio-linguistics is concerned more with 'local', culture-specific conditions on language use, and thus is closer to the pole of performance. I will further discuss the pragmatic functions of lexical bundles, frequent co-occurring word collocations (Biber, Johansson, Leech, Conrad, & Finegan, 1999), which are theoretically based on pragmatic competence.

Sinclair (1991) also disagreed that generative grammar is the only way of meaning creation since this type of simplification can not satisfy the need to explain unrandomness of apparent word associations in native-speakers' language. He doubted that native-speaker introspections, either by informants or by the linguists, can “give evidence about usage” (p. 39). Instead of avoiding answering questions about unrandomness of language patterns, he maintained that meaning is created by both the open-choice principle and the
idiom principle.

The open-choice principle is a “slot-and-filler” model based on generative rules. “At each point where a unit is completed (a word, or a phrase, or a clause), a large choice opens up and the only restraint is grammaticalness” (Sinclair, 1991, p. 109). To make it simple, this “slot-and-filler” model allows a range of grammatical choices to be filled at different slots in a language unit. This whole process of filling the slots is analytical in nature since lexis and grammar are divided, where the latter is used to provide “a string of lexical choice points” (Sinclair, 1991, p. 109).

Sinclair (1991) argued that language meaning is established by the idiom principle, which can be used to explain unrandomness of word choices occurring in a text. “The principle of idiom is that a language user has available a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments” (p. 110). That principle can be reflected by the holistic process of language use, where we see that speakers or writers use prefabricated sequences as a whole without considering grammar issues. Wray (2002) also noted the coexistence of analytic and holistic processing in her dual-system model, which is similar to Sinclair’s dichotomy model. Wray stated that “Analytical processing entails the interaction of words and morphemes with grammatical rules, to create, and decode, novel, or potentially novel, linguistic material. Holistic processing relies on prefabricated strings stored in memory” (p. 14).

The relationship between the open-choice principle and the idiom principle is a dichotomous but complementary one. Dichotomy means that the two principles are incompatible with each other when in use. The two principles operate in complementary
ways, with the idiom principle dominating over the open-choice principle. In the process of interpreting a text, speakers/writers depend on the idiom principle first, and "with good reasons", they switch to the open-choice principle, and back again. Here idiom has the broad meaning equivalent to word association patterns where "the choice of one word affects the choice of others in its vicinity" (Sinclair, 1991, p. 173).

Due to the significant role that the idiom principle plays in language use, unsurprisingly, formulaic language has become the center of many corpus-based studies. Formulaic language is commonly defined as multiword strings "stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray & Perkins, 2000, p. 1). Despite the fact that speakers and writers automatically produce formulaic language at a level below conscious control, researchers have managed to propose a range of linguistic descriptors with the purpose of describing formulaic language in a systematic way. Those labels include "lexical phrases" (Nattinger & DeCarrico, 1992), "formulaic sequences" (Wray, 2002), "lexical bundles" (Biber et al., 1999), "collocation" (Partington, 1998), "recurrent word combinations" (Altenberg, 1998), "phrasal lexemes" (Moon, 1998b), "composite units " (Howarth, 1996), etc. In spite of the nuances among those labels, they have been utilized to examine pragmatic functions of formulaic language in various contexts (Altenberg, 1998; Biber et al., 1999; Nattinger & DeCarrico, 1992).

Until now, corpus linguists have depended on perceptual, statistical, or mixed approaches to pin down formulaic language from collections of text in corpora. Moon (1998a) used a perceptual method for identifying formulaic language. She determined that the pre-existing collection of 6,000 expressions in Collins COBUILD English
Dictionary (1987) were formulaic, with the rationale that those expressions "[…] for reasons of semantics, lexico-grammar, or pragmatics are regarded as holistic units rather than compositional strings" (Moon, 1998a, p. 79). Biber et al. (1999) assumed that formulaic language can be detected by its high frequency in natural discourse, and thus took a frequency-driven approach to the identification of lexical bundles, namely continuous multi-word clusters that occur frequently in natural discourse and spread across at least five texts. Simpson (2004) used a mixed statistical and perceptual method to identify formulaic expressions in the Michigan Corpus of Academic Spoken English (MICASE). First she used a text analysis software program to generate statistical frequency lists of three- and four-word sequences that satisfied the cut-off frequency of 20 times per million words. Then she relied on native speakers' intuition to locate formulaic sequences that showed structural and idiomatic coherence. As she explained, "only strings that constitute complete syntactic units, sentence stems, or that intuitively look, sound, and feel like idiomatically independent expressions were included in the set " (Simpson, 2004, p. 42).

With formulaic language located in natural language, many researchers with special interest in natural discourse shift their focus to the investigation of pragmatic functions of formulaic language, aiming to discover how formulaic language has been used to construct spoken/written discourse in a variety of registers. Oakey (2002) reported how a lexical phrase it is/has been (often) _ that X functioned differently in academic texts from three broad discipline branches, social science, medicine, and technical engineering. The author found that writers in medical science used the lexical phrase more frequently than writers in the other two disciplines. When using this lexical phrase, writers preferred the
present perfect form over the simple present form. Among the occurrences of the phase in present perfect form, 70% were used to introduce outside information as support without mentioning the source. Medical writers used the lexical phrase in this way with the purpose of adding authority to their own stance or position. Only occasionally did the writers note outside sources in the written text. O'Keefe, McCarthy and Carter (2007) demonstrated that the same formulaic expression *we need to* served different functions in different types of spoken discourse. In business dialogues, *we need to* framed requests from an authority (the board of a company) in a polite way for the purpose of making less direct commands or directives for company subordinates. However, in a professor's lecture, the same expression signalled a gap in knowledge that students should pay extra attention to. Biber and Baribieri (2007) gave a thorough description of how lexical bundles functioned to express stance, organize discourse, and frame referential information in both spoken and written discourse in university registers.

A few researchers have observed that the intuitions of language textbook authors can lead to distorted representations of target language use (Biber, 1999). With the availability of corpora, like BNC, which provides samples of natural language use across a variety of registers, ranging from conversations, fiction, newspapers to academic prose, linguists are able to carry out empirical studies to identify the gap of language use between pedagogical materials and language in use. Researchers have carried out studies investigating whether grammar rules presented in ESL/EFL textbooks actually reflect real language. For example, Römer (1996) suggested the mismatch of the meaning distribution of two modal verbs, *can* and *could*, between the Green Line series of textbooks used in EFL courses in German elementary schools and the spoken subcorpus
of the BNC. The Green Line suggested using *can* and *could* more often for the meaning of ability while the spoken sub-corpus of the BNC representing real spoken language indicates using the two modal verbs for the meaning of possibility. Many EFL/ESL grammar books state that the two common types of postverbal elements of *remember* are complement clauses with *to*-infinitive and –ing. However, the corpus-based study by Tao (2001) showed that the two common types of postverbal structure suggested by pedagogical grammar textbooks only accounts for 6% of all the uses.

In many advanced-level language courses at tertiary settings, the pedagogical focus has shifted from grammar to the development of discipline-specific language skills which students need for their studies. Advanced learners at the university level tend to have dual roles, being both language learners and discipline novices: they may take both English language courses and entry-level discipline courses. On the one hand, for students needing discipline-specific language learning, ESP textbooks offer authentic, discipline-specific readings as the main content and take a skill-based approach by including a variety of activities emphasizing key functions (making definition, describing processes, etc.) (McDounough, 1998).

On the other hand, the discipline introductory courses are supposed to equip the students with the core and basic concepts in the disciplines. In the discipline studies for those first-year novices, discipline entry-level textbooks are their primary readings (Carson, 2001) and important for the development of their academic literacy skills. Not only do those textbooks hold the position of authority in disciplines (Olson, 1989), they also explicitly introduce norms and basic concepts to the novices. Textbook authors take the role of discipline experts guiding readers to important lines of arguments and
propositional facts (Hyland, 2002a). With an increasing amount of research in textual analysis that indicates the strong association between language use and discipline specificity, researchers focus on exploring academic written texts (lab reports, textbooks, journal research articles) at different academic levels (undergraduate vs. graduate) for different groups of audience (discipline novices vs. experts). Specialized corpora have been set up for researchers to discover how formulaic language serves to construct text in particular disciplines, linking pragmatic functions of formulaic language with the broader context of the reader/author relationship. Findings of specialized corpus-based research can be used to inform discipline novices about the language use in their discipline studies, develop a deeper understanding of disciplinary text and help them distinguish what is meant from what is said or written.

Questions loom about whether the ESP textbooks reflect target language use and to what extent those pedagogical materials provide students with essential language guides relevant to their discipline literacy skills. Since reading materials in ESP textbooks determine the language input that the ESP students are exposed to in language classes and construct the basis of learning activities, it is essential to evaluate whether there is a gap in language use between ESP textbooks and discipline introductory textbooks. If there is such a gap, ESP teachers might consider including more target language in classroom teaching. Lexical bundles and their pragmatic functions in specialized corpora that are part of discipline-specific materials can also be used effectively in EAP and ESP classrooms to prepare for students’ entry into the discipline reading and writing appropriately.

1.1 Research questions

Previous evaluation of ESP textbooks or courses seems to focus more on issues of
whether the teaching materials have simulated tasks in real workplace or learning environments (Dlaska, 1999; McDounough, 1998). This type of evaluation does not consider whether the language used in ESP textbooks is consistent with target language use in introductory textbooks. My research explored potential differences between lexical bundles in English for Specific Purposes (ESP) textbooks and in Electrical Engineering (EE) introductory textbooks used by undergraduate students. This study addressed the following research questions:

1) What are the lexical bundles and their pragmatic functions in introductory textbooks of Electrical Engineering?

2) Do these lexical bundles occur in ESP textbooks? Do the pragmatic functions of the bundles differ?

3) Are lexical bundles and their pragmatic functions highlighted in ESP textbooks?

1.2 About this research

To answer these questions, I initiated a corpus-based study with the aim of identifying the lexical bundles and their pragmatic functions in ESP textbooks and EE introductory textbooks, adopting both quantitative and qualitative methods. My analysis was based on data from two corpora made up of materials respectively from ESP textbooks and EE textbooks. First, I used the frequency and distribution level of multi-word units as two criteria for the identification of lexical bundles in the two corpora. Wordsmith Tools (Ari, 2006) was used to count the frequencies of multi-word units and record the distribution levels of those units across different texts. I considered multi-word units as lexical bundles if those units show occurrence frequencies no less than the cut-off point of 20 times per million words, and at the same time are spread across no fewer than
five texts in the register. Second, I applied a qualitative method to analyze and compare pragmatic functions of those lexical bundles in original contexts.

With the completion of this study, we may obtain a relatively clear picture of how lexical bundles in Electrical Engineering introductory textbooks construct propositional facts and norms in the discipline. Findings from this research can help language instructors generate language teaching materials that incorporate awareness of the pragmatic use of lexical bundles in particular disciplines. Having an idea of the pragmatic functions of the lexical bundles, learners may improve their reading skills and have a better understanding of textbooks by distinguishing what is said from what is meant. They may also have a chance to observe how textbook authors use lexical bundles to set up particular reader/author relationships from which they might draw ideas when writing in the discipline. In addition, this research reports on language differences between ESP textbooks and EE introductory textbooks in terms of lexical bundles and their corresponding pragmatic specifications. Knowing these differences may help English instructors and material writers improve their awareness of lexical bundle use in specific disciplines so that they can take a critical view toward ESP textbooks, reconsidering what they can do better to meet their students’ needs.
Chapter 2 Corpus and Formulaic Language

A corpus is defined by Francis (1982) as "a collection of texts assumed to be representative of a given language dialect, or other subset of a language, to be used for linguistic analysis" (p. 7). Computer technology has rapidly pushed forward the development of corpora, allowing texts to be stored on a computer for qualitative and quantitative analysis (Conrad, 1996; O'Keefe et al., 2007). Not surprisingly, O'Keefe et al. (2007) incorporated this technical aspect into their definition of a corpus: "a corpus is a principled collection of electronic texts usually stored on a computer, available for qualitative and quantitative analysis" (p. 2).

In the past three decades, computerized corpora have rapidly moved from the first generation, characterized as small-sized and made up of sampled texts, to the second generation tagged as "mega-corpora", which are larger and include full texts (Flowerdew, 2004). One example of the first generation corpus is the half-million-word London-Lund corpus developed in the 1970s and 1980s, which sampled spoken language such as conversations and radio broadcasts. A good example of the second generation corpus is the 450-million-word Bank of English, evolving from several corpora, among which is the 40-million-word Longman Spoken and Written English (LSWE) corpus. The Bank of English is also called a monitor corpus due to its constantly growing size. The corpus size noted by Biber, Conrad, and Reppen (1998) was 100-million words; this has grown to 450-million words in recent years. Another example of the second generation corpus is the 100-million-word British National Corpus (BNC), which consists of both written and spoken language, representing language use mainly across four registers, news, fictions, conversations, and academic prose (O'Keeffe et al., 2007).
2.1 General corpus vs. Specialized corpus

Hunston (2002) described a general corpus as "a corpus of texts of many types", including "written or spoken language, or both (that) may include texts produced in one country or many" (p. 14). Both the first- and second- generation of corpora (the London-Lund corpus, the Bank of English) are general corpora since they represent language as whole and include a spread of texts from various language situations. The LSWE corpus is a general corpus, representing language use as whole. In order to identify grammatical variations across four primary registers: conversation, fiction, news and academic prose, Biber et al. (1999) developed the LSWE corpus with four core subcorpora, conversation, fiction, news, and academic prose. The conversation subcorpus represents natural conversation of British and North American speakers, who represent an even demographic distribution of age, sex, social group, and religion. The fiction subcorpus is a collection of major types of British and North America’s contemporary fiction. The news subcorpus includes texts on various topics (local news, social news, world news, etc.) from different newspapers aiming at different reader groups in British and North America. The academic prose subcorpus consists of book extracts and research articles from academic disciplines of sciences, social sciences and humanities. Most of the books are for professional readers who have experience in the trade, and the journal articles are also written for the same professional audience.

Representing language as whole, general corpora have been used for the compilation of dictionaries and grammar books. The Bank of English Corpus has facilitated the compilation of the Collins COBUILD English dictionary (Meyer, 2002; Sinclair, 1991). The LSWE corpus, a subset of the Bank of English, has provided Biber and his group
with an enormous amount of evidence, enabling them to construct a systematic description of grammatical structures across primary registers as described in their landmark book *Longman grammar of spoken and written English* (Biber et al., 1999).

Besides the large-scale description of language use in dictionary and pedagogical grammar books, applied linguists have used general corpora for the investigation of a defined number of lexical, grammatical, or lexico-grammatical aspects of language use. Römer (1996) has identified the inaccurate description of modal verb usage in an elementary textbook series used in German elementary schools after comparing the usage of nine central model verbs in the textbooks with the modal verb usage in the one-million-word British National Corpus (BNC). Oakey (2002) explored how the lexical phrase *it is/has been (often) asserted/believed/noted that X* implements pragmatic functions in academic writing of three disciplines (social science, medicine, and technical/engineering), represented by a subset of the BNC. Coniam (2004) studied his own writing through quantitative comparison of a range of linguistic features, including first person pronouns ("we" and "I"), passives and hedges, across three corpora: Academic A (published research articles in Applied Linguistics), Academic B (the author’s own writing related to teacher training), and the BNC written corpus.

However, general corpora may not satisfy the research needs of the group of linguists who have an interest in investigating language use in narrower domains. For instance, the BNC did not offer useful data for Charles (2007), who intended to compare the usage of the *noun + that* pattern in theses of political science and in material science. Flowerdew (2004) explained that there are four reasons why general corpora are not suitable for the investigation of specialized language. First is that general corpora may lack
representativeness of particular genres and registers since in general corpora a variety of
texts has been chose in a balanced way to guarantee the "representativeness of language
as whole" (p. 14). Tribble (1997) pointed out the pointlessness of using BNC for
investigation of the lexis in the health domain. He said, "in the written component of the
BNC sampler (1,000,000 words) there are no instances of "cancers". An Encarta ®
micro-corpus of health articles (24805 words) gives 33 usefully contextualized examples”
(p. 112). Secondly, a general corpus may contain specialized subcorpora, but their
accessibility was not the main consideration during the construction of the larger general
corpus. Thirdly, even when there are specialized subcorpora in a general corpus, the
subcorpora may not contain specific types of discourse that researchers are interested in.
Some types of discourse relevant to particular workplaces and professions are more likely
to be excluded from general corpora. In the case of a specialized corpus, the 100-million-
word Corpus of Professional English (CPE) (Noguchi, Orr, & Tono, 2006), 20% of the
corpus materials are from legal/workplace documents that normally can only be obtained
through professional networks in the field and thus semi-excluded from public access.
The final reason is that some general corpora contain samples of texts (the London-Lund
Corpus) rather than full texts; this may not satisfy researchers’ interests in examining
lexico-grammatical patterns across different sections of a text.

As we see, it is the four main characteristics of general corpora (general
representativeness of language as whole, difficult accessibility of subcorpora, limited
coverage of specialized text types, and truncated text samples) that prevent researchers
from using general corpora to investigate specialized language use. This explains the
need for specialized corpora to facilitate the investigation of language features in
particular domains.

Hunston (2002) described a specialized corpus as "a corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc" (p. 14). She also pointed out that a specialized corpus "aims to be representative of a given type of text" (p. 14). Compared with larger general corpora, specialized corpora maximally represent target language use in specific areas (O'Keefe et al., 2007) because compilers of specialized corpora often bring their professional/insider knowledge of a particular domain to bear on corpus compilation and data analysis procedures (Flowerdew, 2004). Consequently, specialized corpora represent specific language use in different modes (spoken language vs. written language), disciplines (political science vs. material science), and/or genres (research articles vs. academic textbooks). Specialized corpora also make detailed qualitative data analysis more manageable (Flowerdew, 2004), and thus can be used to capture language features with small amount of data (O'Keefe et al., 2007).

The size of specialized corpora can vary, depending on the broadness of language use domains that those corpora represent. Two examples of specialized corpora covering mutli-domains of language use are the Michigan Corpus of Academic Spoken English (MICASE) and the TOEFL 2000 Speaking and Writing Academic Language (T2K-SWAL) Corpus. The MICASE is a spoken language corpus, representing academic speech events across four major academic divisions (humanities and arts, social sciences, physical sciences and engineering, and biological sciences) (Simpson & Leicher, 2006). The 2.7 million-word T2K-SWAL Corpus includes various language use situations in both spoken and written registers in tertiary settings. The spoken subcorpus consists of a
variety of speech events on campus, such as classroom teaching, lab sessions, and study groups (Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay, & Urzun, 2004b). The written subcorpus contains texts from textbooks, course packs, and institutional materials (i.e. university catalogues). Researchers have explored those specialized corpora for studies of lexical, grammatical, and lexico-grammatical features in specific fields. For example, Simpson (2004) identified unique formulaic expressions in the MICASE, discovering that most of those expressions have pragmatically functioned as discourse organizers (i.e., *and in fact* functions as a focuser/introducer) and interactional markers (i.e., *and you can see* for explanation/demonstration) in academic speech.

Compared with the MICASE and the T2K-SWAL Corpus, small-scaled corpora (around 20,000-250,000 words) are more likely be used by researchers who investigate language features in very limited scopes. To investigate linguistic feature differences between biology textbooks and published biology research articles, Conrad (1996) set up a 32,000-word corpus with 20 text samples respectively chosen from two textbooks and tier one research journals. Charles (2007) recognized that the pattern of *noun+that* allows writers’ assessment of propositions to be integrated into the text and thus expresses writers’ stances. In order to examine how this pattern has been used to express stances in politics/international relations and materials science, she constructed two small-scale corpora by choosing eight full-length theses respectively from the two disciplines.

2.2 Representativeness: general vs. specialized corpora

A corpus, whether a general or specialized one, is assumed to have been designed to represent specific types of discourse. Representativeness is not only a key feature of a corpus, but also a key principle to follow during the construction of a corpus. Biber (1993) argued that to maximally represent language in use, corpus designers need to follow both
external and internal criteria, carefully evaluating text types and linguistic distribution within each text. He suggested using stratified sampling for selection of a probabilistic sample. In this method, "subgroups are identified with the target population (in this case, the genres), and then each of those 'strata' is sampled using random techniques " (p. 244).

He proposed a set of situational parameters as sampling strata, (see Table 2.1), with the purpose of defining "the target population as completely as possible "(p. 245).

**Table 2.1** Situational parameters listed as hierarchical sampling strata (Biber, 1993, p. 245)

1. **Primary channel.** Written/spoken/scripted speech
2. **Format.** Published/not published (+ various formats within 'published')
3. **Setting.** Institutional/other public/private/personal
4. **Addressee.**
   (a) Plurality. Unenumerated/plural/individual/self
   (b) Presence (place and time). Present/absent
   (c) Interactiveness. None/little/extensive
   (d) Shared knowledge. General/specialized/personal
5. **Addressor**
   (a) **Demographic variation.** Sex, age, occupation, etc.
   (b) **Acknowledgement.** Acknowledged individual/institution
6. **Factuativity.** Factual-informational/intermediate or indeterminate/imaginative
7. **Purposes.** Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, ..
8. **Topics....**
He also defined situational parameters of published writing, arguing that

Published writing always has unenumerated addressees, is always written for nonpresent addressees, and is almost always non-interactive (except for published exchanges of opinion). It can require either general or specialized background knowledge (e.g. popular magazines versus academic journals) but rarely requires personal background knowledge (although this is needed for a full understanding of memoirs, published letters, diaries, and even some novels and short stories) (p. 246).

In addition to carefully balancing various text types within a corpus, designers of a representative corpus should also take into account linguistic representativeness, which "depends on issues such as the number of words per text sample, the number of samples per text, and the number of texts per text type" (Biber, 1993, p. 243).

When setting up a general corpus representing language as whole, corpus compilers use the method of stratified sampling to collect spoken and written texts, aiming to establish a balanced general corpus. The 100-million-word British National Corpus (BNC) has been designed in a balanced way to "characterize the state of contemporary British English in its various social and generic uses" (Aston & Burnard, 1998, p. 28). The BNC compilers had a plan for text selection in advance, arranging a variety of ratios to certain types of texts to be included in the corpus (Aston & Burnard, 1998). The written texts in the BNC account for 90% of the corpus, and spoken texts 10%. Both the spoken and written texts have been chosen in a balanced way. Written texts in the BNC were selected according to three parameters: domain, time, and medium. Informative texts account for 75% of the corpus, imaginative texts 25%. Informative texts cover areas in arts, natural and pure science, commerce and finance, and applied science. Only materials from
contemporary sources were selected. Informative texts were chosen from those written after 1975, and imaginative texts from 1960 onward. The three main types of text medium are books, periodicals and miscellaneous resources, respectively accounting for 60%, 30%, and 10% of the corpus. Miscellaneous resources include published texts (advertisements, brochures, etc), unpublished texts (letters, memos, etc), and written-to-be-spoken scripts (scripted television materials).

Similar to general corpora, specialized corpora covering multi-domains such as the MICASE and the T2K-SWAL Corpus have been structured in a balanced way to represent language use in narrower scope. The T2K-SWAL corpus represents written and spoken registers in university settings, and the MICASE contains spoken language in the academic community. During the construction process of the MICASE, around 200 hours of academic speech events were recorded and transcribed at the University of Michigan between 1997 and 2001. Academic speech events were divided into classroom events (lectures, discussion sections, seminars and students presentations), and non-classroom events (advising sessions, colloquia, dissertation defenses, etc.) (Simpson & Leicher, 2006). A range of factors, such as levels of interactivity (ranging from monologic to highly interactive), academic level of the class (undergraduate vs. graduate) and speaker variables (academic role, language status, age group, gender) have also been taken into consideration in the building-up process of the MICASE.

The T2K-SWAL corpus consists of spoken and written subcorpora. In the spoken subcorpus, texts from classroom teaching, accounting for more than 70% of the subcorpus, are balanced along three levels of interactiveness (low-, medium-, and high-level of interactiveness). In the written sub-corpus, textbook materials account for over
75% of the corpus, and are balanced along dimensions of disciplines (business, education, engineering, humanities, natural science, and social science) and education levels (lower undergraduate, upper undergraduate, and graduate).

The principle of stratified sampling recommended by Biber (1993) seems to have worked for both general corpora and some specialized corpora. However, the case may not be the same for all small-scale specialized corpora. Based on quantitative analysis of linguistic features, Williams (2002) suggested that internal criteria are an objective approach to guarantee the representativeness of specialized corpora. However, most of the small corpora compilers have applied external criteria, regarded as somewhat subjective, for text selection of their corpora. Consequently, compilers of small corpora commonly deal with issues such as selection of text types, involving their intuition and judgments of who produced the text, who the audience would be, and how useful the content of the corpora is regarding to the ongoing research questions.

Although larger corpora compliers have used internal textual criteria to ensure the representativeness of sampled corpus materials (Aston & Burnad, 1998), compilers of small specialized corpora tend to take a more ethnographic approach by involving specialist informants of the discipline in corpus compilation (Conrad, 1996; Cortes, 2004; Hyland, 2004). To ensure the representativeness of small corpora, researchers have relied on insiders’ insights in selecting materials for their corpora. Cortes (2004) noted that one of her main research goals was to identify lexical bundles (recurrent word clusters in multi-texts of a corpus) and their pragmatic functions in published research articles of history and biology. So she interviewed professors from both history and biology departments for their opinions on appropriate selection of journal articles in a specialized
corpus. Using corpus-based technique, Conrad (1996) explored linguistic variations between biology textbooks and articles in biology journals. The corpus that she set up includes 40800-word samples, respectively drawn from biology textbooks and journal articles. The 20 written samples of the textbook sub-corpus came from two introductory biology textbooks used in an Ecology course for second-year biology students. Conrad argued that the samples from the two textbooks exhibit enough representativeness since previous research proved that those textbooks have linguistic features similar to other biology-relevant textbooks. Besides, Conrad noted that the two textbooks were commonly and repeatedly used by undergraduate students. These two merits ensured that the textbook corpus offered "a solid basis for examining the textual variation faced by introduction ecology students at this university" (p. 303).

2.3 Formulaic language and specialized corpora

Specialized corpora have become the main enquiry area as well as a tool for the exploration of formulaic language. Samples of authentic spoken and written texts have been collected for the empirical investigation of formulaic language in specific genres and registers. Through concordance lines and frequency lists of formulaic language, researchers use computer software (i.e. WordSmith Tools 4.0) to make direct observations of formulaic language.

O'Keefe et al. (2007) found that the acquisition of native-like collocations created great difficulties for English language learners, who asked questions such as "why is something expressed the way it is?" and were not satisfied with responses by native speakers such as "that's the way we say it." Bearing in mind the research goal of exploring formulaic language in natural spoken discourse, O'Keefe et al. took the same
position as Sinclair’s (1991) dichotomous view of language, proposing that language consists of the idiom principle and the open-choice principle, with a focus on the “idiom” principle. Prefabricated chunks, rather than syntax, are put at the center of language patterning since researchers believe that “comparison of chunks across different data sets can reveal interesting ‘fingerprints’ of particular text type, modes of communication, or group of uses” (p. 62).

O’Keefe et al. (2007) compared the pragmatic functions of recurrent chunks in the Cambridge and Nottingham Corpus of Business English (CANBEC) with those in academic segment (ACAD) of Cambridge and Nottingham Corpus of Discourse in English (CANCODE). The CANCODE is a 5-million-word collection of spontaneous speech across British Isles covering a wide variety of situations with different levels of formality, ranging from casual conversations to spoken language in academic settings. The researchers found that pragmatic functions of formulaic chunks highly correlated with the original contexts. For example, we need to was found to have a high frequency in both corpora, CANBEC and ACAD; however this chunk serves different functions in different registers. In CANBEC, we need to is used to frame "corporate requests for information and for action issued by individuals with authority" (p. 213). As an indirect form, it is "protecting face and less direct than potentially face-threatening demands or directives"(p. 213). However, in ACAD, we need to introduces new topics and mainly refers to gaps in students’ knowledge during professors' lectures.

The above studies indicate that a happy marriage exists between specialized corpora and formulaic language, with corpora allowing for the exploration of formulaic language in specific registers and genres. The following section provides us an overview of
literature that furthers our understanding of the significance as well as the main features of the prefabricated/semi-prefabricated part of our language.

2.4 Formulaic language

Formulaic language, commonly recognized as “multiword units of language that are stored in long-term memory as if they were single lexical units” (Wood, 2002, p. 2), has become the center of many corpus-based studies. Researchers from second language acquisition (Wray, 2002), psycholinguistics (Schmitt, 2004), and second language writing (Altenberg & Granger, 2001) are interested in analysis of collocations and patterns in written and spoken text. These linguists share the same insight about language: our use of language is a system consisting of generative rules and prefabricated chunks, with speakers at times retrieving those chunks without depending on generative rules.

Although linguists Krashen and Scarcella (1978) believed that prefabricated language plays a minor role in language use, Nattinger and DeCarrico (1992) have argued that one can not deny the predictability and routine of daily language. Even language learners resort to prefabricated language in order express situations beyond their level of linguistic competence (Wray, 2002). Research evidence shows that many learners have experienced the stage of item-learning before proceeding to the system-learning stage. Children in the early stages of language acquisition use unanalyzed chunks to meet specific social needs (e.g. I-want-to-go indicates the child’s desire). Nattinger and DeCarrico argued that memorized chunks are used by children for “later segmentation and analysis in developing the rules of syntax” (p. 25).

Wray (2002) mentioned that two main factors, processing constraints and the interaction of encounter and emulation in communicative situations, seem to determine
the predominance of certain chunks over others in speech. Speakers tend to use formulaic chunks to relieve processing stress and thus save processing time when they speak. Clusters of phrases, word strings, or even longer sentences are remembered by humans to help them facilitate social activities. Wray (2002) maintained that formulaic language serves to relieve the time-demanding pressure in oral traditions, sport commentaries, and auctions. Processing pressure seems to be one reason why formulaic sequences are widely used in these three situations. Oral traditions require poets to improvise content, sport commentators in live programs are pressured to describe fast action on sport fields in a rapid and accurate way, and auctioneers have to quickly and precisely deal with each transaction. Speakers under such enormous pressure use formulaic language to reduce processing time.

Wray and Perkins (2000) explained in detail how formulaic language is used as a tool for social interaction. Speakers use formulaicity for manipulation of others and assertion of separate and group identities. Commonly used in daily conversation, commands (keep off the grass), requests (could you repeat that please?), polite markers (I wonder if you’d mind) and other examples that signal the physical, emotional, and cognitive needs of speakers. Speaker often use personal turns of phrase (you know what I mean, Mary) or story-telling devices (You’re never going to believe this, but) to separate themselves from the crowd or to attract listeners’ attention to topics that the speakers are interested in. In addition, influenced by their current speech community, speakers acquire the implicit ability to make themselves sound like others. They learn what sequences of words sound culturally appropriate, pull those strings out of their linguistic pools and deliver them to the hearers. The hearers filter the received message and may then store
those idiomatic sequences.

Due to the central role of formulaic language, researchers have managed to propose descriptions of formulaic language units in a systematic way (Wood, 2002). These descriptors are termed as "lexical phrases" (Nattinger & DeCarrico, 1992), "formulaic sequences" (Wray, 2002), "lexical bundles" (Biber et al., 1999), and so on. The descriptors have been proposed to capture the "multi-faceted nature of formulaic language" (Wray and Perkins, 2000, p. 3) in terms of its structural, semantic, and functional features.

2.4.1 Forms of formulaic language

Formulaic language can take canonical or non-canonical forms, with the length varying from short phrases to whole sentences. Nattinger and DeCarrico (1992) specified lexical phrases as prefabricated chunks with associated functions, which can be classified into four categories according to structural criteria:

1. polyword: in a nutshell, so far so good
3. Phrasal constraints: a _ ago, to _ this up, as I was _
4. Sentence builder: I think (that) ..., not only X, but also Y

In the above examples of lexical phrases, speakers and writers use formulaic language in various forms, ranging from totally fixed to allowing variations in certain slots, occurring at lexical or/and syntactic levels.

Lexicalized sentence stems noted by Pawley and Syder (1983) can be fixed idioms, such as a stitch in time saves nine. However, most of the lexicalized sentence stems are conventional expressions, allowing variations at specific grammatical slots. Pawley and Syder gave an example below, expressing apology in daily conversations:
1. I'm sorry to keep you waiting.
2. I'm sorry to have kept you waiting.
3. Mr X is sorry to keep you waiting all this time. (p. 210)

The lexicalized sentence stem in above expressions is NP be-TENSE sorry to keep-TENSE you waiting. Lexical variation occurs in the slot of NP, where first person pronoun I and third person pronoun Mr X take the position. Morphological variations happen in two other slots, be-TENSE and keep-TENSE, where we see agreement of copula be with the third person singular noun in the third sentence and perfect progressive of keep in the second sentence.

Wray (2002) also observed that formulaic chunks are both presented in fixed forms (Nice to see you) or in semi-preconstructed phrases (NPi set + TENSE POSSi sights on (V) NPj, I’ve set my sights on winning that cap.), which permit variations at lexical or/and syntactic level. Those variations include morphological changes (it’s been/it will be a devil of job), changes of content words (a piece/slice of the action), or variations of open class items (give SOMEONE a piece of ONE's mind). At the syntactic level, formulaic language allows syntactic variation, such as passivization, without losing the original meaning (Wray, 2002). For example, spill the beans can be transformed to the beans are spilled, without the former meaning changed.

2.4.2 Compositionality of formulaic language

Besides having the continuum of fixedness in its form, formulaic language allows “a continuum of semantic compositionality” (Wray, 2002, p. 57). At one end of this continuum, there are free collocations with complete semantic transparency whose meanings can be decoded with “minimal contextual cues” (O’Keefe et al., 2007, p. 84) (e.g. get the message). At the other extreme, idioms, proverbs, and sayings have holistic
and institutionalized meanings that can not be predicted from the literal word-by-word meanings (e.g. step on the gas, kick the bucket) (Howarth, 1996; Moon, 1998a; Nattinger and DeCarrico, 1992). Non-compositionality can also be true with the case of non-idiomatic expressions, where people are more likely to think of the practical meanings, motivated by pragmatic functions of those expressions, rather than the literal meanings decoded by individual words in the expressions (Moon, 1998a). For example, in many cases expresses epistemic modality, recognized as a hedge of the following proposition, as in:

In many cases, Hume says, actions are virtuous because of the motives which produce them. (OHPC: non-fiction). (italics added, quoted from Moon, 1998a, p. 229)

2.4.3 Pragmatic functions of formulaic language

Wray (2002) and other researchers have noted that formulaic language not only shortens processing time in both spoken and written discourse but also accomplishes pragmatic functions in the discourse (Pawley and Syder, 1983; Wray, 2000, 2002; Wray and Perkin, 2000). Wray (2002) suggested that the pragmatic meaning of formulaic chunks associates a linguistic event with a specific situation. Contextual factors play a key role in the formation and use of formulaic language; as Wray explained, “we shall ask what the situation is, and what meaning is to be expressed, and then ask which formulaic sequence is most likely to be selected for the job” (p. 59). Pawley and Syder (1983) expressed a similar opinion on the provenance of formulaic language and suggested that additional knowledge about lexicalized sentence stems underlies our knowledge of generative grammar. Distinct from conventional sentence stems, a lexicalized sentence stem has the important feature of being a social institution. “In saying that a lexical item is a social institution we meant that the expression is a
conventional label for a conventional concept, a culturally standardized designation (term) for a socially recognized conceptual category” (p. 209).

Nattinger and DeCarrico (1992) agreed that it is important to include pragmatic competence and socio-pragmatic competence in a description of language. They noted that besides grammatical competence, our knowledge of language and our language use abilities should include “knowing how to select and retrieve ready-made form/function composites (pragmalinguistic competence) for appropriate situations or contexts (socio-pragmatic competence)” (p. 13). In specific language use context, only those chunks with pragmatic appropriateness of the context are selected from a pool of possible prefabricated chunks. For instance, a speaker may prefer to choose *if I were you* to give personal advice.

2.4.3.1 An introduction to formulaic language in spoken and written discourse

The “striking” differences between formulaic language in spoken and written discourse reflects varying communicative practices in the two areas. Compared with spoken discourse, writing is characterized by integration and detachment. “The integration that characterizes written language occurs generally through complexity of noun groups, conjoined parallel phrases, sequences of prepositional phrases, and relative clauses” (Nattinger & DeCarrico, 1992). Nattinger and DeCarrico proposed the concept of lexical phrase, a lexico-grammatical unit associated with particular pragmatic function(s), with varying lengths of phrases or clauses (i.e., on the other hand, the _er X, the _er Y).

Lexical phrases in both spoken and written discourse have been grouped into three categories according to their pragmatic functions: social interactional markers, necessary
topic markers and discourse device markers. Social interaction markers describe social
relations and are used by speakers to maintain conversations through summoning (*good
morning*), nominating a topic (*Have you heard about X*), or clarifying the information
spoken or heard (*pardon me*). Social interactional markers can describe conversational
purpose, expressing politeness (*thanks*), requesting responses (*May I X*?), and so on.
Necessary topic markers (i.e.,) introduce common topics such as quantity (*lots of _*), time
(*when is X ?*), locations (*where is X ?*), etc. To make discourse flow, discourse device
markers act as logical/temporal/spatial connectors (*in spite of X, and then, over there*),
exemplifiers (*for example*), evaluators (*there's no doubt that*), summarizers (*in a nutshell*),
etc.

Spoken discourse has a larger number of discourse device markers than in written
discourse (Nattinger and DeCarrico, 1992). For the same functional types of discourse
device markers, written and spoke discourse have their own sets of phrases. For instance,
both spoken and written discourse use logical connectors. *As a result* and *therefore*, two
logical connectors commonly seen in written text, indicate causative relationship. In
spoken discourse, the same logical relationship is represented by a large group of
discourse device markers, including *so, this means that X, the upshot is X, so then, what
happens is X*, and more.

2.4.3.2 Formulaic language in conversational and transactional spoken discourse

Nattinger and DeCarrico (1992) found that social interactional markers and
necessary topic markers are dominant in conversation while discourse device markers are
most commonly seen in transactional spoken discourse, such as academic lectures and
teacher/student conferences. In my opinion, the common sense of what academic speech
looks like seems to contradict the conclusion made by Nattinger and DeCarrico (1992) who suggested that there are more necessary topic markers in conversation. My understanding is that the main duties of university teachers are to transfer knowledge to students. Especially in the disciplines of hard science, teachers need to describe research processes and provide accurate information of time, location, quantity, which can be framed by necessary topic makers. Why do instructors use fewer topic markers than these are found in daily conversation? The lack of empirical evidence in Nattinger and DeCarrico’s (1992) study seriously weakens the reliability of their analysis regarding the functional differences of formulaic language between daily conversation and transactional discourse. Their conclusions seemed to be based on qualitative analysis of a limited number of transcripts of an unknown data source since issues such as the source of the data, research method, and representativeness of their data remained as mysteries to the readers.

Corpus-based research offers another picture of how formulaic language functions differently in academic speech and conversation. Based on her corpus-based studies, Simpson (2004) suggested that formulaic expressions function mainly as discourse organizers and interactional markers in academic speech, which is inconsistent with what Nattinger and DeCarrico (1992) suggested, namely, the predominance of discourse device markers in transactional spoken discourse. Instead, Simpson argued that discourse organizers (equivalent to discourse device markers in Nattinger & DeCarrico’s study) and interactional markers play equally important roles in academic speech. Discourse organizers in classroom speech have a variety of functions, serving as focusers/introducers (and in fact), meta-discourse markers (you could say),
enumerators/temporal sequencers (*first of all*), markers of contrast and comparison (*in the same way, at the same time*), cause-effect markers (*the reason why*), and summarizers (*it turns out that*). Interaction expressions develop explanations and demonstrations (*and you can see*), command listeners or give them advice (*look at this*), or hedge statements made by speakers (*it seems like*). The research also mentions formulaic expressions specifying location, time, and space (*in the book, on the left, at the end*).

Simpson (2004) simply depended on her intuition of idiomaticity to locate formulaic expressions from a frequency list of computer-generalized multi-word strings. The problem with this approach is that the researcher's judgment of formulaic expression may not be accurate and thus is not likely to offer the overall picture of formulaic expression in academic speech. Unlike Simpson (2004), Biber et al. (2004a) identified lexical bundles on a purely statistical basis. Only multi-word clusters that reached specific frequency- and distribution- criteria were considered as lexical bundles (more details regarding identification of lexical bundles will be discussed in the next chapter). The study of lexical bundles by Biber et al. draws a relatively comprehensive picture of language use in conversation and academic speech, being consistent with the communicative focuses of the two types of spoken discourse. The former type uses stance bundles to emphasize the equal, peer-like communication with their listeners. The latter type favors using more discourse organizers and enables speakers, in most cases teachers, to keep students focused on the discourse; it also uses stance bundles to express speakers' confidence in the content that they deliver and to guide students through academic lectures or seminars.

Both conversations and classroom teaching contain stance bundles that express
epistemic certainty/uncertainty or attitudinal stances towards specific statements (Biber et al., 2004a). However, instructors use stance bundles to establish their authority while in conversation the relationship between interlocutors tends to be equal. Compared with classroom discourse, conversations tend to use more epistemic stance expressions (i.e., *I don't know how, I don't know if, I don't know what, I think it was*) for the purpose of expressing personal certainty/uncertainty, in most cases uncertainty, towards following propositions.

In classroom teaching, more directive expressions (i.e., *I want you to, you don't want to*) are used to urge students to carry out actions expected by instructors. Bundles with directive forces are not common in conversation. In addition, classroom teaching has more discourse organizers than conversations do, reflecting the communicative need that speakers attend to listeners in classrooms. Those discourse organizers mainly introduce new topics (*what do you think, if you look at*) or elaborate current topics (*has to do with, I mean you know*).

In the above, I discussed pragmatic functions that formulaic language carries out in two sub-types of spoken discourse: conversation and academic speech. In written discourse, a different collection of formulaic expressions exists (Biber et al., 2004a; Hyland, 2008). This is the concern in the next chapter where I focus on the pragmatic functions of lexical bundles in textbooks and research articles across a variety of disciplines.
Chapter 3 Lexical bundles

Lexical bundles, "sequences of word forms that commonly go together in natural discourse" (Biber et al., 1999, p. 990), were first identified by Biber et al. (1999) through quantitative computer analysis of the 40-million-word Longman Spoken and Written English (LSWE) Corpus, which represents language in four registers, conversation, fiction, news and academic prose. It is almost impossible to ignore lexical bundles due to their considerable proportions in discourse across the two main registers (conversation and academic prose) of the LSWE corpus (Biber et al., 1999): 28% of words in conversation occur in 3- and 4-word lexical bundles, and 20% of words in academic prose.

Since I aim to investigate how lexical bundles function in textbooks of English for Specific Purposes and in Electrical Engineering textbooks, it is worthwhile to take a look at previous literature on lexical bundles. In this chapter, I first explain that lexical bundles are different from conventional phraseological units. Next, features unique to lexical bundles are introduced. I then focus on the grammatical structures of lexical bundles and point out the strong association between grammatical structures and pragmatic specialization. In the next section, I describe the functional taxonomy developed by Biber et al. (2004a), based on their study of lexical bundles across four registers, classroom teaching, textbooks, conversation and academic prose. This chapter ends with a discussion of issues relevant to lexical bundles, such as frequency and formulaicity and other researchers’ study of pragmatic functions of formulaic chunks.

3.1 Lexical bundles are not phraseological units

Lexical bundles are characterized by their significant statistical features, or as Biber
et al. (1999) first described them "bundles of words that show a statistical tendency to co-occur" (p. 989). Consequently, lexical bundles are different from pure idioms or multi-word fixed expressions commonly seen in dictionaries (Moon, 1998a), which have been the conventional research areas of phraseology (Howarth, 1996; Moon, 1998a). Pure idioms are defined by Moon (1998a) as "fixed and semantically opaque or metaphorical" (p. 4), like kick the bucket, and spill the beans. These pure idioms and fixed expressions are phraseological units that have been investigated under different rubrics, such as composite units (Howarth, 1996) and fixed expressions and idioms (FEIs) (Moon, 1998a) by researchers interested in the syntactic, grammatical and semantic restrictions on phraseological units. For instance, Moon studied FEIs along three main dimensions: institutionalization, and lexicogrammatical fixedness, and non-compositionality.

Institutionalization is "the process by which a string or formulation becomes recognized and accepted as a lexical item of the language (Bauer 1983: 48 and passim)" (Moon, 1998a, p. 7). Institutionalized expressions are often conventional phrases, such as go to school/church. As they have inevitably been the product of language evolution in certain speech communities, both geographical and historical factors have had an impact on the process of institutionalization. For example, swim between waters (meaning: be impartial) was once a conventionalized expression no longer in use (Moon, 1998a). Lexicogrammatical fixedness reflects a certain degree of ill-formedness of FEIs since the level of fixedness decreases when the FEIs are grammatical both on the lexical and syntactic levels. An example is by and large which is frozen, allowing no variation at any slot in the phrase (Moon, 1998a). Canonical FEIs were also investigated for their syntactic patterns and the results showed a wide-spread syntactic diversity among FEIs.
Nomination is one type of syntactic transformation. One form of nomination is where an FEIs is condensed to one of its components, maintaining the allusions attached to the original expressions. The expression *a new broom sweeps clean* was not found in Moon’s (1998a) corpus data; however *new broom*, the truncated form, was found to be used in newspaper articles, maintaining the metaphorical meaning of the longer expression. Non-compositionality refers to the fact that a string, when treated as a whole, has its institutionalized meaning that does not arise from word-to-word interpretation (i.e., sayings) (Moon, 1998a).

Since lexical bundles are sorted out mainly by the statistical feature of high frequency, they are not constrained by the three main features used to describe phraseological units. To have a better understanding of what lexical bundles are, it is worthwhile to have an overview of the available literature on lexical bundles.

**3.2 Research on lexical bundles**

Biber et al. (1999) used a frequency- and distribution-based methodology to identify lexical bundles. For word sequences to be qualified as lexical bundles, they need to satisfy both the frequency and distribution level criteria. The minimal cut-off frequency varies from 10 times per million words for three- or four- word sequences to 5 times per million words for five- or six- word sequences due to the fact that three- or four- word sequences are far more frequent than five- or six- word sequences. With the purpose of avoiding idiosyncrasies by individual speakers/writers, the criterion of distribution level for occurrences of word sequences requires that “these occurrences must be spread across at least five different texts in the register” (p. 992). Biber et al. also
investigated the grammatical structures of lexical bundles across four registers, and briefly explained discourse functions of those bundles in their original context.

Biber, Conrad, and Cortes (2003) initiated a functional taxonomy of 4-word lexical bundles in speech and writing. Four-word lexical bundles, such as *I don't know what* (in classroom teaching), and *on the other hand* (in textbooks) were first identified and then categorized into three main categories: referential expressions, discourse organizers and stance expressions. Cortes (2004) applied the initial taxonomy to discipline-specific texts and demonstrated the discourse functions of lexical bundles in history and biology journal articles.

Biber, Conrad, and Cortes (2004a) expanded the functional taxonomy (Biber et al., 2003) by including TOEFL 2000 Spoken and Written Academic corpus (T2K-SWAL) in their data. This new taxonomy offers a more complete picture of how bundles carry out pragmatic functions across university spoken and written registers, including classroom teaching, textbooks, conversation, and academic prose. Using this taxonomy as the main analytic framework (Biber et al., 2004a), Biber & Barbieri (2007) expanded the scope of lexical bundles by adding more university spoken and written registers. The newly added spoken registers are class management, office hours, study groups and service encounters, and the new written registers are course management and institutional writing. The study by Biber & Barbieri (2007) once again confirms that lexical bundles are pervasive in university language, used in particular sets of registers to serve specific discourse functions.

### 3.3 Features of lexical bundles

Recurrent word combinations, ancestors of lexical bundles, were first identified
through a frequency-driven approach by Altenberg (1998), who noticed the three main features of recurrent word combinations: semantic transparency, fragmental grammatical structures and pragmatic specialization. Altenberg (1998) examined recurrent word combinations in the London-Lund Corpus of Spoken English with the goal of determining if there is certain amount of prefabricated language in the spoken corpus. He set the minimum frequency threshold at ten times per million and limited his investigation scope to recurrent word combinations with the length of three or more. For instance, *I don't think that, do you know, on the other hand, yes I did* were identified as recurrent word combinations. A look at these bundles demonstrates that the majority of those combinations are transparent in meaning, unlike pure idioms; these bundles cannot be paraphrased with a single word as is the case for many idioms (i.e., *look forward to* = *anticipate*).

It was also found that those combinations obviously do not have complete grammatical structures. Altenberg's results showed that only 10 percent of recurrent word combinations took the form of full clauses, and the majority of the word combinations have incomplete grammatical structures, with 76% being clause constituents and 14% incomplete phrases.

Biber et al. (1999), following Altenberg's (1998) idea of recurrent word-combinations, proposed the concept of lexical bundles, which "can be regarded as extended collocations: bundles of words that show a statistical tendency to occur" (Biber et al., 1999 p. 989). Besides adopting the frequency criteria first established by Altenberg (1998), Biber et al. (1999) introduced the distribution criteria to avoid individual idiosyncrasy in the identification process of lexical bundles.
Undoubtedly lexical bundles are considered a modified version of recurrent word combinations, and thus they inherit the three main features characterizing recurrent word combinations, semantic transparency, fragmental grammatical structures and pragmatic specialization. Since semantic transparency is an obvious and easy-to-understand issue, as seen in the examples in the first paragraph of this section, the following section focuses on grammatical structures and pragmatic specialization of lexical bundles.

3.4 Grammatical structures of lexical bundles

Biber et al. (1999) found that lexical bundles in different registers tend to have particular structural types. They also argued that the structure types of lexical bundles have strong association with particular registers. Forty-four percent of lexical bundles in conversations take the form of verbal and clausal units, as in I don’t know why and I thought that was. However, lexical bundles in academic prose are more likely to be nominal with the commonly seen form as noun phrases with parts of prepositional phrases, such as the nature of the and the size of the in academic writing.

Even within the same register, say academic prose, the most widely used structural type of lexical bundles in published texts varies among disciplines. Cortes (2004) investigated 4-word lexical bundles in professional writing of history and found that the most frequently used grammatical structure is a prepositional phrase with an embedded ‘of’ phrase (i.e. as a result of, as part of a). Second comes the structure taking the form of a noun phrase with ‘of’ phrase fragment (i.e., the history of the, a member of the). However, in published biology writing, the pattern of noun phrase with ‘of’ phrase fragment is the most widely used, as seen in lexical bundles such as a function of the, a large number of, and so on.
Despite the differences in grammatical structures of lexical bundles, those bundles share a commonality, the incompleteness of their grammatical structures. It is unusual to see a lexical bundle that has a complete grammatical structure; as Biber et al. (1999) noted, “lexical bundles extend across structural units” (p. 991). The incompleteness of grammatical structure in these lexical bundles has been associated with the nature of the bundles as “more or less prefabricated or routinized building blocks that are at the speaker’s disposal in the production of discourse” (Altenberg, 1998, p. 115). Altenberg’s consideration of speech production as a process of linear composition was based on the observation that most lexical bundles have fragmental structures of multiple clause constituents. O’Keefe et al. (2007) recognized that the fragments are “important clues as to how interaction unfolds and how grammar emerges rather than being pre-existent in interaction” (p. 70). Biber et al. (1999) also considered lexical bundles as frameworks of spoken discourse, and argued that specific content can be added in the following slots to meet communications needs under various situations, as shown in the following examples listed below

*I don’t know why he didn’t play much at the end of the season.*

*I don’t know why Catherine finds that sort of thing funny.*

(Biber et al., 1999, p. 991, the lexical bundles are in italics)

The next section considers the pragmatic specialization of lexical bundles. First I discuss pragmatic specialization of recurrent word combinations, ancestors of lexical bundles. Next, I will focus more on the function taxonomy of lexical bundles mentioned in Biber et al.’s (2004a) landmark article, used as the main analytical framework of this research.
3.5 Pragmatic specialization of lexical bundles

Altenberg (1998) reported pragmatic specialization of recurrent word combinations in spoken discourse. He argued that those combinations are closely associated with particular speech situations and can only be fully interpreted in the context. For instance, *I should think* actually functions as personal evaluation of a situation, and *so I said* functions in the speech context to initiate a report. In addition, he noted that "Each expression has a distinctive conventionalized discourse function, serving to express the speakers' view of how an utterance is to be interpreted" (p. 110).

Biber et al. (2003) developed a functional taxonomy of lexical bundles in speech and writing. Lexical bundles were first identified in three sub-corpora of the Longman Spoken and Written English Corpus, the 4-million-word corpus of English conversation, the 3-million-word corpus of American English conversation, and the 5.3 million-word academic prose corpus of both American and British English. Next the primary functions that those lexical bundles performed were grouped into four main categories: referential bundles, text organizers, stance bundles and interaction bundles. Those four categories construct the framework of the function taxonomy, representing discourse functions realized by the lexical bundles in two registers, conversation and academic prose. Referential bundles “make direct reference to the physical, temporal, and contextual context” (Biber et al., 2003, p. 81). Text organizers refer to the lexical bundles that indicate the logical relationship between discourse segments. Stance bundles are used to indicate evaluation of following propositions that express either epistemic certainty or attitude/modality.

Based on the initial functional taxonomy of lexical bundles mentioned above, Biber
et al. (2004a) expanded the scope of the taxonomy of lexical bundles to four registers, classroom teaching, textbooks, conversation, and academic prose. Based on the analysis of two corpora, the TOEFL 2000 Spoken and Written Academic Corpus (T2K-SWAL) (for main details of corpus composition, see section 2.2) and the three main sub-corpora of Longman Spoken and Written English Corpus (see the composition details of this corpus in the above paragraph), four-word bundles and their discourse functions were studied. Those bundles have frequencies higher than 20 times per million words, occurring in at least 5 different texts.

Biber et al. (2004a) categorized lexical bundles that perform the same discourse function into the same group and identified three primary functions of lexical bundles: (1) stance expressions, (2) discourse organizers, and (3) referential expressions (See Fig. 3.1). Fig. 3.1 also shows the sub-categories which have realized more specific functions.
Figure 3.1 Functional taxonomy of lexical bundles across four registers (Biber et al., 2004a)
3.5.1 Stance expressions

Biber et al. (2004a) see stance bundles as being used to “express attitudes or assessments of certainty that frame some other proposition” (p. 384). They divided these bundles into two major semantic categories: attitudinal/modality stance bundles vs. epistemic stance bundles.

Within each category, the bundles are labelled either personal or impersonal, based on whether it is the writer/speaker or a broader audience who express (es) stance. Most personal epistemic stance bundles (i.e., I don’t know if, I don’t know I, I don’t think so) occur in classroom teaching and conversation to express uncertainty, as in the following sentence:

\[
I \text{ don’t know if it will mean revolution in the same sense of the word, I don’t think so because I think there are other political factors involved. (classroom teaching, Biber et al., 2004a, p. 389)}
\]

In the opposite way, impersonal epistemic stance bundles (are more likely to, the fact that the) are often used to exhibit certainty, as in:

Yet there was irony in the fact that the Russian Revolution, one of the most important Western revolutions, proclaimed itself to be Marxist in aims and character but happed in violation of Marxist historical logic. (textbook, Biber et al., 2004a, p. 390)

Attitudinal/modality stance bundles express “speaker attitude towards actions or event described in the following proposition” (Biber et al., 2004, p. 390). Those bundles are divided into four major sub-categories, expressing desire (i.e., if you want to, I don’t want to), obligation/directive (i.e., I want you to, you have to be, it is important to), intention/prediction (I’m not going to, it’s going to be), and ability (i.e., to be able to, can be used to).
It is interesting that stance bundles are commonly seen in classroom teaching and conversation. Only a total of five stance bundles were identified in textbooks and academic prose in Biber et al. (2004a). Those stance bundles carrying impersonal tones are rare in textbooks and academic prose (Biber et al., 2004a; Biber & Barbieri, 2007). They are used for particular purposes such as showing degrees of certainty (i.e., the fact that), “directing the reader to carry out some action” (i.e., it is important to), or indicating the possibility from an objective perspective (i.e., can be used to) (Biber et al., 2004a).

Within academic prose, discipline variations in language use have been exhibited by Cortes’ study (2004) of pragmatic specialization of lexical bundles in history and biology, where she set up a specialised corpus made up of published journal articles in the two disciplines, history and biology. No stance markers were identified in academic history writing. In biology research articles, like the written discourse of textbooks and academic prose (Biber et al., 2004a), the number of stance bundles was much fewer than that of referential bundles and discourse organizers. Biologists used stance bundles (i.e., it is possible that, it is likely that, has been shown to, was found to be) in impersonal form to express their stance without personal attitudes or assessment, and thus they hedged their affirmation as in:

The SIC method was found to be the most appropriate information criterion for identifying the best model among six alternative models considered. (Cortes, 2004, p. 410).

3.5.2 Discourse organisers

Discourse organizing bundles “reflect relationships between prior and coming discourse” (Biber et al., 2004a), categorized as topic introduction/focus bundles, and
topic elaboration/clarification bundles. In classroom teaching, instructors use topic introduction/focus bundles to introduce new topics in classroom, as in:

But, before I do that, I want to talk about Plato. (classroom teaching, Biber et al., 2004a, p. 392)

The fact that topic introduction/focus bundles often occur with we, the first person collective pronoun or start with you, the second person pronoun, shows that teachers use those bundles to encourage students' participation, as in:

Today we are going to talk about testing hypothesis. (classroom teaching, Biber et al., 2004a, p. 392)

If you look at development and the jobs that are created, it says nothing first of all of the type of jobs that are created. (classroom teaching, Biber et al., 2004a, p. 392)

Speakers/writers use topic elaboration/clarification bundles to further explain topics. It is common for textbook writers to use lexical bundles such as as well as the, and on the other hand, to elaborate their topics through comparison and contrast, as in:

Section 3.5 [...] illustrates how the techniques are employed together as well as the range of resulting execution characteristics that are presented to an architecture, ... (textbook, Biber et al., 2004a, p. 393)

3.5.3 Referential bundles

Referential bundles occur more widely in textbooks and academic prose (Biber et al., 2004a; Biber & Barbieri, 2007; Cortes, 2004). The pragmatic functions of referential bundles are further divided into five categories: identification/focus referential bundles, imprecision bundles, attribute specifying bundles, time/place/text referential bundles and multi-functional referential bundles (Biber et al., 2004a). Identification/focus bundles are used to focus readers'/listeners' attention on the noun phrase after the bundle, as in:

for those of you who came late I have the, uh, the quiz. (classroom
When researchers analyze the discourse functions of lexical bundles, they may take different perspectives, either at the sentential level or at the broader discourse level. Biber et al. (2004a) took the sentential level view and categorized *one of the things, that's one of the* as identification/focus referential bundles. However, they admitted those two bundles functioned in the original text as discourse organizers since the first bundle introduced a topic, and the second one summarized the main point (for more details, see Biber et al., 2004a, p. 277). That explains why Biber & Barbieri (2007) later recategorized those two bundles as discourse organizers.

Imprecision bundles indicate vagueness or approximation of reference in classroom teaching or conversation. *And things like that* is an example of imprecision bundle. Attribute specifying bundles can specify the quantity of the following noun phrase (i.e., *a lot of the, than or equal to*) or frame the tangible/intangible properties of the noun phrases (i.e., *the size of the, the nature of the*). Interestingly, Biber et al. (2004a) mentioned that *a little bit about* not only specifies the quantity of the following noun, but also carries out a hedging function as a downtoner to “minimize the expectations required from students” (p. 395) as in:

> So I want to talk a little bit about process control from that point of view. (classroom teaching) (quoted from Biber et al., 2004, p. 395)

More interestingly, the same phenomenon was noticed by O’Keefe et al. (2007), who found that *a bit of a* is pragmatically specialised as a downtoner and is often associated with negative situations (e.g., *It was a bit of a mess/problem/hassle*).

Time/place/text referential bundles are used to indicate specific places, times or parts of the text itself. Multi-functional bundles (i.e., *the end of the*) may have time/place/text
referential functions at the same time.

3.6 Discussion of lexical bundles

This section concerns two relevant issues of lexical bundles: one is about the formulaicity status of lexical bundles, and the other is about the pragmatic functions of lexical bundles in written academic discourse. With regard to the first issue, although lexical bundles are distinguished by their high frequencies in natural discourse, researchers are cautious about their formulaicity, arguing that high statistical frequencies do not necessarily prove the pre-fabricatedness of lexical bundles. Psycholinguists like Schmitt, Grandage, and Adolphs (2004) came up with experiments to test whether lexical bundles are produced as whole in our mental lexicon. However some obvious limitations with the experiment did not allow the three researchers to deny the formulaicity of lexical bundles.

The second issue of lexical bundles links bundles with the functions they accomplish in the context associated with writing purposes, the reader/author relationship and discipline practices. The discussion of this issue may help us to understand that lexical bundles are useful indicators of registers and genres. Understanding the interaction between text and context would help discipline novices to observe how discipline practices/norms are reflected in language use.

3.6.1 Frequency and formulaicity

Wray (2002) took a cautious approach to the relationship between frequency and pre-fabricated language. She agreed that “if a string is required regularly, it is likely to be stored whole for easier access […] but it does not have to be” (p. 31). Thus she suggested that supplementary criteria are needed to distinguish between pre-fabricated and non pre-
fabricated word strings. There seems to be a need to answer the question of whether recurrent clusters identified by corpus analysis are stored as holistic formulaic sequences in the mind. To answer this question, Schmitt et al. (2004) carried out psycholinguistic investigation to determine if corpus-based recurrent word clusters actually have the prefabricated status.

Schmitt et al. (2004) chose 25 recurrent three- and four-word strings from a pool of recurrent word strings already identified by previous researchers, including lexical bundles identified by Biber et al. (1999) and lexical phrases by Nattinger and DeCarrico (1992). The researchers also created a narrative with those 25 embedded word clusters. Next they designed an oral-response dictation task for research participants to orally retrieve those word clusters. Finally the participants' oral responses were recorded and it was determined whether those clusters were reproduced in a holistic way based on the intonation contour of those word clusters. Only the clusters spoken with a normal intonation contour were considered to have pre-fabricated status. The results showed that only eight clusters were pre-fabricated, which suggests that there was no link between frequency of word clusters and their pre-fabricated status.

However, the main limitation of this research is that the researchers mixed the modes in the research design, which has negatively affected the effectiveness of this study. A closer look at the study results shows that the frequent word strings that were identified as not being pre-fabricated, such as aim of this study, as a consequence of, in a variety of, and so on, are closely associated with written academic prose. When these word clusters were taken out of the formal, academic registers and used in a story created by a group of researchers, participants reported that they had difficulty reconciling their use “with the
informal tone of the narrative” (Schmitt et al., 2004, p. 143). Consequently, it was not surprising to hear false starts or hesitations when participants attempted to retrieve these clusters from memory. Since the researchers used a normal intonation contour as the only criterion for formulaicity, the pauses and hesitations had excluded the written-language-based clusters from possibly being qualified as pre-fabricated, formulaic sequences.

Biber et al. (2004a) chose frequency as the main criteria for the identification of lexical bundles because their research was based on the theoretical assumption that “high frequency is a reflection of pre-fabricated or formulaic status” (Biber & Barbireri, 2007, p. 265), which has been supported by the analysis of pragmatic specialization of lexical bundles. Their research shows that lexical bundles have been consistently used to fulfill particular referencing, discourse organizing and stance expressing functions in real language across registers (Biber et al., 2003; Biber et al., 2004a; Biber & Barbireri, 2007; Cortes, 2004). This consistency association between lexical bundles and their pragmatic functions may suggest that “they are stored whole in memory and used as unanalyzed language chunks for textual and interpersonal discourse functions” (Biber & Barbireri, 2007, p. 283). However, future empirical research is needed to determine if frequency-featured lexical bundles are actually comprehended and produced as holistic chunks.

3.6.2 Written text, context, and lexical bundles

Halliday & Hasan (1989) view written texts as both product and process, recognizing that on the one hand text is an output and on the other hand “it is a process in the sense of a continuous process of semantic choice, a moment through the network of meaning potential, with each set of choice constituting the environment of a further set” (p. 10). In other words, text is a final product of the writing process during which writers
consider the context of situation in terms of the field, the tenor, and the mode of the
discourse, making appropriate language choices to align the text with the situation. The
field, the tenor, and the mode respectively refer to what is happening, who is doing what
to whom, and what rhetorical functions (persuasiveness) or forms of channel that the
discourse is taking. As Hyland (1999) explained, "Writers must anticipate the audience’s
likely background knowledge, processing problems and reactions to the text, with the
understanding that readers are likely to examine it for relevance, informativity, and
interest" (p. 5).

Many researchers (Charles, 2007; Hyland, 1999, 2004; Smyth 2001) have suggested
that genre and discipline variations have direct impact on the shape of text. Text, at the
same time, reveals traces of the contextual features of a variety of disciplines and genres
to researchers who want to learn more about the performance of communication through
textual analysis. Hyland (1999) compared metadiscourse use in textbooks and research
articles across three disciplines: biology, applied linguistics, and marketing.

Metadiscourse, known as “discourse about discourse” (Van de Koppel, 1985), signposts
discourse organization, makes manifest what is said and indicates what is meant. The
audience of textbooks are neophytes of the discipline, having little background
knowledge and discipline-specific literacy skills. Consequently, textbook authors are
expected to state explicitly discipline propositions and facts. Textbooks in all three
disciplines showed a high use of textual forms, especially logical connectives and code
glosses (in other words, namely). For instance, code glosses in the original text added
more information or provide new definitions and were primarily used to “reduce the
cognitive load of propositional material for novices and present unfamiliar content more

In contrast, the audience for research papers are peer experts familiar with the concepts and practices of the discipline. Correspondingly, research articles use fewer textual forms but far more interpersonal expressions to boost the persuasiveness of the writing. These include hedges (*might*, *perhaps*), emphatics (*in fact, it is clear*), and personal markers (*I/we/my*) (Hyland, 1999). The usage of interpersonal forms in research articles serves writers' needs to hold a particular position in a credible and persuasive way, which means that they are concerned with qualifying their arguments, persuading their expert peers of a new interpretation, as well as predicting the possibility of their arguments being proved unreliable.

The frequent use of relational and personal markers in research articles reflects the egalitarian reader/author relationship where writers treat readers as equals and invite them in the discussion. Such equal relationship is not a concern for textbook writers, evidenced by the lack of personal markers. Rather, textbook writers take on the role of knowledge experts in this unequal author/reader relationship, which permits them “greater freedom in expressing their opinions towards propositional content” (Hyland, 1999, p. 20).

Hyland's analysis of metadiscourse use in research articles and textbooks associated specific language units with contextual situations where the text is disseminated to readers. The analysis also indicated that contextual factors in research articles and textbooks are different. Besides genre difference demonstrated by metadiscourse devices, disciplines in the tertiary setting have their own communicative practices and norms; writers in the discipline are expected to establish appropriate reader/author relationships. Smyth (2001) mentioned that introductory psychology textbooks present not only
autonomous facts marking the reality of science, but also contain the hedges which soften the tones of the statements or refer back to previous research. The hedges are reflected in the use of modality expressions (we can ...) and the past tense mainly to refer to previous research. In contrast, introductory biology textbooks "collect what is taken to be accepted within disciplines for passing on to new members, deleting the process of discovery and taking part in the ongoing construction of their sciences" (Smyth, 2001, p. 389). Hedges are rarely seen in biology textbooks since the statements are made with assertion and there is no need to refer back to previous studies for the formation process of the statements.

Charles (2007) examined how the pattern of noun + that constructs stance in two corpora of written texts from political science and materials science. She discovered that nouns in the noun+that pattern mainly associate with semantic meanings of ARGUMENT (i.e., argument, contention) while IDEA (idea, assumption, belief) has higher occurrences in politics texts than in materials science, reflecting the nature of political science as a discipline "that builds knowledge through the examination of ideas and the constructions of arguments" (p. 208). At the same time, nouns in the noun+that pattern in written texts of materials science tend to express meanings of Evidence (evidence, indication), reflecting the norm of hard science where knowledge is constructed on the basis of using experimental results as evidence to support or invalidate the hypothesis under investigation.

The investigation of lexical bundles and their pragmatic functions in writings of a variety of disciplines further raises our awareness of language variations across disciplines. Hylands (2008) investigated lexical bundles in research articles, PhD
dissertations, and M.A. theses from four disciplines: biology, electrical engineering, applied linguistics, and business studies. Half of the lexical bundles identified in one discipline do not occur in any of the other disciplines. Only a small collection of bundles (on the other hand, as well as the, in the case of, at the same time) occur in all four disciplines.

The studies by Biber et al. (2004a) and Hyland (2008) indicate that writers in different disciplines draw on particular lexical bundles to describe propositional facts, organize the flow of discourse, express their evaluation, and engage readers. Referential bundles defined by the functional taxonomy (Biber et al., 2004a) actually correspond to research-oriented bundles defined by Hyland (2008), which “help writers to construct their activities and experiences of the real world” (p. 13). Cortes (2004) found that there are a large number of referential bundles in the 1-million-word corpus composed of biology journal articles. Those referential bundles specify information of time/location (at the beginning of, at the end of, at the University of), describe certain attributes (the depth of the, the length of), or state quantities (a large number of, a measure of). Hyland (2008) found that in science/technology corpora, half of the bundles are research-oriented, indicating quantity, time/location and specific attributes. The use of those bundles meets the writers' need to “demonstrate their ability to handle research methods appropriately and stake a claim to being comfortable with the subject content of the discipline” (Hyland, 2008, p. 15). Those bundles function to explain clearly the research practices, methods, procedures (equipment and materials), and results to the audience, for the purpose of validating hypotheses. The large amount of bundles closely associated with research procedures indicates that the hard sciences are based on grounded, experimental research
Discourse organizers in the functional taxonomy (Biber et al., 2004a) are the counterparts of text-oriented bundles, which are “concerned with the organisation of the text and its meaning as a message or argument” (Hyland, 2008, p. 13). Two thirds of the bundles in applied linguistics and business studies are text-oriented (in the case of, with respect to the, on the basis of), reflecting the interpretative and persuasive nature of soft knowledge fields “which often recast knowledge as sympathetic understanding, promoting tolerance in readers through an ethical rather than cognitive progression” (Hyland, 2008, p. 16). Seventy-five percent of the text-oriented bundles in the social science texts are framing signals, specifying limiting conditions for argument by indicating connections among arguments, directing readers’ attention to aspects of a statement, or clarifying those aspects.

Stance bundles, as defined by Biber et al. (2004a), are equivalent participant-oriented bundles (Hyland, 2008). According to Biber et al.’s (2004a) taxonomy, stance bundles were further categorized into two main subcategories: epistemic stance bundles expressing assessments of following propositions and attitudinal/modality stance bundles where writers/speakers indicate their attitudes or direct readers/listeners to do. In order to emphasize the author/reader relationship, Hyland regrouped stance bundles into two main subcategories, engagement bundles which “address readers directly” (it should be noted, as can be seen) and bundles of stance features which “convey the writer’s attitudes and evaluations” (are likely to be, it is possible that). Hyland’s research found that research articles from the hard sciences (biology and electrical engineering) employ more engagement bundles (we can see that, it is important) than those articles of the soft
sciences (applied linguistics and business studies). Since hard sciences have "the more linear and problem-oriented approach to knowledge construction" (Hyland, 2008, p. 19), writers use directives to help readers get direct access to lines of arguments. Compared with the hard sciences, the soft sciences (applied linguistics and business studies) use more stance bundles (*may be due to, are more likely to*), which allow writers to hedge their propositions and express their opinions instead of accepted facts.
Chapter 4 Methodology

Before moving into a detailed explanation of my research methodology, I first review my research questions: 1) what are the lexical bundles and their pragmatic functions in introductory textbooks of Electrical Engineering (EE)? 2) Do those lexical bundles occur in ESP textbooks? Do the pragmatic functions of those bundles differ? 3) Do ESP textbooks have a focus on lexical bundles and their pragmatic functions in learning activities?

To answer those questions, I used a corpus-based approach to examine potential gaps of language use between Electrical Engineering introductory textbooks and ESP textbooks. In this section, I first explain how I selected text materials for the two corpora, the Electrical Engineering Introductory Textbook Corpus (EEITC) and the English for Specific Purposes Textbook Corpus (ESPTC). I then describe how I set up the two corpora.

I also mention the frequency- and distribution- criteria for identification of 4-word lexical bundles in the EEITC. In addition, I explain how WordSmith Tools 4.0 has been used to identify lexical bundles and extract concordance lines or bigger stretches of text for further analysis. After that, I talk about using the functional taxonomy by Biber et al. (2004a) as the main analytical framework for analyzing the pragmatic functions of lexical bundles. In my research, I also checked the presence of lexical bundles in the ESPTC, as well as the treatment of those bundles in the ESP textbooks.

4.1 Setting up the EEITC

To set up the EEITC, I first chose representative introductory textbooks since the corpus is expected to reflect language use in EE introductory textbooks. Secondly, I
transferred the selected paper-based text into machine-readable text for the final set-up of the EEITC.

4.1.1 Selection of representative introductory textbooks of Electrical Engineering

As I discussed in Chapter 2, a corpus is a collection of electronic texts representing specific or general language use. To develop a corpus, the first step is to collect text materials in line with individual research goals. Since my research project concerns language use in introductory Electrical Engineering textbooks, I went to the websites of engineering faculties in three Ontario Universities, the University of Toronto, the University of Waterloo, and Carleton University. The three universities have Electrical Engineering undergraduate programs with good academic reputations.

First I checked the on-line curricula of the three EE undergraduate programs. In all three, the first two years of the programs offer students a foundation in science and mathematics, as well as important concepts in Electrical Engineering. In the two remaining years, students choose specialization courses, depending on their individual interests and needs. Those subject-relevant courses cover microelectronics, digital communications, wireless systems, signal processing, and so on.

The introductory courses cover material in mathematics, material, electrical fundamentals, and computer fundamentals. Courses in electrical fundamentals are considered as core introductory courses because they teach students basic concepts in circuit analysis needed by all Electrical Engineering students regardless of their future specialization. So I collected textbooks used in courses of electrical fundamentals as the text sources for the EEITC.
Table 4.1 Textbooks used in introductory courses of Electrical Engineering at three Ontario universities

<table>
<thead>
<tr>
<th>The University of Toronto:</th>
<th>The University of Waterloo:</th>
<th>Carleton University:</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Electrical Fundamentals</em> (ECE110H1)</td>
<td><em>Fundamentals of Electrical Engineering</em> (ECE100)</td>
<td><em>Circuits and Signals</em> (ELEC 2501)</td>
</tr>
<tr>
<td><em>Electric circuits</em> (Nilsson &amp; Riedel, 2004)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 lists the introductory courses at the three universities and the textbooks required in those courses. At the University of Toronto, the University of Waterloo, and Carleton University, the introductory courses respectively are *Electrical fundamentals* (ECE110H), *Fundamentals of electrical engineering* (ECE100), and *Circuits and signals* (ELEC 2501) (see Appendix A). As we see in Table 4.1, there is some overlap in the textbooks required in those courses. *Basic engineering circuit analysis* (Irwin, 2002) is the textbook for introductory courses at the University of Toronto and Carleton University. This textbook covers basic concepts in circuit analysis, including electrical elements, resistive circuits (where Ohm's and Kirchhoff's Laws are introduced), nodal and loop analysis (where the basic circuit analysis techniques are explained), equivalent circuits (superposition, Thévenin and Norton's Theorems), capacitance and inductance, and so on.

In addition to *Basic engineering circuit analysis* (Irwin, 2002), *Fundamentals of physics extended* (Halliday, Resnick, & Walker, 2001) (see Appendix A) is required in introductory courses at the University of Toronto and the University of Waterloo. The course outline of ECE100 at the University of Waterloo specifies that the topics from Chapter 21 to Chapter 28 are covered during the course; topics include electric charge,
electric fields, Gauss' laws, capacitance, magnetic fields, etc.

*Microelectronic circuits* (Sedra & Smith, 2004) is required in the introductory course ECE100 at Waterloo. The first five chapters of the book introduce basic concepts in microelectronic design, covering topics of devices and basic microelectronic circuits, operational amplifiers, diodes, MOS field-effect transistors (MOSFETs), and bipolar junction transistors (BJTs). *Electric circuits* (Nilsson & Riedel, 2004) (see Appendix A) is recommended by the introductory course ECE1101H at the University of Toronto.

I initially selected the four textbooks mentioned above as possible source texts for the EEITC. Then it became necessary to have course instructors look at those textbooks for an accurate evaluation of their representativeness as introductory course books. In November 2007, my supervisor and I asked a professor, who is currently teaching introductory courses in Department of Electronics at Carleton University, to evaluate the textbooks. Before the interview, I designed an evaluation questionnaire (see Appendix A).

In addition to the questionnaire, we took the four textbooks listed in the questionnaire to the interview. With the questionnaire functioning as the basis of the interview, the professor examined the tables of content and chose two books as representative, *Basic engineering circuit analysis* by (Irwin, 2002) and *Microelectronic circuits* (Sedra & Smith, 2004). He thought that Irwin's text was a good introductory course book because it covers most basic concepts in circuit analysis and thus is suitable for novices who have completed prerequisites in university calculus and physics. His comment is consistent with the statement made by the textbook author that the book is “designed as a core text for students majoring in electrical and computer engineering” so that learners can gain “a thorough understanding of the fundamental concepts of circuit analysis and their
applications to real-world problem” (Irwin, 2002, p. i). As for the second book, the professor recommended that the first five chapters fit in a one-term introductory course. Besides, he mentioned that the book has been extremely popular with about two million copies sold.

The professor's evaluation of the four textbooks located two textbook as the text base for the development of the Electrical Engineering Introductory Textbook Corpus (EEITC). Although it is true that the representativeness of the EEITC would be increased with more textbook material incorporated into the corpus, the two textbooks provide a solid base for the EEITC due to their representativeness, which has been confirmed by the course syllabi and the professor's evaluation.

4.1.2 The compilation of the Electrical Engineering Introductory Textbook Corpus (EEITC)

The decision was made to draw text materials from two textbooks, Basic engineering circuit analysis (Irwin, 2002) and Microelectronic circuits (Sedra & Smith, 2004). For the EEITC, the entire textbook by Irwin and the first five chapters of the textbook by Sedra and Smith were compiled into the corpus. The professor who evaluated the texts offered the lab manual for Circuits and Signals (ELEC2501) in the introductory course; this was included in the corpus.

As discussed in Chapter 2, a corpus is a collection of electronic text representative of target language use, available for computer-based qualitative and quantitative analysis. Since the two introductory textbooks are paper-based, it was necessary to transfer the content into electronic-based text that could be analyzed by computer software programs. I scanned the two textbooks and the lab manual into a computer. Next the initial scanned-
in image files were edited as text files (saved in the format of *.txt files) so that they can be recognized by the analysis software program (i.e., WordSmith Tools 4.0). Then I manually cleaned the text and kept only textual information. Figures and complex formulae were eliminated because the original characteristics of visual images and mathematical representations were lost or were unrecognizable as symbols when they were saved as part of the *.txt files, the format required by WordSmith Tools 4.0. Headers were also deleted since they have no connection with the text.

The cleaned text files were compiled into the EEITC, as indicated in Appendix C. The EEITC is made up of 22 texts, covering all 16 chapters of Basic engineering circuit analysis (Irwin, 2002), the lab manual for Circuits and Signals (ELEC2501), and five chapters of Microelectronic circuits (Sedra & Smith, 2004).

4.2 Setting up the comparison corpus: the ESP Textbook Corpus (ESPTC)

The content of the ESPTC, a comparison corpus, is drawn from selected ESP textbooks for Electrical Engineering students, and has been developed for the comparison of language use in ESP textbooks and entry-level course books of Electrical Engineering. Although there are many English for Academic Purposes (EAP) textbooks on the market with the aim of introducing general academic language to university ESL students from a variety of disciplines, there are only a few textbooks designed for Electrical Engineering students. The two ESP textbooks that I found are Oxford English for electronics (Glendining & McEwan, 1993) and Oxford English for electrical and mechanical engineering (Glendining & Glendining, 1995), which are designed for college/university students of Electronics/Electrical engineering. The publishers' websites emphasize that the two textbooks include authentic reading materials from a wide range of sources and
cover a variety of topics. The website for *Oxford English for electronics* (Glendining & McEwan, 1993) says that the book features “authentic reading and listening passages from a wide variety of sources”, and covers topics, such as “electronics in the home, block diagrams and circuits, alarm systems, transistor characteristics, computers, data transmission, and many more” (quoted from http://www.oup.com/elt/catalogue/isbn/0-19-457384-2?cc=gb). *Oxford English for electrical and mechanical engineering* (Glendining & Glendining, 1993) also includes authentic reading materials on a variety of topics, including materials, mechanisms, forces, safety at work, engineering design, automated systems, robotics, hydraulics, process control, and CAD.

To further explore whether those two ESP textbooks actually reflect language use in entry-level Electrical Engineering textbooks, it was necessary to compile the contents of the two ESP books into a specialized corpus, termed the ESP Textbook Corpus (ESPTC). To obtain electronic-based text which could be analyzed by computer software programs, I scanned the two ESP textbooks (listening transcripts were excluded), and transferred the scanned-in image files into text files. Next I cleaned the text files by removing repetitive headlines, pictures, tables, and other non-textual materials. Finally, I got two text files with content respectively drawn from the two textbooks (see Appendix B). The two text files construct the 99,774-running-word ESPTC.

4.3 The identification criteria of lexical bundles

To investigate how lexical bundles function in introductory course books of Electrical Engineering, I first identified lexical bundles in those textbooks. For multi-word clusters to be qualified as lexical bundles, Biber et al. (1999) specified that both frequency- and distribution- criteria have to be satisfied. As indicated earlier the cut-off
frequency for lexical bundles seems to be slightly arbitrary, depending mainly on researchers' evaluation of data manageability. For example, Biber et al. (1999) set the cut-off frequency of four-word lexical bundles at 10 occurrences per million words while Biber et al. (2004a) took a more conservative view by setting the cut-off frequency at 40 times per million. There seem to be strong correlations between the cut-off frequency of lexical bundles and the specialization of the corpus. Compared with the 15-million-word general corpora used in the study by Biber et al. (2004a), the 2-million-word corpus in Cortes' (2004) study is much smaller and specialized with materials drawn from academic writing in history and biology. Surprisingly, the cut-off frequency for four-word lexical bundles was set at the high level of 20 times per million words, indicating that a narrower and more specific scope of target language tends to have more lexical bundles that reflect formulaic language use.

Similar to the corpus used in Cortes' (2004) study in terms of the size and the specificity of language use, the EEITC is a small corpus with 247346 running words. I used the same cut-off frequency adopted in Cortes' (2004) study, a minimum frequency of 20 occurrences per million words for four-word lexical bundles. This normalized frequency can be adjusted to equivalent raw frequency counts for various sized corpora. In my research, the minimum cut-off frequency, 20 times per million words, equals a raw frequency of 8 times for the 247346-word EEITC.

In addition to the frequency-criterion for lexical bundles, the distribution-criterion in the operational definition of lexical bundles (Biber et al., 1999) mandated that a lexical bundle should occur in at least five texts. In my study, I followed the same criterion to identify lexical bundles from the EEITC.
4.4 Using WordSmith Tools 4.0 for identification of lexical bundles and textual analysis

I chose WordSmith Tools 4.0 (Scott, 2007) as the main instrument for identification of lexical bundles and textual analysis. This decision is based on Ari's (2006) study, where he compared the efficiencies of three concordance software programs, which are kfNgram 1.2.03, N-Gram Phrase Extractor, and WordSmithTools 3. Among the three programs, WordSmith Tools 3 has shown the best performance because it has the function of making frequency lists of multi-word clusters, as well as showing concordance lines and multi-text occurrences.

WordSmith Tools 4.0 has three main functions, Concord, Keywords and Wordlist (Scott, 2007). I used two functions, Concord and Wordlist, for the identification of lexical bundles and the provision of concordance lines. To make a frequency list of four-word clusters as shown in Figure 1, the first step was to choose all the 22 text files of the EEITC under the menu of Wordlist, followed by choosing the option of "make a word list". The statistical information of the text files (i.e. running words, types, etc) is available in the pop-up window. A three-word-cluster frequency list can automatically be generated by simply clicking on the concordance option from the drop-down menu of Compute. It is necessary to go to the cluster option under Compute to specify the word cluster length as 4, as well as the minimum cut-off frequency count as 8 times. This process generated a list of four-word clusters with frequency counts higher than 8 in the EEITC (see Figure 4.1).

---

1 More information about the three software programs, kfNgram 1.2.03, Phrase Extractor, and WordSmithTools3, can be respectively obtained from the following websites: http://www.kwicfinderxorn/kfNgram/kfNgramHelp.html, http://www.lectutor.ca, and http://www.lexically.net/wordsmith/index.html.
The sub-menu of Concord in Wordsmith Tools 4.0 can be used to check if the clusters meet the distribution-criterion of spreading over at least five texts. Based on the list of four-word clusters with raw frequency counts higher than eight, I entered individual four-word clusters from the list to obtain the text-distribution information of those clusters. Only those occurring in more than five texts are considered as lexical bundles. Concord can also extract concordance lines of lexical bundles from original text. Take the most frequent four-word cluster the current in the as an example. Clicking on Concord displays a concordance window of this cluster, as shown in Figure 4.2.
I used Biber et al.'s (2004a) functional taxonomy as the analytic framework in my...
research, and examined concordance lines of those lexical bundles for their pragmatic functions in the original text. Biber et al. (2004a) proposed the functional taxonomy of lexical bundles based on their corpus analysis of university discourse across four registers, classroom teaching, textbooks, conversation, and academic prose. The discourse functions of the lexical bundles are categorized into three main groups: referential bundles, stance bundles and discourse organizers. For better data management, I encoded the discourse functions noted in the taxonomy (Biber et al., 2004a) into a range of code numbers (See Appendix C). In my analysis, I depended on the taxonomy to determine the pragmatic functions of lexical bundles by examining concordance lines and sometimes longer stretch of original text in the EEITC. Although a lexical bundle can have multiple functions, only its primary function was categorized for further quantitative analysis according to the functional taxonomy. I consulted my supervisors to confirm the decisions I made about the pragmatic functions of lexical bundles. The structural types of lexical bundles in the EEITC were also analyzed to see if there existed a strong association with language use in introductory textbooks of Electrical Engineering.

Lexical bundles identified in the EEITC were examined for their presence in the pedagogical corpus, the ESPTC. The bundles available in the ESPTC were also scrutinized for their pragmatic functions. I used the statistical computer software SPSS to record, organize, and analyze the original data in the EEITC and ESPTC. The data includes functional characteristics, structural characteristics and raw frequency counts of lexical bundles. Comparison of the results of lexical bundles regarding their structure and functions in the two corpora will be reported in detail in the next section. The bundles available in the ESPTC were also examined to see if any attention has been given to these
bundles by checking the exercise and task sections of the two ESP textbooks, *Oxford English for electronics* (Glendining & McEwan, 1993) and *Oxford English for electrical and mechanical engineering* (Glendining & Glendining, 1995).
Chapter 5 Results and discussion

To answer the central research question of if there is a gap in language use in terms of the lexical bundles that exist in introductory Electrical Engineering textbooks and ESP textbooks, I have organized the research findings and discussion into eight sections. In the first part, I review the theoretical background of how particular linguistic features help construct text. Next I report on the structural characteristics of lexical bundles in the EEITC and the strong associations between structural types and pragmatic functions. In the third part, I focus on lexical bundles in the EEITC (the target corpus) and their pragmatic functions in the original context. I then identify the bundles present in the ESPTC and analyze their pragmatic functions in the ESP textbooks. In the fifth part, I compare the proportional distribution of lexical bundles in the two corpora, the EEITC and the ESPTC, and examine how lexical bundles construct texts differently in the two corpora. In the sixth part, I report my investigation results on how the two ESP textbooks, *Oxford English for electronics* (Glendining & McEwan, 1993) and *Oxford English for electrical and mechanical engineering* (Glendining & Glendining, 1993) treat lexical bundles in learning tasks. In the seventh section, I critique Biber et al.’s (2004a) functional taxonomy and suggest that language teachers have a cautious attitude toward the way pragmatic functions of lexical bundles are categorized by the taxonomy. Finally, I summarize the findings from the previous seven parts to provide an overview of the results of this corpus-based research.

5.1 What does text do in discipline textbooks?

Textbooks present “the authorized version of a society’s valid knowledge” (Olson, 1989, p. 238). The authority of textbooks means that they “[...] play a central role in the
learners' experience and understanding of a subject by providing a coherently ordered epistemological map of the disciplinary landscape and, through their textual practices, can help convey the norms, values and ideological assumptions of a particular academic culture" (Hyland, 1999, p. 3). Instructors depend on textbooks to organize their teaching and transmit basic and core concepts to students. Students spend their time and energy absorbing knowledge from textbooks.

The fact that most readers of introductory textbooks are disciplinary novices determines that the reader/writer relationship is a hierarchical one that empowers the writer to take "an expert role towards both information and readers" (Hyland, 1999, p. 21) and positions readers as learners of established knowledge in the field. Textbook writers present the basic concepts, conventions, and core skills of the discipline to readers. Textbooks are supposed to "say what they mean and mean precisely, neither more nor less than, what they say" (Olson, 1989, p. 237). At the same time, writers need to explicitly guide the neophytes through the content, initiate them into the discipline, and engage them.

To initiate readers into the discipline, textbook authors present propositional facts, norms and conventions for the purpose of making meaning explicit and accessible to novices (Hyland, 1999). They use particular linguistic forms that emphasize term definitions and explicit logical relationships among clauses. Hyland found that introductory textbook writers of biology, marketing, and applied linguistics more often used logical connectives (in addition) and code glosses (in other words), two types of textual metadiscourse devices, for the purposes of organizing discourse and clarifying propositional facts. Biber et al. (2004a) discovered that university textbook writers
frequently use referential bundles to identify important entities (one of the most) or to specify important attributes of an entity (the sizes of the, in the case of). These referential bundles serve to construct detailed, accurate propositional facts in academic texts. The prevalent use of logical connectives, code glosses, and referential bundles in introductory textbooks demonstrates textbook writers’ concern for presenting information clearly and explicitly. Besides presenting propositional facts, textbook writers make efforts to engage readers in the content. Hyland (2002b) discovered that textbooks, compared with research articles and student writings, more often use collective first person pronouns we and us to invite readers to participate in actions that the writers have emphasized.

The authority of textbooks is realized when textbook authors help readers to construct their role as learners in the learning process. Authors need to present the ‘true’ and ‘valid’ discipline knowledge keeping readers away from arguments and dialogues among experts. Hyland (2002a) demonstrated how Science and Engineering textbook authors use directives to guide neophytes through the text, putting an obligation on readers to learn what is important and directing their actions mainly on the cognitive level. Utterances expressing a directive force are realized by (1) imperatives (let’s consider), (2) modals of obligation (should), or (3) predicatives followed by a complement to clause (it is important to). “While the devices may convey different degrees of emphasis, they all carry the authority to the writer in specifying how the reader should participate in the text or perform some action outside it” (Hyland, 2002a, p. 217). Biber et al. (2004a) found that the fact that the was frequently used in textbooks to introduce outside support and boost the objectivity of the following proposition. Hyland (1999) noted that science textbook writers used many hedges featuring modality indicators (i.e., possible, may,
The hedges function like buffer areas so that authors do not need to constantly upgrade their claims. The hedges as well protect the text against refutation from possible expert readers, and furthermore, they provide novices with models which indicate cautious attitude toward facts in a discipline’s rhetorical conventions.

Introductory textbooks are an important type of primary text in academic settings because they offer the content on which professors’ lectures and evaluated tasks (exams, graded assignments, oral and written reports) are based (Carson, 2001). In addition, previous studies have shown that writers in the discipline have preference for certain sets of linguistic features and use these devices to construct the body of knowledge (Biber et al., 2004a; Biber & Barbieri, 2007; Cortes, 2004).

The introductory textbooks might be peripheral to third- and fourth-year undergraduates and graduate students, who are more concerned with establishing their positions in research reports or papers and thus need to acquire rhetorical features of published research articles (Hyland, 1999). However, for discipline novices, understanding how textbook authors construct disciplinary texts through particular linguistic features, as well as distinguishing what is said and what is meant in textbooks, will help neophytes understand how authors generate meanings and locate key knowledge points.

Consequently, there is a need to analyze introductory textbooks from specific disciplines to understand how discipline writers use particular linguistic features to construct the text in order to initiate and engage novices into the discipline. So far, there have been few studies that investigate the use of lexical bundles in Electrical Engineering introductory textbooks.
My investigation of lexical bundles in the EEITC can begin to bridge this research gap. Based on the frequency- and distribution- criteria specified in the methodology section (Chapter 4), 105 lexical bundles in the EEITC were identified and categorized according to their grammatical structures (see Table 5.2) and discourse functions (See Table 5.3). Later in this section, I make reference to these tables for a detailed discussion of the grammatical and functional characteristics of lexical bundles. First I look at the structural type distribution of lexical bundles in the EEITC as listed in Table 5.1.

5.2 Structural characteristics of lexical bundles in the EEITC

Introductory textbook writers use more phrasal bundles than clausal bundles to introduce the propositional facts and discipline norms. As shown in Table 5.1, phrasal lexical bundles (e.g., in the case of, the sum of the) predominate, with 66.67%, more than twice the percentage of clausal lexical bundles (e.g., is connected to the, is equal to the). Consistent with the observation that “the lexical bundles in academic prose are phrasal rather than clausal” (Biber et al., 2004a, p. 382), introductory Electrical Engineering textbook authors frequently use phrasal lexical bundles to explicitly describe transactional facts to novices.

Reppen (2004) and Conrad (1996) grouped concurrent linguistic features that reflect same functions along five dimensions, Dimension 1 (involved vs. informational production), Dimension 2 (narrative vs. non narrative discourse), Dimension 3 (situation dependent vs. elaborated reference), Dimension 4 (other expressions of persuasion), and Dimension 5 (impersonal vs. non-impersonal style). Along Dimension 1, they found that textbook authors use nouns and prepositions for the purpose of transmitting information, suggesting that presenting factual information is a key feature of textbooks. Not
surprisingly, 64% of lexical bundles in the EEITC have one of the two structural patterns: noun phrase with embedded fragment (e.g., *the remainder of the, the magnitude of the*) and prepositional phrase with embedded fragment (e.g., *in the design of, from the fact that*).

**Table 5.1** Structural types and proportional distribution of lexical bundle in the EEITC

<table>
<thead>
<tr>
<th>Basic structural type</th>
<th>Structural categories</th>
<th>Percentage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical bundles with phrasal structure</td>
<td>Noun Phrase with of-phrase fragment</td>
<td>30.48%</td>
<td>66.67%</td>
</tr>
<tr>
<td></td>
<td>Noun Phrase with other post-modifier fragment</td>
<td>13.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prepositional phrase with of-phrase fragment</td>
<td>6.67%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other prepositional phrase</td>
<td>13.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verb phrase with active verb</td>
<td>2.86%</td>
<td></td>
</tr>
<tr>
<td>Lexical bundles with clausal structure</td>
<td>Anticipatory it + Verb/Adjective phrase</td>
<td>0.95%</td>
<td>31.43%</td>
</tr>
<tr>
<td></td>
<td>Passive verb + prepositional phrase fragment</td>
<td>12.38%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coupula Be+Noun/adjective phrase</td>
<td>7.62%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Verb/Adjective+) to-clause fragment</td>
<td>2.86%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adverbiacl Clause fragment</td>
<td>0.95%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>First person plural pronoun + verb phrase</td>
<td>6.67%</td>
<td></td>
</tr>
<tr>
<td>Other expressions</td>
<td></td>
<td>1.9%</td>
<td></td>
</tr>
</tbody>
</table>

In Table 5.2, 105 lexical bundles are categorized into 12 structural types. These lexical bundles are not complete grammatical units but rather phrasal or clausal fragments with other fragments embedded.

**Table 5.2** Structural classification of lexical bundles in the EEITC

1. **Noun phrase with of-phrase fragment** (N=30)
   The sum of the, the remainder of the, the magnitude of the, the value of the, the gain of the, the operation of the, the ration of the, the slope of the, the frequency response of, input resistance of the, the output of the, the input resistance of, the values of the, gain of the amplifier, the polarity of the, the voltage gain of, the direction of the, voltage gain of the, the amplitude of the, a plot of the, a sketch of the, value of the current, the
location of the, the resistance of the, a function of the, the details of the, a result of the, the effect of the, the use of the, the case of the, the presence of the, the end of the

2. **Noun phrase with other post-modifier fragment (N=13)**
   the fact that the, the voltage across the, the current in the, the voltage at the, the current through the, voltage across the capacitor, equivalent circuit for the, the difference between the, the relationship between the, energy stored in the, an expression for the, the energy stored in, the expression for the, power delivered to the

3. **Prepositional phrase with embedded *of-phrase* fragment (N=7)**
   in the case of, in the design of, on the value of, of the input signal, in terms of the, in the range of, in the circuit of

4. **Other prepositional phrase (N=14)**
   from the fact that, on the other hand, of the circuit in, for the network in, with the result that, with respect to the, in series with the, in the frequency domain, in the time domain, in parallel with the, in the circuit in, in the network in, as well as the

5. **Verb phrase with active verb (N=3)**
   Consider the circuit shown, let us determine the, let us find the

6. **Passive verb + prepositional phrase fragment (N=13)**
   can be used to, can be obtained by, can be written as, can be found from, be found from the, is known as the, referred to as the, is defined as the, is connected to the, delivered to the load, be connected to the, given by the expression, is given by the

7. **Copula Be+Noun/adjective phrase (N=8)**
   is equal to the, is proportional to the, is the sum of, is directly proportional to, is the magnitude of, is determined by the, is independent of the, is of the form

8. **Personal pronoun + lexical/verb phrase (N=7)**
   we note that the, we see that the, we find that the, we assume that the, we wish to determine, we wish to find, we will use the

9. **(Verb/Adjective+) to-clause fragment (N=3)**
   To note that the, wish to determine the, wish to find the

10. **Adverbial Clause fragment (N=1)**
    As long as the

11. **Other expressions (N=2)**
    As well as the, present in the network

12. **Anticipatory it + Verb/Adjective phrase (1)**
    It follows that

5.2.1 Phrasal lexical bundles

Phrasal lexical bundles consist of five structural subcategories (see Table 5.2), noun phrase with *of-phrase* fragment, noun phrase with other post-modifier fragment, prepositional phrase with *of-phrase* fragment, other prepositional phrase, and verb phrase with active verb. The two biggest groups are noun phrase with embedded fragments and prepositional phrase with embedded fragments.
In the first group of lexical bundles, the majority have an *of*-phrase as head of the post-modifying component, as in the *sum of the*, the *remainder of the*, the *magnitude of the*, etc.. The rest of the bundles take the form of a noun phrase with other post-modifier fragment, like the *fact that the*, the *voltage across the*, the *current in the*, and so on.

The noun phrases in many of these bundles are discipline specific, such as *gain*, *frequency response*, *input resistance*, *output*, *current*, *voltage*, etc.. The high frequency of those discipline-specific bundles in introductory textbooks not only indicates their importance as basic concepts, but also their necessity in describing processes in the textbooks. For instance, *the current in the* occurs 31 times in the EEITC and is an essential part in explaining a particular process, as in:

This will obviously be satisfied if *the current in the* divider is made much larger than the base current. (MEchapter05)

Lexical bundles with *of*-phrase fragment, like *the slope of the*, *the direction of the*, *the ration of the*, assist with the description of characteristic curves of circuit parameters and the composition and operation of the circuits. For instance, *the slope of the* is used in the original text to describe the frequency-magnitude relationship in the composite plot, as in:

At *w = 10 rad/s*, which is the break frequency of the pole, *the slope of the* magnitude curve changes back to - 40 dB/decade. (CAchapter11)

The lexical bundles headed by an *of*-phrase also describe the physical/abstract attribute of the following noun, including quantity, physical entity, abstract quality, etc., as we can see in the following examples:

1. which states that *the sum of the* currents entering a node is equal to *the sum of the* currents leaving the node. (CAchapter2)
2. Since *the gain of the* op-amp configuration is 1, vo=vs. (CAchapter3)
3. Thus, the bridge can be roughly modeled as a second order system. let us design an RLC resonance network to demonstrate the bridge's vertical movement and investigate *the effect of the* wind's frequency. (CAchapter11)
The second biggest group of phrasal lexical bundles have the structure of prepositional phrases with embedded *of-phrase* or other fragments, as shown in Table 5.2. Bundles starting with *in* often mark physical existence of location and connection, examples of which include *in the design of, in the frequency domain, in parallel with the,* etc. In the first five concordance lines below, for example, *in parallel with the* in line 4 specifies the physical connection between electrical components, such as resistors and current sources, as in:

1. current source is 40 KΩ, i.e., 60 kΩ, *in parallel with the* series combination
2. 12-kΩ, resistor. This resistor is *in parallel with the* 4-KΩ resistor, and
3. are in series and their combination *in parallel with the* 6-kΩ resistor.
4. with a current source of value \( i = \frac{v}{R} \) *in parallel with the* resistor R.
5. is convened to a 4-mA current source *in parallel with the* 3-KΩ resistor. If

In addition, many lexical bundles, such as *in the case of, on the other hand, and with the result that,* explicitly mark a diversity of logical relationship of elucidation, contrast/comparison, or cause/effect, as in:

1. We found that capacitors and inductors were capable of storing electric energy. *In the case of a* charged capacitor, the energy is stored in the electric field that exists between the positively and negatively charged plates. (CAchapter06)
2. Digital applications, *on the other hand,* make use of the gross non-linearity exhibited by the VTC. (MEchapter01)
3. In actual practice, it is usual that one side of the junction is much more heavily doped than the other, *with the result that* the depletion region exists almost entirely on one side (the lightly doped side). (MEchapter03)

### 5.2.2 Clausal lexical bundles

Accounting for 31 percent of all lexical bundles in the EEITC, clausal lexical bundles are mainly present in three structural types: passive verb + prepositional phrase fragment (e.g., *can be obtained by*), copula *be* + noun/adjective phrase (e.g., *is proportional to the*), and personal pronoun + lexical/verb phrase (e.g., *we see that the*) (see Table 5.2).
The first structural type of lexical bundles, with a passive verb followed by a prepositional phrase fragment, indicates definitions, scientific formulas, locations, or possibilities. Bundles such as is known as the, referred to as the, and is defined as the introduce new terms and concepts in original texts, as in:

1. Such a characterization of amplifier performance is known as the amplifier frequency response. (MEchapter1)
2. The angle $\theta$ is the phase angle of the load impedance and is often referred to as the power factor angle. (CAchapter09)
3. The power factor is defined as the ratio of the average power to the apparent power and is said to be leading when the phase of the current leads the voltage. (CAchapter09)

Bundles like can be written as and given by the expression direct readers’ attention to the following scientific formulas as in:

If the power dissipated in the amplifier circuit is denoted $P_{\text{dissipated}}$, the power-balance equation for the amplifier can be written as $P_{\text{dc}}+P_{1}=P_{L}+P_{\text{dissipated}}$ where $P_{1}$ is the power drawn from.... (MEchapter1)

It is as well noticeable that bundles of this structural type describe physical connections among electrical components, such as is connected to the. Other bundles of this structural type, such as can be used to, can be found from, and can be obtained by, indicate the authors’ cautious attitude toward the possibility of the proposition, as in:

The manner in which the current $i(t)$ from the source divides between the two branches is called current division and can be found from the preceding expressions. (CAchapter02)

Eight lexical bundles fall into the second structural type, starting with copula be, followed by subject predicative, being either noun or adjective phrase (See Table 5.2). Four bundles in this category, is equal to the, is proportional to the, is the sum of, and is directly proportional to, describe quantitative relationship, as in:

1. The right-hand side of the equation is the sum of the currents entering node I through current sources. (CAchapter03)
2. This name is derived from the fact that if the voltage or current is 1/2 of its maximum value, then the power, which is proportional to the square of the voltage or current, is one-half its maximum value. (CAchapter11)
The third type of clausal lexical bundles (e.g., we note that the, we see that the), which takes the form of personal pronoun + lexical/verb phrase, start with the collective pronoun we. They in most cases serve to show writer's desire of inviting reader's attention to what is stated in the complement clause, as in:

In addition, we note that the solution of sinusoidal steady-state circuits would be relatively simple if we could write the phasor equation directly from the circuit description. (CAchapter07)

5.2.3 Discussion of structural characteristics of lexical bundles

The richness of lexical bundles in the EEITC and the dominance of phrasal lexical bundles over clausal lexical bundles distinguish introductory textbooks of Electrical Engineering from general university textbooks. Biber and his fellow researchers observed that university textbooks are short on lexical bundles since only a total of 32 lexical bundles was identified in the textbook subcorpus (760,000 words) of the T2K-SWAL corpus (Biber et al., 2004a). They attributed the lack of lexical bundles to the preference textbook authors have for full clauses over phrasal lexical bundles and posited that the authors tend to have more language choices without the need to depend on repertoires commonly seen in real-time production under time pressure. However, the findings of my research seem to contradict this conclusion. In the much smaller EEITC (247,346 running words), 67% of the 105 lexical bundles are phrasal bundles.

The high percentage of phrasal lexical bundles suggests that Electrical Engineering introductory textbook authors heavily rely on phrasal lexical bundles as building blocks for meaning generation, framing new information to readers and allowing meaning construction in the text. Fragmental structure of lexical bundles allows for integration of other structures, such as noun phrases, that complement clauses and other clauses to generate new ideas. Lexical bundles having the structure of noun phrase and fragments
(the magnitude of the, the value of the) connect following noun phrase with the embedded one and specify particular attributes of the following noun phrase. Lexical bundles with the structure of prepositional phrase and fragments specify location information (in the frequency domain) or allow complement clauses or noun phrases that provide extra descriptive information to the embedded noun phrase (from the fact that, in the case of).

Why are there fewer phrasal bundles in the textbook subcorpus in the study by Biber et al. (2004a)? The lack of specialization in the target language can explain why more four-word lexical bundles were identified in the EEITC than in the larger textbook subcorpus of the T2K-SWAL corpus. The textbook subcorpus in Biber et al.'s study is a collection of 87 texts from six major academic disciplines (Business, Engineering, Education, Humanities, Natural Sciences and Social Science) across three academic levels (lower division undergraduate, higher division undergraduate, and graduate). Previous research by Hyland (2008) proved that writers have discipline-specific repertories of lexical bundles. This discipline specificity results in a limited overlapping of lexical bundles in academic writings across four disciplines (Biology, Electrical Engineering, Applied Linguistics, and Business studies). Among the 50 most frequent lexical bundles in the four disciplines, only 5 bundles (on the other hand, as well as the, in the case of, at the same time, and the results of the) occur across all the disciplines.

The 247346 running-word EEITCC represents more specific language use, including only introductory textbooks of Electrical Engineering. Although the size of the textbook subcorpus (760,000 words) of the T2K-SWAL is more than three times larger than the EEITC, it is less specific. Correspondingly there are fewer lexical bundles identified in the T2K-SWAL.
The findings as well exhibit strong associations between structural types of lexical bundles and pragmatic functions of those bundles. These bundles have been used to transmit both transactional and relational information to textbook readers. In the EEITC, phrasal lexical bundles and the majority of clausal bundles indicate quantity (*the sum of the*), discipline specificity (*input resistance of the*), location (*in the circuit of*), logical relationship (*a result of the*). They also serve to structure definitions (*is known as the*), to display scientific formulas (*can be written as*), and to describe processes (*the direction of the*) and physical connection among electrical components (*in parallel with the*). Some clausal bundles are associated with expressing possibilities toward particular propositions or getting readers' involved in the subject content or guide readers to take actions.

**5.3 Functional characteristics of lexical bundles in the EEITC**

To make a more systematic analysis of the functions of EEITC lexical bundles, I used the functional taxonomy of lexical bundles in university teaching and textbooks (Biber et al., 2004a) as the analytical framework to identify pragmatic functions. Table 5.3 shows the 105 four-word lexical bundles identified in the EEITC. These bundles, classified into a set of functional categories and subcategories according to the functional taxonomy (Biber et al., 2004a), were divided into three main types: stance bundles, discourse organizers, and referential bundles. Each main function category has several subcategories with more specific meanings and functions (See Chapter 3). Within each functional subcategory, the lexical bundles are sequenced in descending order of frequency. For example, among the nine epistemic stance bundles in Table 5.3, the most frequent bundle is *the fact that the*, while the least frequent is *be found from the*. 
Table 5.3 Functional classification of 4-word lexical bundles in the Electrical Engineering Introductory Textbook Corpus (EEITC)

<table>
<thead>
<tr>
<th>Categories</th>
<th>Lexical bundles (sequenced in the order of descending frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. STANCE BUNDLES</strong></td>
<td></td>
</tr>
<tr>
<td>A. Epistemic stance</td>
<td></td>
</tr>
<tr>
<td>A1) impersonal:</td>
<td>the fact that the, from the fact that, to note that the, we assume that the, can be used to, can be obtained by, can be written as, can be found from, be found from the</td>
</tr>
<tr>
<td>A2) personal:</td>
<td>we note that the, we see that the, we find that the</td>
</tr>
<tr>
<td>B. Attitudinal/modality stance</td>
<td></td>
</tr>
<tr>
<td>B1) desire</td>
<td>we wish to determine, wish to determine the, we wish to find, wish to find the</td>
</tr>
<tr>
<td>B2) obligation/directive</td>
<td>consider the circuit shown, let us determine the, let us find the</td>
</tr>
<tr>
<td>B3) intention/prediction Personal:</td>
<td>we will use the</td>
</tr>
<tr>
<td><strong>II. DISCOURSE ORGANIZERS</strong></td>
<td></td>
</tr>
<tr>
<td>A. Topic introduction/focus</td>
<td></td>
</tr>
<tr>
<td>B. Topic elaboration/clarification on the other hand, as well as the, as long as the</td>
<td></td>
</tr>
<tr>
<td><strong>III. REFERENTIAL EXPRESSIONS</strong></td>
<td></td>
</tr>
<tr>
<td>A. Identification/focus</td>
<td>is known as the, is determined by the, in the case of, referred to as the, is independent of the, is defined as the</td>
</tr>
<tr>
<td>B. Imprecision</td>
<td></td>
</tr>
<tr>
<td>C. Specification of attributes</td>
<td></td>
</tr>
<tr>
<td>C1) Quantity specification</td>
<td>is equal to the, in the range of, the sum of the, is proportional to the, is the sum of, is directly proportional to, the remainder of the</td>
</tr>
<tr>
<td>C2) Tangible framing attributes</td>
<td>the magnitude of the, the value of the, the voltage across the, in the design of, the current in the, the gain of the, the operation of the, the ratio of the, the slope of the, the frequency response of, the voltage at the, input resistance of the, the output of the, the input resistance of, the values of the, gain of the amplifier, on the value of, the current through the, of the input signal, voltage across the capacitor, the polarity of the, the voltage gain of, equivalent circuit for the, the difference between the, energy stored in the, for the network in, the relationship between the, the direction of the, voltage gain of the, an expression for the, the energy stored in, the expression for the, the amplitude of the, a plot of the, a sketch of the, value of the current, the location of the, the resistance of the, power delivered to the, a function of the, the details of the</td>
</tr>
</tbody>
</table>
Categories | Lexical bundles (sequenced in the order of descending frequency)
--- | ---
C3) Intangible framing attributes | in terms of the, with the result that, with respect to the, a result of the, the effect of the, the use of the, the case of the, the presence of the
D. Time/place/text reference | is connected to the, in series with the, in the frequency domain, in the time domain, in parallel with the, in the circuit in, in the circuit of, in the network in, present in the network, delivered to the load, be connected to the
D1) Place reference | is of the form, it follows that the, given by the expression, in this case the, is given by the
D2) Time reference |
D3) Text deixis | at the end of, the end of the
D4) Multi-functional reference |

5.3.1 Referential bundles in the EEITC

Three main subcategories of referential bundles were identified in the EEITC: referential identification/focus bundles, attribute specification bundles and time/place/text bundles (see Table 5.3). Bundles of these three types in the EEITC respectively identify important entities, specify concrete or abstract characteristics, and make reference to time, place and text.

5.3.1.1 Referential identification/focus bundles in the EEITC

Referential identification/focus bundles direct readers' attention to the noun phrase following the bundles, and thus are widely used for the purpose of introducing concepts/definitions or identifying particular situations in the EEITC. In this category, the bundle with the highest frequency count is *is known as* the, which occurs 31 times in the corpus. *Is known as the* introduces new definitions, naming the process or entity right before the bundle with terms commonly accepted in the discipline. This function can be observed from the following concordance lines of the bundle:

1. a special loop techniques and the other *is known as* the supermesh approach
2. facilitate plotting, PSPICE uses what *is known as* the PROBE utility, which
at the output is $V_o$. The point $Q$ is known as the quiescent point, the
degree of amplifier performance is known as the amplifier frequency
response of the amplifier; the first is known as the magnitude or amplitude
The other two bundles, referred to as the, and is defined as the, function the same way as
is known as the in the original context, reflecting the conventional practice of introducing
terms and definitions in introductory textbooks.

*In the case of* signals that readers should pay attention to the noun phrase after the
bundle, as we can observe in the first five concordance lines of the bundle:

1. g.2.3c. Note that *in the case of* a short circuit where $R=$
2. Via a single algebraic equation. *In the case of* the single-node-pair circuit
3. to specify capacitors and inductors. *In the case of* capacitors, the
4. capable of storing electric energy. *In the case of* a charged capacitor, the
5. Current and voltage are in phase. *In the case of* a resistor shown in Fig.

Expanding the 4th concordance line, we get:

We begin our discussion by recalling that in Chapter 5 we found that capacitors
and inductors are capable of storing electric energy. *In the case of* a charged
 capacitor, the energy is stored in the electric field that exists between the
positively and negatively charged plates. This stored energy can be released if ...
(C:chapter06)

In the beginning of the above paragraph, the author mentions two types of energy storage
components: capacitors and inductors. *In the case of* guides readers' attention to the
particular situation of a charged capacitor, where afterwards the detailed charging process
is described.

Some of the referential identification bundles show multi-functions in original
context. In addition to guiding readers' attention to the noun phrase after, the two bundles
*is determined by*, and *is independent of* the identify cause/effect relationship in the
original context, as in:

1. zero voltage; the current in the short *is determined by* the rest of the circuit
2. the voltage across the open terminals *is determined by* the rest of the circuit
3. rate at which this exponential decays *is determined by* the time constant $t$. A
4. RC is the network's time constant. $K_2$ *is determined by* the initial condition of
5. the frequency selectivity of the circuit *is determined by* the value of Q. A
Besides, the referential bundles with the passive form (e.g., *is known as the*, and
referred to as the) reflect the impersonal style of technical and scientific writing, as in:

Thus we will demonstrate two techniques dealing with this type of situation. One
is a special loop technique and the other *is known as the* supermesh approach.

(CAchapter03)

5.3.1.2 Attribute specification bundles in the EEITC

Attribute specification bundles in the EEITC specify quantity or frame
tangible/intangible attributes of the noun phrases after the bundles. Table 5.3 lists seven
quantity specification bundles found in the EEITC, including *is equal to the*, *in the range
of*, *the sum of the*, *is proportional to the*, *is the sum of*, *is directly proportional to*, and *the
remainder of the*. The first five concordance lines of *is equal to the* below exemplify its
function of specifying quantity characteristics:

1. Note that the power supplied *is equal to the* power absorbed.
2. the sum of the currents entering a node *is equal to the* sum of the currents
3. of two resistors connected in parallel *is equal to the* product of their
4. or, since Vb *is equal to the* voltage drop across the
5. to point a in the wye (i.e., Ra) *is equal to the* product of the two

In addition to quantity specification bundles, most of the attribute specification
bundles in the EEITC either frame physical characteristics or abstract properties of
following noun phrases, and correspondingly they are labelled as tangible or intangible
attribute framing bundles. Examples of the former subcategory include *the magnitude of
the*, *the value of the*, *the voltage across the*, etc., and bundles of the latter type are like *in
terms of the*, with the result that, with respect to the, and so on.

*The magnitude of the*, a tangible attribute referential lexical bundle with the highest
frequency in the EEITC, occurs in 63 concordance lines and captures *magnitude* as a very
important specification parameter for electric current, voltage, gain, and other many
variables in circuit analysis. To better understand its functions, the first five concordance
lines of this bundle are given as follows:

1. it is important to specify not only the magnitude of the variable
2. yields the equation ... Therefore, the magnitude of the current is 0.1mA,
3. under these condition is ... Thus, the magnitude of the gain is ... Thus, the
4. the magnitude of the gain is ... Thus, the magnitude of the gain is 165.29. At
5. When a voltage or current is selected, the magnitude of the phasor will be

Intangible attribute framing bundles frame abstract properties of noun phrases after

the bundles, as shown in the concordance lines of a result of the:

1. between the two nonreference nodes as a result of the presence of the voltage
2. in the last section). This factors is a result of the input capacitance Ci,
3. and Va will correspondingly increase as a result of the high (ideally infinite)
4. (Eq. 2.5). This is no coincidence; it is a result of the fact that both the
5. The input offset voltage arises as a result of the unavoidable mismatches

Beside attribute specification functions, it is also found in both tangible and

intangible attribute framing bundles that some bundles are closely associated with the

cause/effect logical relationship in the original text. Those bundles consist of on the value

of, with the result that, a result of the, the effect of the, and the presence of the. For

instance, the first 5 concordance lines of on the value of are listed below:

1. or may not be the case, depending on the value of the various voltage
2. this equation specifies an upper limit on the value of the mutual inductance.
3. Since the transfer function is dependent on the value of the capacitor, let us
4. To minimize the effect of this variation on the value of the signal that appears
5. Thus the voltage Vo will not depend on the value of the current that might

On the value of in the above lines presents the cause/effect relationship in the original

text. Among the total of 18 concordance lines with on the value of in the EEITC, depend

(v.), dependence and dependent occur 13 times in front of on the values of, clearly

indicating the bundles as part of the cause in a cause/effect relationship; this is illustrated

in concordance lines 1, 3, and 5. In the remaining 5 concordance lines, the three nouns

found in front of the bundle are limit, effect, and influence, positioning the bundle as the
effect in a cause/relationship, as seen in the line 2 and 4 above.

The association of lexical bundles such as with the result that, a result of the, and
the effect of the with the effect in the cause/effect relationship seems to be self-evident, but it is not the same case for the presence of the. In the original context, the presence of the marks the existence of a particular entity, which in the three concordance lines below marks specific consequences, such as destroying the symmetrical nature of the node equation, blocking the flow of the conduction current, or having deleterious effect on the integrator circuit.

1. Note that the presence of the dependent source has destroyed the symmetrical nature of the node equation. (CAchapter03)
2. Because of the presence of the dielectric, the conduction current that flows in the wires that connect the capacitor to the remainder of the circuit cannot flow internally between the plates. (CAchapter05)
3. It should be clear from this discussion that the integrator circuit will suffer deleterious effects from the presence of the op-amp input dc offset voltage and current. (MEchapter2)

5.3.1.3 Place/text/multi-functional referential bundles in the EEITC

The number of place/text/multi-functional referential bundles is second only to that of attribute specification bundles. Among the total of 18 place/text/multi-functional referential bundles (See Table 5.3), 12 are place referential bundles, indicating location information. The rest are text deixis bundles and multi-functional referential bundles, the former of which make reference to particular sections of the text and the latter of which make time or place reference.

Is connected to the, in series with the, in the frequency domain together with other place referential bundles indicate place information, exemplified by the first 5 concordance lines with is connected to the:

1. range. When the unknown resistor is connected to the bridge, R3, is
2. in which an independent voltage source is connected to the reference node.
3. sonic simple electronic device that is connected to the ac wall plug in our
4. area. Suppose now that a source is connected to the capacitor shown in
5. the passive electrode of the stimulator is connected to the ground electrode of

Place referential bundles like in series with the, in parallel with the, is connected to
the, delivered to the load, and be connected to the serve as describing the exact spatial connection among electrical components in electrical circuits, as we can see from the concordance lines with in series with the:

N Concordance

1. with a voltage source of value $v=\alpha R$ in series with the resistor $R$. The
2. parallel impedance for a voltage source in series with the impedance, as
3. as an ammeter and must be placed in series with the branch current of
4. the resistors. 8. Connect a diode in series with the 560 Ohms resistor
5. In this case, a large capacitance in series with the signal will create a DC

Text deixis bundles in the EEITC (is of the form, it follows that the, in this case the, given by the expression) make reference to text or formulas, as in:

1. Since the roots are real and unequal, the circuit is overdamped, and $v(t)$ is of the form

   $$v(t)=Ke^{-2t}+K'e^{-0.5t}$$

   (CAchapter06)

2. Since $n$ is a function of temperature (Eq. 3.36), it follows that the concentration of the minority holes will be a function of temperature whereas that of the majority electrons is independent of temperature. (MEchapter3)

3. An important circuit is the single-node-pair circuit. In this case the elements have the same voltage across them and, therefore, are in parallel. (CAchapter02)

   It is interesting to see that the second example, it follows that the makes both anaphoric and cataphoric references at the same time, where it refers back to the preceding text and that makes reference to the following proposition. In this case the makes anaphoric reference by referring back to the clause preceding the bundle.

Lexical bundles mentioned above function as endophorics by connecting written texts with diagrammatic texts (formulas, diagrams), which “interact with the text and the reader in complex ways, providing a scientific rendering of knowledge which moves learners from their person first-had experience of the world towards the world of abstract scientific knowledge” (Hyland, 2004, p. 119). The bundles in the above sentences act like glue by integrating the two modes of texts together and create meanings out of the new combination. Recognizing lexical bundles as endophorics offers learners an awareness of how disciplines create new meanings out of different modes of text.
Multi-functional referential bundles refer to time or place information determined by the original context. In the first 5 concordance lines of \textit{at the end of}, the first four lines refer to the particular location and the fifth line locates to a specific time point:

N Concordance
1. numerous resistors, it is recommended that the analysis begin \textit{at the end of} the network opposite the terminals. Two or more resistors are
2. vehicle used to demonstrate each facet of this approach. Finally, \textit{at the end of} this section, we will outline a strategy for attacking any
3. and systematically proceed to those that are more difficult). Then \textit{at the end of} this section we will outline a general strategy fore employing
4. symbolic pencil. To connect the top of V I to R I, point the pencil \textit{at the end of} the wire stub protruding from V I, click once and release. Next,
5. requires that the energy stored in the inductor must be the same \textit{at the end of} each switching cycle. Recalling that the inductor energy is relate

5.3.2 Stance bundles in the EEITC

A total of 20 stance bundles were identified in the EEITC, accounting for about 20% of the 105 bundles. This seems to contradict the notion by Biber et al. (2004a) that stance bundles are rare in the textbook register, with only 5 stance bundles found in the textbook sub-corpus of the T2K-SWAL corpus. The fact that introductory textbooks of Electrical Engineering have more stance bundles suggests that the authors more often express assessment or attitudes towards following propositions. The stance bundles identified in the EEITC fall into two categories: epistemic stance bundles for evaluation of the knowledge status of following propositions and attitudinal/modality stance bundles for the authors' attitudes toward ongoing events.

5.3.2.1 Epistemic stance bundles in the EEITC

Taking both impersonal and personal forms, epistemic stance bundles in the EEITC assess following propositions. As we see in Table 5.3, the majority of epistemic stance bundles are impersonal ones, indicating that textbook authors prefer to assess following propositions without involving their personal opinions. This feature of language use is
consistent with the impersonal style of discipline textbooks (Conrad, 1996; Reppen, 2004), where personal stances are avoided in arguments and claims. The underlying reason can be attributed to textbook writers’ intention of maintaining objectivity, “allowing them to adopt an authoritative command of their data and their audience” (Hyland, 2002a, p. 232).

Impersonal epistemic stance bundles in the EEITC indicate logical certainty/possibility/ positive evaluation, implicitly emphasizing what writers deem as essential and important or hedging the claims that have been made. From the fact that and the fact that the introduce outside supports without explicitly speaking out authors’ certainty towards arguments and claims, as we see in the following two concordance lines:

1. For the time being note that the op amp is a differential-input single-ended-output amplifier, with the latter term referring to the fact that the output appears between terminal 3 and ground. (MEchapter02)
2. The usefulness of the arrow notation stems from the fact that we may want to label the voltage between two points that are far apart in a network. (CAchapter2)

With the form of a modal verb + passive verb, impersonal epistemic stance bundles (can be used to, can be obtained by, can be written as, and can be found from) express logical possibility toward following propositions, hedging statements made in the original text without blocking the possibility of potential alternatives, as we can see in following examples:

1. Ebers and Moll, two early workers in the area, have shown that this composite model can be used to predict the operation of the BJT in all of its possible modes. (MEchapter05)
2. Another way of looking at this point can be obtained by observing that each half of the secondary winding of the center-tapped transformer is utilized for only half that time. (MEchapter3)

Among the impersonal epistemic stance bundles, one noticeable issue is that we assume that the is categorized as an impersonal bundle although the bundle starts with the
first person plural pronoun, *we*. *We assume that the* directs readers’ attention to particular lines of argument that writers have assessed as certain. However, it is hard to conclude that *we* represents both readers and writers. Let us examine two text extractions of this bundle below:

1. *We assume that the* voltage source delivers V volts regardless of what is connected to its terminals. Theoretically, we could adjust the external circuit so that an infinite amount of current would flow, and therefore the voltage source would deliver an infinite amount of power. (CAchapter01)
2. Figure 2.6(a) shows the equivalent circuit, and the analysis proceeds as follows: The gain A is very large (ideally infinite). If *we assume that the* circuit is "working" and producing a finite output voltage at terminal 3, then the voltage between the op amp input terminals should be negligibly small and ideally zero. (MEchapter2)

In the above two examples, *we* in fact refers to the textbook authors or their peers since the assumptions in the two examples can only be made by discipline experts, not by novices. In the second example, when the author sets up a particular situation for readers to consider, instead of saying “*if I assume that the circuit is “working”...*”, the author uses *we* to start the sentence for the purpose of focusing readers’ attention on the assumption. This assumption can hardly be made by novices since they have not reached the level which allows them to apply particular restraints for the purpose of simplifying circuit analysis. So this usage distances readers from discipline experts, and thus in an impersonal way elaborates particular conditions for following propositions.

Additionally, it is interesting to observe how the linguistic environment affects the pragmatic functions of *to note that the*. The linguistic environment around *to note that the* is often filled by adjectives with positive meanings, such as *interesting, useful, important,* and so on, as in the concordance lines of *to note that the* below:

N Concordance
1. book. In the meantime, it is interesting *to note that the* magnitude of currents
2. (...) Therefore, (...) it is interesting *to note that the* 3-O resistor never enters
3. in Fig. 13. 11. C=cos 0. It is important to note that the damping ratio and the
damping ratio and the Vds curves in saturation. It is important to note that the model only predicts
with a common ground. It is interesting to note that the reference grounding
Using the three personal epistemic bundles, we note that the, we see that the, and we
find that the, textbook authors explicitly and confidently state knowledge claims (as in
the first example below) or analysis results obtained so far (as in the second and third
textbooks below) for the purpose of establishing convincing arguments/analysis. At the
same time verbs such as find, see, and note guide readers’ attention to the propositional
facts and the use of collective pronoun we considers readers as active participants of the
analysis/argument procedure.

1. In effect, we are now faced with solving a set of algebraic equations for the
unknown phasors. The phasors are then simply transformed back to the time
domain to yield the solution of the original set of differential equations. In
addition, we note that the solution of sinusoidal steady-state circuits would be
relatively simple if we could write the phasor equation directly from the circuit
description. (CAchapter7)
2. Combining the model in Fig. 3.25, the values in Table 3.1, and these I/O
limitations, we can produce the graphs in Fig. 126, which show the output-input
relationship for each op-amp outlined in Table 3.1. From the graph we see that
the LM6492 and MAX 4240 have rail-to-rail outputs and the LM324 and MAX
4240 have rail-to-rail outputs and the LM324 and PA03 do not. (CAchapter3)
3. We repeat the simulation with the dc differential input voltage set to -1V to
find the negative-output saturation voltage. Accordingly, we find that the uA741
op amp has a maximum output voltage $V_{omax} = 14.8V$. (MEchapter 2)

5.3.2.2 Attitudinal/modality stance bundles in the EEITC

Attitudinal/modality stance bundles in the EEITC (see Table 5.3) mark textbook
authors' attitude towards following propositions, including desire, obligation/directive
and intention. We wish to determine and we wish to find, categorized as desire bundles,
often occur in the instructions for learning tasks where authors require learners to
demonstrate their ability to apply the analysis of circuit parameters. Desire bundles are
used by writers to clarify the ends of those learning tasks; they plan the discourse and
clearly structure the text by signposting the writers’ purposes. The functions of these
bundles can be observed in following concordance lines:

1. The circuit in Fig. 9.1 has the following parameters: ... We wish to determine equations for the current and the instaneous power as a function of time and plot these functions with the voltage on a single graph for comparison. (CACchapter09)
2. We wish to find all the currents and voltages labelled in the ladder network shown in Fig. 2.30a. (CACchapter02)

Obligation/directive bundles, including consider the circuit shown, let us determine the, and let us find the express the writer's stance of directing readers to participate in particular learning tasks. These bundles appear in imperative clauses and become the formulaic part of textbook language. In the first following concordance lines with consider the circuit shown, the authors explicitly guide readers' attention to particular circuits. In the second and third concordance lines below with let us determine the and let us find the, readers are explicitly instructed about the goals of the learning tasks and are reminded that the following analysis centers around those ends.

N Concordance
1. At this point we wish to extend our analysis to include a multiplicity of voltage sources and resistors. For example, consider the circuit shown in Fig.2.18a. (CACchapter02)
2. Once again, let us determine the current in the RL circuit examined in Example 7.3. (CACchapter07)
3. Given the network in Fig. 2.34a, let us find the source current $I_s$. (CACchapter02)

Intention bundles indicate the writers' intention. The only intention bundle in the EEITC is we will use the, which emphasizes the issues that the author considers as essential for explicitly explaining a particular procedure. In the first example below, we will use the clarifies the focus of the statement as the initial condition value, which becomes the focus of the following discourse. In the second example, we will use the clearly states the model used for circuits analysis.

N Concordance
1. The value of the voltage source for the capacitor or the current source for the inductor is a known value at one instant of time. In general, we will use the
initial condition value since it is generally the one known, but the value at any instant could be used. This value can be obtained in numerous ways and is often specified as input data in a statement of the problem. (CAchapter06)

2. We are not ready to consider the analysis of BJT circuits to which only dc voltages are applied. In the following examples we will use the simple model in which, $|V_{BE}|$ of a conducting transistor is 0.7 V and the $V_{CE}$ of a saturated transistor is 0.2V, and we will neglect the Early effect. Better model can, of course, be used to obtain more accurate results. (MEchapter5)

5.3.3 Discourse organizing bundles in the EEITC

In the EEITC, only three bundles, *on the other hand*, *as well as the*, and *as long as* the are identified as discourse organizers. *On the other hand* signals the contrast between coming and prior discourse, as in:

..., the power delivered to the load (from the secondary side of the transformer) is less than or at most equal to the power supplied by the signal source. *On the other hand*, an amplifier provides the load with power greater than that obtained from the signal source. (MEchapter01)

*As well as the* juxtaposes the coming and prior discourse, as in:

This charge equilibrium enables us to calculate the average diode current *as well as the* details of the output waveform. (MEchapter3)

*As long as* the specifies the following proposition as one condition for prior proposition to become true, as in:

Breakdown of the CBJ in either the common-base or common-emitter configuration is not destructive *as long as* the power dissipation in the device is kept within safe limits. (MEchapter5)

5.4 Lexical bundles in the ESP textbooks

McDounough (1998) evaluated the design of six ESP textbooks, two of which were included in the ESPTC. The two textbooks are *Oxford English for electronics* (Glendinning & MacEwan, 1993) and *Oxford English for electrical and mechanical engineering* (Glendinning & Glendinning, 1995). In McDounough’s research, the two ESP textbooks feature four main characteristics: 1) the content is discipline specific, 2) authentic readings from the discipline are used and both technical and nontechnical topics are included, 3) a skill-based approach has been taken, and 4) key functions are identified
and emphasized. For instance, *Oxford English for electronics* (Glendinning & MacEwan, 1993) adapts authentic readings from the discipline and covers a broad range of topics. Most of the topics are relevant to discipline contents, such as circuits, transistor characteristics, data transmission and so on. The textbook also covers non-technical components, such as choosing a course, and careers in electronics.

The textbook takes a skill-based approach and includes a variety of activities that aim to develop the four language skills. In language practice, key functions are emphasized, such as making definitions, describing processes, and making predictions. The ESP textbook offers exercises that simulate real learning tasks in the discipline. Diagrams, figures and equations, essential parts of Electrical Engineering introductory textbooks, are used to supplement written information and facilitate the understanding of the analysis process of electrical circuits. If the students can explain the visual information in introductory textbooks, both in spoken and written modes, they may be able to explain the process and key information in the diagrams to their peers or instructors. In Unit 23 of *Oxford English for electronics* (Glendinning & MacEwan, 1993), task 9 offers the diagram of a process control system. Then through the fill-in-the-blanks exercise, learners need to explain the control process.

However, there is little empirical research that has investigated whether language in reading materials in ESP textbooks is consistent with target language use, in this case, introductory Electrical Engineering textbooks. Reading materials play a key role in determining the language input that the ESP students are faced with and construct the basis for learning tasks in the ESP textbooks. Although the two ESP textbooks included in the ESPTC state that they use authentic materials, it is not clear how authentic these
pedagogical materials are when compared with Electrical Engineering introductory textbooks. Consequently, I posed questions about lexical bundles in the ESPTC: Do lexical bundles in the EEITC (the target corpus) also occur in the ESPTC? What are the functional characteristics of lexical bundles in the ESPTC?

In this part, I report on the lexical bundles present in the ESPTC and their pragmatic functions, and later I compare the bundles and their functions in both corpora, the EEITC and the ESPTC, to examine if there is a gap of language use between the two corpora. Table 5.4 includes the lexical bundles present in the ESPTC, which are classified into functional subcategories according to their pragmatic functions in the original text.

<table>
<thead>
<tr>
<th>Table 5.4 Functional classification of 4-word lexical bundles in the ESP Textbook Corpus (ESPTC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories</td>
</tr>
<tr>
<td>I. STANCE BUNDLES</td>
</tr>
<tr>
<td>A. Epistemic stance</td>
</tr>
<tr>
<td>A1) impersonal:</td>
</tr>
<tr>
<td>B. Topic elaboration/clarification</td>
</tr>
<tr>
<td>II. DISCOURSE ORGANIZERS</td>
</tr>
<tr>
<td>C. Identification/focus</td>
</tr>
<tr>
<td>C1) Quantity specification</td>
</tr>
<tr>
<td>C2) Tangible framing attributes</td>
</tr>
<tr>
<td>C3) Intangible framing attributes</td>
</tr>
<tr>
<td>D. Place/time reference</td>
</tr>
<tr>
<td>D1) Place reference</td>
</tr>
<tr>
<td>D2) Time reference</td>
</tr>
</tbody>
</table>
5.4.1 Stance bundles in the ESPTC

The three stance bundles, *the fact that the*, *from the fact*, and *can be used to* are impersonal epistemic stance bundles indicating assessment of following propositions (see Table 5.4). As we see in the following two concordance lines from the ESPTC, the bundle in the first example expresses certainty towards the following proposition by emphasizing it as a fact. The bundle in the second example indicates the logical possibility of following propositions and thus hedges the statement made by the author without excluding other possible solutions:

1. The values may seem illogical at first sight, but this is not so. They stem from *the fact that the* tolerance extremes of a value reach the extremes of adjacent values, thereby covering the whole range without overlap. (English for electronics)
2. In a plotter, for example, by using two motors running at 90 degrees to each other, they *can be used to* drive a pen an exact distance in all directions. (English for electrical & mechanical engineering)

5.4.2 Discourse organizers in the ESPTC

Similar to the EEITC, which has only three four-word lexical bundles as discourse organizers, the ESPTC has only one discourse organizer, *as long as the*. The lexical bundle serves to connect two propositions and clarifies that the latter proposition is the necessary condition for the former proposition to happen, as in the following example:

Once again, the attraction and repulsion between it and the field magnet make it turn. The armature continues turning *as long as the* direction of the current, and therefore its magnetic poles, keeps being reversed. (English for electrical & mechanical engineering)

5.4.3 Referential bundles in the ESPTC

In the ESPTC, the four identification/focus referential bundles are *is known as the*, *is determined by the*, *in the case of*, and *referred to as the*. As we can see in the four samples below, the four bundles in the original text guide readers' attention to the noun phrases after the bundles, which in the following concordance lines respectively refer to a
new concept, a determinant of an event, a new situation, or an established term:

1. The overall effect of this phenomenon for the whole amplifier is known as the total harmonic distortion (THD). (English for electronics)
2. This frequency is determined by the values of the inductance and the capacitance. (English for electronics)
3. In the case of diodes for specialized applications, the third letter does not generally have any particular significance. (English for electronics)
4. These managers may be referred to as the Management Team. (English for electrical & mechanical engineering)

Attribute specification referential bundles in the ESPTC specify quantity, tangible or intangible features of particular entities. Is proportional to the is the only quantity specification bundle found in the ESPTC (see concordance line 1 below). Tangible attribute framing bundles like the magnitude of the, in the design of (see concordance line 2, 3 below), and so on specify physical characteristics of the following entities. Intangible attribute framing bundles such as with the result that, with respect to the, and so on, specify the abstract characteristics of the following entities (see concordance line 4 below).

N Concordance

1. The change in voltage is proportional to the load. (English for electrical & mechanical engineering)
2. This causes the electron beam to be deflected vertically in proportion to the magnitude of the input signal. (English for electronics)
3. Qualified to degree level and preferably holding professional status you will have excellent knowledge of and expertise in the design of HV and LV distribution and control systems, gained ideally during your time in a manufacturing and operational/maintenance environment. (English for electrical & mechanical engineering)
4. The load to be moved is lifted only a few centimetres and as a result of the low pressure (1-2 bar) no clouds of dust are formed and the floor cannot be damaged. (English for electrical & mechanical engineering)

Place/text referential bundles in the ESPTC indicate spatial relationship between entities and accurate time points. Place referential bundles are connected to the, in series with the, and in parallel with the, functioning the same way as in the EEITC and indicating the spatial connections among components with the purpose of explaining the
circuit diagrams, as in the following text samples:

1. The positive terminal of the capacitor is connected to the output of the AF amplifier and the wiper of the pot is connected to the output of the AF amplifier and the wiper of the pot is connected to the power amp. (English for electronics)
2. The charger is switched on and off by a double-pole switch connected in series with the mains input. (English for electronics)
3. It consists of a large electrolytic capacitor connected in parallel with the rectifier as shown in Fig. 4. (English for electronics)

In the EEITC, the two bundles, at the end of the war they were further developed as mobile phones for use by the emergency services and other services such as taxis. (English for electronics)

Overall, the lexical bundles in the ESP textbooks have the same functions as they do in introductory textbooks of Electrical Engineering. However, the total number of bundles in the ESP textbooks is much lower than that in the EEITC, which demonstrates that the ESP textbooks for university students of Electrical Engineering may not offer readers enough language input that is similar to target language use of lexical bundles in the entry-level Electrical Engineering textbooks. The next section discusses the distribution differences of lexical bundles between the two corpora with the aim of identifying the gap of language use.

5.5 The comparison of lexical bundles in the ESPTC and the EEITC

The previous four sections report on structural and functional characteristics of lexical bundles in the EEITC and the ESPTC; however, two research questions remain to be answered: 1) Is there a gap of language use between the EEITC and the ESPTC? 2) How do lexical bundles help structure the text in the EE introductory textbooks and the ESP textbooks? In this section, I compare lexical bundles and their pragmatic functions in the two corpora. The following discussion centers on the role that lexical bundles play in
constructing the text in the discipline introductory textbooks and the ESP textbooks.

An initial comparison of the bundles in the ESPTC and the EEITC shows that only 33 bundles, one third of the 105 bundles in the EEITC, appear in the two ESP textbooks. Figure 5.1 compares the number of lexical bundles in the EEITC and in the ESPTC. We can see that the number under each function category in the ESPTC is much lower than the corresponding number in the EEITC.

**Figure 5.1** Functional distributions of lexical bundles in the EEITC and the ESPTC

Figure 5.1 also shows the gap between the numbers of stance bundles in the two corpora; there are 20 stance bundles in the EEITC while the ESPTC only has 3 stance bundles. In Figure 5.1, the ESPTC has far fewer referential bundles than the EEITC; there are around 80 referential bundles in the EEITC while the number shrinks to 29 in the ESPTC. Not only is the number of lexical bundles in the ESPTC much lower than that in the EEITC, the subcategories of the ESPTC are fewer than those in the EEITC. As we can see in Figure 5.2, there are a total of 13 subcategories of lexical bundles in the
EEITC but only 8 in the ESPTC.

Compared with the Electrical Engineering introductory textbooks, the ESP textbooks overemphasize the use of elaborated reference while at the same time not fully representing stance bundles that express assessments or attitudes towards following propositions. Table 5.5 below lists the percentages of lexical bundles under each subcategory in the EEITC and ESPTC. It is noticeable that 87.88% of the lexical bundles are referential bundles in the ESPTC, much higher than in the EEITC (78.10%). The percentage of stance bundles in the ESPTC is 9.09%, much lower than 19.05%, the percentage in the EEITC.

**Table 5.5** Proportional function distribution of lexical bundles in the EEITC and ESPTC

<table>
<thead>
<tr>
<th>Main function types of lexical bundles</th>
<th>Sub-category functions of lexical bundles</th>
<th>the percentage in the EEITC</th>
<th>the percentage in the ESPTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stance bundles</td>
<td>Epistemic stance Bundles: personal</td>
<td>2.86%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Epistemic stance bundles: impersonal</td>
<td>8.57%</td>
<td>9.09%</td>
</tr>
<tr>
<td></td>
<td>Attitudinal/modality stance bundles: desire</td>
<td>3.81%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Attitudinal/modality stance bundles: directive/obligation</td>
<td>2.86%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Attitudinal/modality stance bundles: intention</td>
<td>0.95%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>19.05%</td>
<td>9.09%</td>
</tr>
<tr>
<td>Discourse organizers</td>
<td>Topic clarification</td>
<td>2.86%</td>
<td>3.03%</td>
</tr>
<tr>
<td>Referential bundles</td>
<td>Identification/focus</td>
<td>5.71%</td>
<td>12.12%</td>
</tr>
<tr>
<td></td>
<td>Attribute specification: quantity</td>
<td>6.67%</td>
<td>3.03%</td>
</tr>
<tr>
<td></td>
<td>Tangible attribute specification</td>
<td>40.95%</td>
<td>42.42%</td>
</tr>
<tr>
<td></td>
<td>Intangible attribute specification</td>
<td>7.62%</td>
<td>15.15%</td>
</tr>
<tr>
<td></td>
<td>Place referential bundles</td>
<td>10.48%</td>
<td>9.09%</td>
</tr>
<tr>
<td></td>
<td>Text deixis bundles</td>
<td>4.76%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Time referential bundles</td>
<td>0%</td>
<td>6.06%</td>
</tr>
<tr>
<td></td>
<td>Multi-function bundles</td>
<td>1.90%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>78.10%</td>
<td>87.88%</td>
</tr>
</tbody>
</table>
To further identify the distribution differences of functional categories of lexical bundles in the two corpora, I mainly used the data in Figure 5.2 and Table 5.5 to examine those differences along the dimension of subcategory functions. Table 5.5 lists the percentages of functional subcategories in the EEITC and the ESPTC. Figure 5.2 stacks the percentages of subcategories of lexical bundles in the ESPTC with those in the EEITC, and thus offers a direct view of the comparison of the subcategory distribution of lexical bundles in the two corpora.

Figure 5.2 Proportional distributions of function subcategories of lexical bundles in the EEITC and the ESPTC
5.5.1 The comparison of referential bundles in the EEITC and the ESPTC

In Electrical Engineering, problem-solving skills are highly valued and expected to be key characteristics of professional electrical engineers. Introductory textbooks have a clear focus on applying well-established principles to simplify and analyze the circuits for solutions to problems. Each chapter starts with the introduction of core concepts and basic principles and moves to the section of learning examples sequenced in ascending order of difficulty to demonstrate how to analyze electrical circuits. To get the best understanding of those items, readers need to refer back to the written text and graphics, such as blocks, diagrams, equations and numeric expressions, and figure out how to apply those to determine particular parameters in a circuit (i.e., the amplitude of the current, voltage charges, etc.). At the end of each chapter, there are more exercises designed to teach students how to solve a variety of problems for the purpose of solidifying their newly acquired knowledge of core concepts and improving their analytical skills. In most Electrical Engineering introductory courses, there are lab sessions which further students' understanding of the core concepts through experiments and reporting results in lab reports.

Lexical bundles and their pragmatic functions in the EEITC show that the corpus to some extent has commonality with the university textbook subcorpus of the T2K-SWAL corpus (Biber et al., 2004a). In both corpora referential bundles are most commonly used to describe referential properties of propositions, which is in line with the goal of academic textbooks to present the norms of the field and construct a body of knowledge. Introductory EE textbook writers commonly use referential bundles to describe content or context and to specify attributes of variables, circuits, equipments and other aspects of
circuit analysis. Referential bundles in the EEITC identify items that writers deem as important (is known as the, is determined by the), specify concrete/abstract attributes of variables or electrical components (is equal to the, the magnitude of the, with the result that), or make reference to time, location and text (at the end of, is connected to the, is of the form). These bundles describe technical details relevant to circuit analysis and explain how the analysis is conducted.

On the surface, the overall predominance of referential bundles in both the EEITC (78.10%) and the ESPTC (87.88%) seems to show that for both types of texts information production is the primary task for textbook writers. Information about tangible attributes (i.e., the magnitude of the), quantity (i.e., is equal to the), place (i.e, is connected to the), time (i.e., at the end of) and so on is framed by those referential bundles. However, the introductory textbooks and ESP textbooks differ significantly in how they make use of those referential bundles to construct precise information in the text. This can be observed from the subcategory distribution of those bundles in the two corpora.

Compared with the subcategories of referential bundles in the EEITC, the ESPTC lacks two subcategories: text deixis bundles and multi-functional reference bundles. Without text deixis bundles in the ESPTC, readers do not have the chance to observe the reader/author interaction and text referential functions in written text. The use of text deixis bundles in the EEITC (is of the form, it follows that the, given by the expression) indicates interaction with readers (Hyland, 2002a; Thompsons, 2001), where the bundles engage readers in textual acts (Hyland, 2002a) by referring to certain texts or formulae. This usage also demonstrates the explanatory nature of introductory textbooks by directing novices to review both text and diagrammatic information to understand the
content. It also offers novices models of connecting prior discourse with text/formulas for elaboration. Text referential bundles function as cohesive devices (Widdowson, 2007) by linking prior text with following text/formulas so that the new content is understood and "a context is created in the mind and signaled in the text in the process of its production" (p. 43). These bundles contribute to the construction of knowledge by facilitating the creation of a context in readers' minds.

Additionally, no multi-functional referential bundles are present in the ESPTC. Multi-function referential bundles in the EEITC are *at the end of the* and *the end of the*, which make both time and place references. However, in the ESPTC the two bundles only refer to time information, omitting the usage of place references commonly seen in EE introductory textbooks.

Not only does the ESPTC lack the two functional subcategories of referential bundles mentioned above, the existing subcategories also show proportional distribution differences from those in the EEITC. Tangible attribute specification bundles are the only type of referential bundles that are commonly used in both corpora with a percentage of 40%. Language use of other subcategories of referential bundles in ESP textbooks does not reflect target language use in introductory EE textbooks. Compared with the EEITC, the ESPTC overemphasizes three subcategories of lexical bundles: intangible attribute specification bundles, identification/focus referential bundles, and time reference bundles. The three types of bundles are used for description of abstract characteristics, identification of important concepts, and reference to particular time.

In addition to the dominant use of tangible attribute specification bundles (e.g., *the magnitude of the, the size of the*) for the description of concrete properties of following
noun phrases, writers of introductory EE textbooks use six other subcategories of lexical bundles which are sequenced in the following order of descending percentage (as we see in Figure 5.2): place referential bundles → intangible attribute specification bundles → quantity specification bundles → identification/focus bundles → text deixis bundles → multi-function bundles. The first five subcategories have a rather even distribution: the percentage of place referential bundles is 10.48%, with the next coming four categories having percentages descending from around 8% to 5% (see Table 5.5). Multi-function bundles take a rather small percentage of all identified bundles. The first five subcategories of lexical bundles respectively specify place and quantity, introduce important concepts, identify abstract attributes of entities, and make reference to a particular section of text.

In the ESPTC, in addition to intangible attribute specification bundles, when sequenced in an order of descending percentages, the six subcategories of referential bundles shown in Figure 5.2 are: intangible attribute specification bundles → identification/focus referential bundles → place referential bundles → time reference bundle → quantity specification bundles. Intangible attribute specification bundles and identification/focus referential bundles are the top two common types of bundles: the percentage of intangible attribute specification bundles in the ESPTC is around 15%, twice the percentage in the EEITC, and the percentage of identification/focus referential bundles in the ESPTC is around 12%, more than twice the percentage in the EEITC. The percentage of place referential bundles drops from its top place in the EEITC to third place in the ESPTC with 9.09%. Quantity specification bundles are the least commonly used in the ESPTC with 3.03%, compared with 6.67% in the EEITC. Time reference
bundles in the ESPTC account for 6.06%; however their presence in the EEITC is rare, limited to multi-function lexical bundles with a negligible 1.9%.

The mismatch of subcategory distribution between the EEITC and the ESPTC indicates that the two types of textbooks do not share the same focus on the presentation of referential information. In Electrical Engineering introductory textbooks, it is important to describe quantity characteristics, make textual reference and depict the physical connections among electrical components in circuit analysis, evidenced by the high percentages of referential bundles of quantity specification and text/place reference. However, those bundles are absent in ESP textbooks. Instead there are high percentages of intangible attribute specification bundles, identification/focus referential bundles and time referential bundles.

5.5.2 The comparison of stance bundles in the EEITC and the ESPTC

The number and the composition complexity of subcategories of stance bundles in the EEITC surpass those in the ESPTC. As shown in Table 5.5, stance bundles in the ESPTC account for only 9%, less than half of the percentage in the EEITC (20%). As we also see in Figure 5.2, compared with the only subcategory of stance bundles (impersonal epistemic stance bundles) in the ESPTC, the EEITC has a broader coverage of five subcategories of stance bundles. When sequenced in the order of descending percentages, those bundles are impersonal epistemic stance bundles (8.57%), attitudinal/modality stance bundles expressing authors' desire (3.81%), personal epistemic stance bundles (2.86%), obligation/directive stance bundles (2.86%), and attitudinal/modality stance bundles (0.95%) (See Figure 5.2).

Stance bundles in the EEITC show how textbook writers construct their roles as
discipline experts, and how they engage readers as participants in tasks or as observers of particular lines of argument. Textbook writers' preference for the impersonal epistemic stance bundles (can be obtained by to note that the, from the fact that) introduces shared evaluation toward particular propositions, keeps authors' personal opinions out the text, and thus helps to maintain objectivity and authority of the texts (Conrad, 1996; Reppen, 2004).

In addition to impersonal epistemic stance bundles, the other four subcategories of stance bundles account for a total percentage of 10.5% of all bundles in the EEITC. Commonly seen in the four subcategories of bundles (we note that the, we see that the, let us determine the, let us find it), the collective first person pronouns we and us signpost writers' invitation to readers as participants in the text. The frequent use of we/us in introductory textbooks functions as sort of a centrifugal force, serving to decrease the distance between writers and readers and inviting readers to become aware of the intellectual tasks or lines of arguments that writers think are important. Writers as experts use we and us to guide readers toward points deemed essential by the writers.

Electrical Engineering textbook authors commonly use stance bundles to exert directive forces (Hyland, 2002a) on readers in three ways, mainly at the cognitive level, 1) by emphasizing points that writers believe are important in order to direct readers' attention toward the findings or facts (to note that the, we see that the, we find that the), 2) by guiding readers to new lines of argument (we assume that the, from the fact that the), or 3) by directly instructing readers on the next step of analysis procedure (we will use the, we wish to determine, consider the circuit shown, etc.). The use of those bundles strengthens the authoritative effect of the text, enabling writers to display an integrated
and coherent picture of well-established knowledge from the perspective of discipline experts. This reader/author relationship in introductory textbooks helps the initiation of novices into the well-established norms of the discipline, and avoids presenting them with competing or conflicting dialogues among researchers.

In contrast, there is only one subcategory of stance bundles in the ESPTC, impersonal epistemic stance bundles (Figure 5.2). A total of four subcategories of stance bundles are absent in the ESPTC; these are personal epistemic stance bundles, attitudinal/modality stance bundles showing obligations and directives, attitudinal/modality stance bundles showing desire, and attitudinal/modality stance bundles showing intention.

The stance bundles in introductory textbooks are used by authors to position themselves as discipline experts and to form persuasive statements of discipline-specific knowledge. The bundles also clearly guide readers to the basic propositions and applications of the discipline. One of the implications that stance bundles have for ESP learners is: the lack of personal epistemic stance bundles and attitudinal/modality stance bundles in the ESPTC means that the learners lose opportunities to observe how these bundles direct the readers’ attention to points of knowledge or propositional facts (we see that the, we find that the). They would not be able to observe how those bundles explicitly instruct novices about writers’ purposes in the next step of analysis/argument procedure (we will use the, we wish to determine).

As well the learners will miss opportunities to observe the common use of we and us in Electrical Engineering introductory textbooks. Hyland (2002b) identified the gap of use regarding plural reference in research articles and student reports in Science and
Engineering. In per 10000 words, authors of research articles used an average of 31 plural pronouns for self-reference, 6 times higher than the number of plural pronouns in student writing. In interviews students confirmed that they were reluctant to take ownership of their own work. The use of the self-reference plural pronouns in Electrical Engineering introductory textbooks could be used as a model in the sense that it exhibits how textbook authors take up their roles as experts when telling novices the given knowledge and instructing them step by step about the argument/analysis procedure. If exposed to this aspect of language use in ESP materials, learners might be persuaded about the legitimate use of self-reference in writing of Electrical Engineering. They do not need to wait until later in their studies to realize that they are supposed to indicate ownership of their written reports.

This underrepresentation of target stance bundles in the ESP textbooks once again proves that there is a gap between ESP textbooks and Electrical Engineering introductory textbooks. The missing four subcategories of stance bundles in the ESP textbooks means that ESP readings do not offer enough language input in target areas, leading to the lack of opportunities for novices to observe how the interpersonal relationship between readers and writers is constructed by the four-word stance bundles in Electrical Engineering textbooks.

5.5.3 The comparison of discourse organizers in the EEITC and the ESPTC

Discourse organizing bundles in the EEITC and the ESPTC both have similar low percentages, indicating that not many four-word lexical bundles in both corpora serve as discourse organizers to link prior and coming discourses in original context.
5.6 ESP textbook treatment of target lexical bundles in the EEITC

Previous analysis indicated the gap in language use between the EE introductory textbooks and the ESP textbooks. It is worthwhile to examine if and how the ESP textbooks have emphasized lexical bundles and their pragmatic functions in learning activities. Table 5.6 and Table 5.7 show the results for the two ESP textbooks, *Oxford English for electronics* (Glendinning & McEwan, 1993) and *Oxford English for electrical and mechanical engineering* (Glendinning & Glendinning, 1995). The two tables show that a total of only five lexical bundles has been noted as key language points in the instruction for learning tasks. Consequently, those bundles are likely to be used in writing or speaking tasks, such as describing diagrams, components, and cause/effect relationship. For instance, on page 11 of *Oxford English for electronics* (Glendinning & McEwan, 1993), the task instruction for describing block diagrams and circuits suggests using *is connected to (the)* for the description of the links between building blocks, as in: *The turner is connected to the detector.* In the following writing task describing diagrams, students need to state how the building blocks are connected, *is connected to the* is very likely to be used in students' writing. In this way, we say *is connected to the* is featured for its pragmatic function in the ESP textbook.
Table 5.6 The treatment of lexical bundles in *Oxford English for electronics* (Glendinning & McEwan, 1993)

<table>
<thead>
<tr>
<th>Pragmatic functions</th>
<th><em>Oxford English for electronics</em> (Glendinning &amp; McEwan, 1993)</th>
<th>Task</th>
<th>Location</th>
<th>Lexical bundles included in the task instruction</th>
<th>Productive activities using those bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expressing Stances</strong></td>
<td></td>
<td>Certainty</td>
<td>p. 110</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Describing purpose</td>
<td>p. 124</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>Organizing discourses:</strong></td>
<td></td>
<td>Comparison &amp; contrast</td>
<td>p. 16, p. 96</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Topic elaboration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Referential functions:</strong></td>
<td></td>
<td>Describing block diagrams and circuits</td>
<td>p. 11-13</td>
<td>✓ <strong>is connected to the</strong></td>
<td>✓</td>
</tr>
<tr>
<td>Place referential function</td>
<td></td>
<td>Describing diagrams</td>
<td>p. 32, p. 54</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>Text deixis</strong></td>
<td></td>
<td>Linking facts and ideas</td>
<td>p. .24</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actions in sequence</td>
<td>p. 44</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Describing a process</td>
<td>p. 62</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>Identification/focus</strong></td>
<td></td>
<td>Describing components</td>
<td>p. 28</td>
<td>✓ <strong>is known as the</strong></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Tangible framing attributes</strong></td>
<td></td>
<td>Cause and effect</td>
<td>p. 92</td>
<td>✓ <strong>the effect of the</strong></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison &amp; contrast</td>
<td>p. 96</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Describing graphs</td>
<td>p. 97</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td><strong>Intangible framing attributes</strong></td>
<td></td>
<td>Cause and effect</td>
<td>p. 105</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change and result</td>
<td>p. 117</td>
<td>✓ <strong>a result of the</strong></td>
<td>✓</td>
</tr>
</tbody>
</table>
Table 5.7 The treatment of lexical bundles in *Oxford English for electrical and mechanical engineering* (Glendinning & Glendinning, 1995)

<table>
<thead>
<tr>
<th>Pragmatic function</th>
<th>Tasks</th>
<th>location</th>
<th>lexical bundles included in the task instruction</th>
<th>Productive activities using those bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressing Stances</td>
<td>Describing functions</td>
<td>p. 73</td>
<td>√ can be used to</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Describing functions</td>
<td>p. 119</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>Certainty</td>
<td>p. 144</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Organizing discourses:</td>
<td>Topic elaboration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conditions</td>
<td>p. 62</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>Describing contrast</td>
<td>p. 68</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Referential functions:</td>
<td>Place referential function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Connecting components</td>
<td>p. 40</td>
<td>√ is connected to the</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Describing and explaining</td>
<td>p. 57</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Text deixis</td>
<td>Making definitions</td>
<td>p. 24</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Identification/focus</td>
<td>Naming components</td>
<td>p. 40</td>
<td>√ is known as the</td>
<td>√</td>
</tr>
<tr>
<td>Tangible framing attributes</td>
<td>Cause and effect</td>
<td>p. 87, p. 93</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Intangible framing attributes</td>
<td>Cause and effect</td>
<td>p. 125</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
As we see in Table 5.6, four referential bundles have been featured in language instruction in *Oxford English for electronics* (Glendinning & McEwan, 1993); these are *is connected to the* (place referential bundle), *is known as (the)* (identification/focus referential bundle), *the effect of (the)* (intangible attribute specification bundle), and *a result of the* (intangible attribute specification bundle). Those bundles are respectively used in tasks for describing block diagrams and circuits, naming components, indicating effect and result. In Table 5.7, only three lexical bundles have been emphasized for their pragmatic functions in *Oxford English for electrical and mechanical engineering* (Glendinning & Glendinning, 1995). *Can be used to* (epistemic stance bundle expressing possibility), *is connected to (the)* (place referential bundle), and *is known as (the)* (identification/focus referential bundle) are suggested respectively to describe functions, elaborated component connection, and name components.

Table 5.6 and Table 5.7 show that learning tasks in the ESP textbooks lack coverage of lexical bundles. On the one hand, the majority of learning tasks have not addressed any lexical bundles in the instruction sections. Only a few tasks highlight the functions of a limited number of lexical bundles. As we can see in Table 5.6 and Table 5.7, those learning tasks range from indicating certainty, comparing and contrasting, linking facts and ideas, making definitions, to indicate cause and effect relationships. These learning tasks could draw useful materials from a pool of lexical bundles with a focus on their pragmatic functions in the EEITC. For instance, in *Oxford English for electronics* (Glendinning & McEwan, 1993), on page 97, the learning task is to describe graphs. Lexical bundles such as *the slope of the, a plot of the, a sketch of the* could be suggested in the instruction for the description of figure detail, as in:
The slope of this segment, which is equal to the slope of the tangent to the $i$-$v$ curve at the operating point $Q$, is equal to the small-signal conductance. (MEchaptper3)

Another example is the task of describing cause/effect relationships on page 92, the language use would be better presented if the language focus included a number of bundles, such as *is determined by*, *with the result that*, *is independent of the*, *a result of the*, *the effect of the*, which learners can use to describe cause and effect relationships.

A further observation made about the learning tasks in the two ESP textbooks is that some language usage in those tasks is not reflected in the EEITC. In Oxford *English for electrical and mechanical engineering* (Glendinning & Glendinning, 1995), the task on page 125 asks students to describe a cause/effect relationship, suggesting *bring about* to link cause and effect. I searched all form variations of *bring about*, including *brought about*, *bring about*, *bringing about*, and *brings about* in the EEITC but did not find any occurrences. In another task of certainty on page 144, the textbook authors group expressions of certainty together and put them in a list below:

<table>
<thead>
<tr>
<th>Certain</th>
<th>Fairly certain</th>
<th>Uncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Will</td>
<td></td>
</tr>
<tr>
<td></td>
<td>will probably</td>
<td>might</td>
</tr>
<tr>
<td></td>
<td>be likely+vb</td>
<td>may</td>
</tr>
<tr>
<td></td>
<td>be + probable + that</td>
<td>could</td>
</tr>
<tr>
<td>No</td>
<td>will not</td>
<td></td>
</tr>
<tr>
<td></td>
<td>be + unlikely to + vb</td>
<td>be possible that</td>
</tr>
</tbody>
</table>

Among those certainty expressions, besides single modal verbs, the chunk expressions do not occur in the EEITC, reflecting the fact that textbook writers' intuition of language use does not match target language use.
5.7 Critique of the functional taxonomy

In the following section I evaluate the functional taxonomy (Biber et al., 2002a) used as the analytic framework of this research. The main purpose of the critique is to demonstrate the difficulties of fitting lexical bundles perfectly into corresponding subcategories, as well as showing that the taxonomy may fail to capture multiple functions that some lexical bundles have in the original context. Biber et al. (1999) first developed an operational definition of lexical bundles. Later Biber (2003) came up with the initial functional taxonomy of lexical bundles, which are categorized according to their functions as referential bundles, stance bundles and discourse organizers. Other researchers applied this taxonomy for textual analysis in a broad range of registers across casual conversations, textbooks, service encounters, classroom talk and research articles (Biber et al., 2004a; Biber & Baribireli, 2007; Cortes, 2004).

The functional taxonomy has enabled researchers to locate language use differences among different registers and indicated that lexical bundles are useful indicators of these registers because they function as building blocks of discourse in meaning generation and become repertoires of particular registers. For example, previous multi-dimensional analysis of linguistic features (Conrad, 1996; Reppen, 2004) led to the conclusion that classroom teaching demonstrates similar grammatical patterns as conversation does because the distribution patterns of nouns, verbs, and personal pronouns are alike in both registers. That contradicts our common sense that the two registers are assumed to be different since classroom teaching has a clear informational focus and would be expected to share specific patterns with textbooks and other types of academic prose. Biber et al. (2004a) analyzed the distribution and discourse functions of lexical bundles across four
registers, conversation, classroom teaching, textbooks, and academic prose. The results show that language use in classroom-teaching is somewhere in the middle of conversation and academic written registers such as textbooks. Classroom teaching uses comparable amount of stance bundles and discourse organizers as conversations reflecting the pedagogical needs of engaging students and the real-time organizing needs during the lecture. At the same time, classroom teachers depend on a similar number of referential bundles for information construction, the number of which is comparable to those in textbooks and academic prose.

I used the functional taxonomy of lexical bundles (Biber et al., 2004a) for my own research. The taxonomy was a good tool to identify functional categories and subcategories of lexical bundles in Electrical Engineering introductory textbooks. However, three practical issues related with the taxonomy loomed in the process of analysis. The first issue has to do with the possibility of multiple functions of some lexical bundles, which makes it hard to place lexical bundles in one specific functional subcategory. The second issue is that occasionally the linguistic environments around the lexical bundles have an impact on their pragmatic functions. And the third issue is about the feasibility of the functional taxonomy to language use in specific registers.

After examining the concordance lines of lexical bundles in the EEITC, I followed Biber et al.’s taxonomy (2004a), reported the initial result of functional subcategories of lexical bundles to Dr. Ellen Cray and Dr. David Wood, two professors at my department and asked for their suggestions. There were occasions when we were hesitant to decide which functional subcategory a lexical bundle should belong to. A couple of controversial issues were identification/focus referential bundles, such as in the case of and is known as
The professors were right in the sense that the two bundles were used to direct readers' focus on the noun phrase after it at the sentential level. However, in addition to that, it seemed to me that the two bundles in many cases also function as discourse organizers when I observed the lexical bundle in a larger section of text. In the following sample of text, the second sentence indicates the possibility of using an equation to analyze the circuits. Then at the start of the third sentence, in the case of explicitly signals the beginning of an elaboration, where afterwards the author explains how to calculate the value of currents and voltages in two different types of circuits. This effect of elaboration could hardly be achieved if a simple preposition in took the place of in the case of in the original text.

In Chapter 2 we analyzed the simplest possible circuits, those containing only a single-node pair or a single loop. We found that these circuits can be completely analyzed via a single algebraic equation. In the case of a single-node-pair circuit (i.e., one containing two nodes, one of which is a reference node), once the node voltage is known, we can calculate all the currents. In a single-loop circuit, once the loop current is known, we can calculate all the voltages.

Although Biber et al. (2004a) said they considered the primary pragmatic functions when categorizing functions of lexical bundles, in a case like the above example, omitting other pragmatic functions and only sticking to the one never-could-be-wrong functional subcategory of lexical bundles might lead to the reduction of interpretation of pragmatic functions of lexical bundles in the original text. It is not clear how researchers are going to define which function of a lexical bundle is "primary" when they are in the face of a multi-functional lexical bundle.

The linguistic environments around some bundles can make the subcategory classification rather ambiguous. An example is to note that the, categorized as an impersonal epistemic stance bundle. In the following concordance lines of the lexical
bundle, the adjectives in front of to note that the, such as interesting and important, introduce textbook author’s personal assessment to the following propositions and thus weakens the overall epistemic impersonality of the lexical bundle.

N Concordance
1. book. In the meantime, it is interesting to note that the magnitude of currents
2. (...) Therefore, (...) it is interesting to note that the 3-O resistor never enters
3. in Fig. 13. 11. C=cos 0. It is important to note that the damping ratio and the

The third issue is about the broad range of academic spoken and written discourse on which Biber et al.’s taxonomy is based. For researchers concerned with language use in much narrower scopes, this taxonomy may not be suitable, and consequently the solution might be that they start with Biber’s functional taxonomy to have an initial look at how lexical bundles shape the target text. At the same time, they may need to take into consideration the context of specific registers, including the writer’s purposes, audience, and the link between text and context. That means researchers can carefully evaluate the functions of lexical bundles and in an inductive way set up a taxonomy that suits specific registers where those bundles are situated. Hyland (2008) reclassified lexical bundles in research-oriented, text-oriented, and participant-oriented groups. For the purpose of better functional description of those bundles, the reclassification of those bundles helps an understanding of the text construct in published research articles, Master theses, and PhD dissertations from four disciplines: Biology, Electrical Engineering, Applied Linguistics, and Business studies.

Hyland found that “writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers…” (Hyland, 2008, p. 20). Writers in Electrical Engineering have a preference for research-oriented bundles to describe research objects or context; they also heavily use bundles that make references
to text, which is commonly seen in both Electrical Engineering introductory textbooks (the concern of my research) and written products at higher academic levels (journal articles, Master theses, and Ph. D. dissertation), determined by the writers’ needs to constantly refer to visual and numerical information in their arguments. Those writers are also concerned with engaging their readers and directing them to a particular interpretation or perform an action. The substantial presence of engagement bundles shows that hard science disciplines “place considerable emphasis on precision, particularly to ensure the accurate understanding of procedures and results” (Hyland, 2008, p. 19). In contrast, 2/3 of the bundles in applied linguistics and business studies are text-oriented, reflecting the interpretative rather than empirical nature of soft knowledge fields as “producing discourse which often recasts knowledge as sympathetic understanding, promoting tolerance in readers through an ethical rather than cognitive progression” (Hyland, 2008, p. 16). It is consistent that knowledge in the social sciences “[...] is typically constructed as plausible reasoning rather than as nature speaking directly through experimental findings” (Hyland, 2008, p. 16). Fifty percent of the text-oriented bundles in the social science texts are framing signals, specifying logical conditions for argument by indicating connections among arguments, directing readers’ attention to aspects of a statement, or clarifying those aspects.

Previous researchers setting up such a functional taxonomy could organize lexical bundles according to their functions, which would give them a systematic picture of how the bundles construct text in specific registers and genres and carry out specific functions to achieve the communication goals. However, sometimes we have to admit this uniform categorization may not serve the richness of functions determined by the linguistic
environment and contextual factors around the lexical bundles, such as varying author/reader relationship and writing purposes in specific registers and genres. For language teachers who want to apply pragmatic functions of lexical bundles identified by previous research into teaching practice, it is worth the effort to examine the research results in context and take into consideration contextual factors when reevaluating the pragmatic functions of lexical bundles at the discourse level.

5.8 Summary

This corpus-based research investigates lexical bundles and their pragmatic functions in the EEITC and the ESPTC, examining if there is a gap between the introductory electrical engineering textbooks and the ESP textbooks. The research results demonstrate that lexical bundles, functioning as discourse builders, serve to construct the body of knowledge and establish the reader/author relationship in Electrical Engineering introductory books. The findings also tell us that there exists a gap in language use between the entry-level discipline textbooks and ESP textbooks.

To report on the research results, I began with the analysis of structural characteristics of lexical bundles in the EEITC and found a strong association between structural types and their pragmatic functions. The EEITC tends to use more phrasal lexical bundles than clausal lexical bundles, which helps to construct explicit and precise meaning in introductory textbooks. For instance, bundles made up of noun phrase with of-phrase fragment (e.g., the sum of the, the gain of the) indicate physical/abstract attribute of the following noun phrase.

Using Biber et al.'s (2004a) taxonomy as the analytic framework, I further analyzed and reported on functional characteristics of lexical bundles in the EEITC. Based on their
pragmatic functions, lexical bundles are classified into three main categories: referential bundles, discourse organizers, and stance bundles. Referential bundles in the EEITC are most widely used, identifying important entities for the purpose of focusing readers' attention on new concepts or important facts (e.g., *is known as the, referred to as the*). They can also specify quantity, physical, or abstract characteristics of following noun phrases (e.g., *is equal to the, the value of the, the effect of the*), or make references to place or text (e.g., *in the frequency domain, in this case the*). Stance bundles, less commonly used than referential bundles, express certainty toward the knowledge of the status of following propositions (e.g., *the fact that the*) or to mark textbook authors' attitude or logical judgement toward the content (e.g., *we wish to determine, can be used to*). Discourse organizers control the flow of the written discourse by connecting following and prior discourses (e.g., *on the other hand*).

In a similar way, I analyzed the lexical bundles and their pragmatic functions in the ESP textbooks. The results show that lexical bundles in the ESP textbooks have the same functions as their counterparts in introductory textbooks of Electrical Engineering, but that the total number of bundles in ESP textbooks is much lower than that in the EEITC. As a result, the ESP textbooks do not provide readers with enough input, and the readers may not be aware of how entry-level textbook writers of Electrical Engineering organize and demonstrate textual information.

Next I further examined the proportional distribution differences of lexical bundles between the two corpora, the ESPTC and the EEITC. The findings show that the ESPTC has a much narrower scope of functional subcategories than the EEITC; based on their pragmatic functions, the EEITC has lexical bundles classified into 13 subcategories,
while the ESPTC has only eight subcategories. The subcategory classification of stance bundles of the EEITC shows that introductory textbooks use more epistemic stance bundles than attitudinal/modality stance bundles for the purpose of expressing assessments rather than attitudes toward following propositions. With only one subcategory of stance bundles, the ESPTC shows a much narrower scope of language use and thus underrepresents some aspects of the target language. The examination of subcategory distribution of referential bundles shows that identification/focus referential bundles, intangible attribute framing bundles and time referential bundles are overemphasized in the ESP textbooks. While quantity specification bundles and place referential bundles are prominent in the EEITC, they are ignored in the ESPTC. Compared with target language use in Electrical Engineering introductory textbooks, the ESPTC underrepresents descriptions of quantity and place but overemphasizes information description that is not central in target language use. With the limited language use in the ESP textbooks, learners are likely to miss opportunities to understand how entry-level textbook authors of Electrical Engineering use lexical bundles to organize and demonstrate propositional facts and information, as well as how they maintain the authority of the textbooks and keep readers involved in the subject content.

Then I reported on how ESP textbooks treat lexical bundles in learning tasks. The investigation result shows that ESP textbooks do not adequately represent lexical bundles that could help learners accomplish learning tasks. In the two ESP textbooks, only five lexical bundles and their pragmatic functions are mentioned in the instruction for learning tasks. Some language points emphasized in the learning tasks of ESP textbooks do not occur in target language represented by the EEITC, suggesting that textbook writers
depend on their intuitions rather than evidence form target language use to compile their teaching materials.

In the last section, I critiqued the functional taxonomy proposed by Biber et al. (2004a). The uniform taxonomy may be limited in reflecting the functional complexity of lexical bundles in original contexts. Although the taxonomy helped the identification of lexical bundle functions in the EEITC, the data analysis process for this research revealed that three practical factors, the multi-function of lexical bundles, the linguistic environments and the contextual factors made it difficult to place lexical bundles in a single functional subcategory.
Corpora have practical implications for language learning and teaching because they make an important contribution to language description and thus provide "an empirical basis for checking our intuitions about language" (O'Keefe et al., 2007, p. 21). Previous studies show that corpora can be pedagogical resources and tools that language learners can consult about language use in areas of their learning interests (McCarthy, 1999; Yoon & Hirvela, 2004). Corpora are as well databases from which language experts can generate teaching materials (McCarthy, 1999; O'Keefe et al., 2007). There are cases where corpus-based materials have been incorporated into the teaching syllabi for learners with different proficiency levels (Cortes, 2004; Lee & Swales, 2006; Yoon & Hirvela, 2004).

McCarthy (1999) proposed two possible methods of using corpus data in language teaching, which he termed a corpus-driven approach and a corpus-informed approach. "A corpus-driven approach is absolutely faithful to the evidence of the corpus; a corpus-informed approach takes insight from the corpus, but filters that insight through commonsense language teaching practices" (McCarthy, 1999, p. 3). A corpus-informed approach is better suited for elementary-level learners since teachers have to create short, comprehensible materials and encourage students to use frequent phrases such as have dinner/lunch in natural conversations (McCarthy, 1999). The ESL textbook series of Touchstone (McCarthy, McCarten & Sandiford, 2004), which uses a corpus-informed approach, is designed for beginning- to intermediate-level adult or young adult learners with the aim of helping learners develop conversation managements strategies such as how to start and end conversations. Touchstone uses data from the spoken subcorpus of
the 100-million-word Cambridge International Corpus (CIC) to generate language exercises for skills development based on language use in real conversations.

By directly giving learners examples from original texts, a corpus-driven approach appears to fit in the learning needs of advanced level learners. Yoon and Hirvela (2004) used the Collins COBUILD Dictionary Corpus, which has millions of entries from the Bank of English, in two ESL classes at the intermediate and advanced levels. Inductive learning was introduced to ESL learners: authentic texts were given to learners, who used computer software to obtain concordance lines with particular words to expand their understanding of the functions of the words in original contexts. Cortes (2004) used specialized corpora as supplementary materials in an intensive writing course for third- and fourth-year undergraduate students majoring in history, who have English as their first language. The instructor set up a specialized corpus consisting of history journal articles required by an intensive writing course. She identified the lexical bundles and their pragmatic functions in the corpus and provided formal instruction and exercises of the lexical bundles in the writing class. Lee and Swales (2006) designed a corpus-based experimental course for 4 NNS doctoral students, encouraging learners to observe and discover how particular lexico-grammatical features were used in research articles from their disciplines.

An important issue is what kind of corpus data language learners and teachers should focus on. One possibility is to set the focus on the Academic Word List (AWL), consisting of 570 word families, each of which “was defined as a stem plus all closely related affixed form” (Coxhead, 2000). For example, concept is a word family consisting of word members such as conception, concepts, conceptual, conceptualization, and so on.
The AWL was generated from a 3.5-million-word corpus consisting of academic written texts from 28 subject areas in four disciplines: arts, commerce, law and science. All the word families in the AWL were chosen according to the three criteria: 1) the word family is not included in the General Service List (GSL) (West, 1953), 2) each member of the word family had to occur for at least 100 times in the corpus, and 3) each member of the word family had to occur for at least 10 times in all four disciplines and in at least 15 subject areas. Coxhead (2000) developed the AWL with the assumption that there is a core academic vocabulary in university settings evenly spread across texts from a variety of subject areas. Building this repertoire of academic words was expected to help learners manage their discipline learning more efficiently.

However, the AWL has been criticized for being too general. Hyland and Tse (2007) argued that the reliance on this general academic word list would not help students improve their discipline-specific language competencies. They further suggested that the teaching focus can be shifted to lexical bundles since they depict language use in specific disciplines. The focus of EAP teaching on target discipline language is not something new since analyses of learners' academic tasks in discipline studies (Carson, 2001) indicated that language teachers need to identify target language use to help learners attain the literacy skills required by the target disciplines. Depending on their approach to teaching, some Language for Specific Purposes (LSP) teachers suggest a focus on specific language use in language teaching since it was discovered that A-level students in a general language learning program were not competent in expressing their ideas in subject-specific communication situations (Dlaska, 1999). After analyzing learners' needs in an Engineering-specific course taught in German, Dlaska suggested including subject-
specific language use in activities and exercises with an emphasis on developing the language skills needed in specific workplaces and course studies; doing this will help bridge the gap between general and specific language abilities and allow learners to develop their specific language skills.

Benson (1993) recognized the benefits of having corpus materials from the target language and suggested that “familiarity of content is recognized as an important starting point for new learning, and the presentation of new items within accessible contexts should have effective effects on the learning process” (p. 120). Learners tend to have high motivation when choosing their own texts, setting up specialized corpora tailored to their disciplines and exploring language use relevant to their academic interests (Aston, 2002; Jabbour, 2001; Lee & Swales, 2006).

For learners with discipline-specific learning needs, lexical bundles, defined by their high frequencies in target language use and pragmatic functions, indicate the strong associations between discipline specificity and multi-word clusters (Biber et al., 2004a; Cortes, 2004; Hyland & Tse, 2007) and thus constitute good EAP/ESP teaching materials. My research suggests that lexical bundles play specific roles in structuring the texts and maintaining the author/reader relationship in introductory EE textbooks. Introductory textbooks commonly use referential and stance bundles for two main purposes: 1) to present propositional facts and construct the body of knowledge; and 2) to explicitly guide readers through intellectual tasks, directing their attention to lines of argument deemed as essential by writers, or/and telling readers what the learning tasks are. Language teachers can use knowledge of lexical bundles and their pragmatic functions to help students increase their awareness of language use in the discipline. At the same time,
for discipline learners, becoming familiar with the pragmatic functions of the lexical bundles in the entry-level textbooks will help them develop their discipline-specific reading and writing skills since they have opportunities to observe how these bundles help structure the texts.

Referential bundles in the EEITC specify quantity (the sum of the) and concrete/abstract characteristics (the magnitude of the, in terms of the) and make reference to text/place (is of the form, is connected to the). These bundles describe details relevant to research process and circuit analysis. In addition, some bundles indicate a logical relationship (is determined by the, with the result that) and thus link texts in a coherent way. Understanding the functions of these bundles will help students deal with tasks such as explaining diagrams and reporting their lab procedures with ease. Cortes’ study (2004) pointed out that students at third- and fourth-year of their undergraduate programs used fewer referential bundles than academics. For example, instead of using temporal referential bundles (i.e., at the end of, at the beginning of), one student repeatedly mentioned the same exact date in a short stretch of text. Although it is not expected that discipline novices will write like academics, exposing students to referential bundles and their pragmatic functions in EAP classrooms could help the novices become aware of referential bundles in their disciplines and increase their awareness of the use of those bundles through formal instructions and exercises.

Stance bundles in the Electrical Engineering introductory textbooks help textbook authors construct their role as discipline experts and engage readers as participants in intellectual tasks or as active observers of lines of arguments. It is important for students to understand how textbook authors use stance bundles to construct a coherent picture of
foundational knowledge. When introduced to stance bundles and their pragmatic functions in EAP classrooms, students should be allowed time to observe the text samples and discuss the author/reader relationships reflected by these bundles. Discussion like this during the class could make students aware of the role that stance bundles play in shaping the reader/writer relationships in textbooks and lead students to pay attention to key knowledge points recognized by textbook writers. Language teachers need to guide students to analyze and discuss which bundles are applicable to their discipline writing tasks, such as lab reports, and let students critically think about why and how they are going to use those bundles in their writing.

Although there have been many studies demonstrating what type of corpus data can be used in language teaching and how to incorporate corpus-based data in language classroom, there have been little conclusive research that reported the effectiveness of corpus-based EAP/ESP courses in terms of increasing learners’ reading and writing competencies. As indicated in Cortes’ (2004) study, the comparison of how students used lexical bundles in their writing before and after the instruction showed that the students did not make significant improvement. Cortes suggested that the lack of progress in terms of lexical bundle use could be attributed to the fact that the acquisition of lexical bundles in the discipline is a long process and thus learners need more exposure to lexical bundles as it takes time for these novices to improve their writing.

However, students in EAP/ESL programs covered in three previous studies (Cortes, 2004; Lee & Swales, 2006; Yoon & Hirvela, 2004) expressed overall satisfaction with their corpus-based learning experience. Students reported using corpora as a resource to check and confirm their writing (Lee & Swales, 2006; Yoon & Hirvela, 2004). History
students reported an increased awareness of lexical bundles in both reading and writing (Cortes, 2004). Some history students reported having developed new learning strategies, such as having a list of bundles and their functions as reference resources in writing tasks (Cortes, 2004). In Lee and Swales' (2004) class, students were asked to use their own discipline knowledge to set up small-scale corpora, consisting of the students' own writing and expert writing. Through corpus analysis, advanced learners observed how particular lexico-grammatical features were used in norms of particular disciplines. Biostatistics PhD students observed that dropping off the definite article before noun phrase as in administration of the drug is common in medical research articles "due to the increased routinization and standardization of medical procedure" (p. 66). In additions, learners reported increased confidence in writing: they became more independent and confident writers relying less on native speakers regarding language issues. These learners as well reported increased engagement with text when they had access to language use in target disciplines since “the exemplification was often much closer to their contextual and textual circumstances” (p. 71).

As such, we can predict that introducing lexical bundles and their pragmatic functions in EAP/ESP programs may guide novices to think beyond the rhetorical communication between textbook writers and readers. Further benefits may include improved confidence in writing and better learning strategies. However, the pedagogical implication of this research is weakened by the small size of the comparison corpus. The two ESP textbooks included in the ESPTC are the only textbooks available in the market. In the future with more ESP textbooks available, the conclusion I make about the gap of language use between the EEITC and the ESPTC might need to be modified.
It should also be mentioned that this research seems to benefit only EE students in the ESP courses since the focus of the research is language use in entry-level textbooks of Electrical Engineering. Since students in EAP programs are from a diversity of disciplines, computer, business, biology, and so on, it becomes essential to set up corpora of introductory textbooks from a variety of disciplines. With the establishment of these corpora, we can obtain a comprehensive picture of lexical bundles and descriptions of their pragmatic functions in multi-disciplines.
Chapter 7 Conclusion

Lexical bundles identified in the EEITC can be used in EAP/ESP pedagogy because they consistently serve important pragmatic functions of presenting accurate propositional facts and expressing attitudes/assessment of these facts, reflecting the prefabricated status of lexical bundles. As a consequence, the lexical bundles construct the repertoire of EE introductory textbooks by framing the body of knowledge and indicating the reader/author relationships. In addition to the fact that discipline-specific corpora allows the study of language use in particular materials, these corpora can also be used to evaluate the authenticity of text used in ESP/EAP textbooks by using lexical bundles with pragmatic functions as the yardstick.

There has long been the debate on how to evaluate the authenticity of texts used in EAP/ESP materials. Breen (1985) pointed out that the meaning of authenticity be defined four dimensions in terms of the texts, the participants, the tasks and the social situation:

1. Authenticity of the texts which we may use as input data for our learners.
2. Authenticity of the learners’ own interpretation of such texts.
3. Authenticity of the tasks conducive to language learning.
4. Authenticity of the actual social situation of the language classroom.
   (Breen, 1985, p. 61)

An authentic text is often defined as one written for native speakers for real communicative purposes, “a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort” (Morrow, 1977, p.13). It is interesting to see that almost all the prefaces or the promotion websites of EAP/ESP textbooks emphasize that the texts are authentic. The authenticity of EAP/ESP materials can not be the same as an authentic text since the pedagogical materials adapt authentic texts from different sources (course syllabi, course requirements, stretches of texts from textbooks, and so on) in target disciplines and arrange the texts according to
the themes/topics arranged by pedagogical textbook writers. It is not known how authentic these materials are when compared with real language use. An alternative way to judge the authenticity of a text is by determining if there is a match between the lexical bundles with the pragmatic functions in target language and in pedagogical materials. Researchers can examine if the same lexical bundles identified in target language use are present in pedagogical materials and determine whether the functions of lexical bundles in the pedagogical textbooks are the same as in target texts.

This research demonstrates how to use lexical bundles to judge the authenticity of ESP pedagogical materials. In the EEITC a total of 13 subcategories of lexical bundles were identified, only eight of which are present in the ESPTC. The two missing subcategories of referential bundles are text referential bundles and multi-functional referential bundles. Text referential bundles in the EEITC reflect interaction with readers by guiding their attention to certain texts or diagrammatic information to understand the content. Multi-functional referential bundles in the EEITC make both time and place references while the usage of place references is ignored in the ESPTC. Out of the total of five subcategories of stance bundles in the EEITC, only one subcategory is present in the ESPTC. The lack of four subcategories of stance bundles in the ESPTC means that pedagogical textbooks do not offer enough language input for learners to observe the reader/author relationship which directs readers’ attention to important propositional facts or arguments in the textbooks. The proportional distribution differences of lexical bundles in the two corpora indicate that the majority of the subcategories in the EEITC are either underrepresented or overemphasized in the ESPTC. For example, the ESPTC overemphasizes descriptions of time references whose usage does not take a central
position in the EEITC.

The gap of language use between the EEITC and the ESPTC indicates that the ESPTC not only lacks target language use, but also misrepresents language use in introductory textbooks of Electrical Engineering. With the limited and distorted representation of lexical bundles in the ESP materials, the learners would not observe how introductory textbook writers present the basic concepts in Electrical Engineering and how they engage readers as participants in intellectual tasks.

What needs to be done in the future is in-depth investigation of the relationship between lexical bundles and the context. For instance, for Electrical Engineering students, their learning experience is multi-faceted: they have introductory textbooks in a number of disciplines as primary texts. At the same time they attend lectures given by professors, participate in lab sessions and deal with all sorts of assignments and quizzes. The analysis of lexical bundles in textbooks can tell us which points are deemed as important by textbook writers. But are those points also emphasized in professors’ lectures, assignments, labs and quizzes? For example, Electrical Engineering students can not directly copy the lexical bundles and their usage identified in this research for their lab reports because the context has been changed. The author/reader relationships in EE introductory textbooks emphasize the authority of the text since the textbooks represent valid shared knowledge in the disciplines. However in the case of writing lab reports, the author/reader relationships are different from those in discipline textbooks; the students have the instructors as their readers, who already have the expectations of what the lab reports should look like. The students need to understand the contextual differences before they apply their knowledge of lexical bundles into discipline writing. This also
implies that future work needs to be done to identify the possible links between textual analysis and the real learning context students are immersed in.
Appendix A: Questionnaire for evaluation of introductory textbooks in electrical engineering

For my thesis in Applied Language Studies, I am exploring the linguistic features of introductory Electrical Engineering (EE) textbooks for undergraduate students, and plan to compare those features with the linguistic features of English for Specific Purposes (ESP) textbooks. Based on my initial study of the EE websites at three Canadian universities, Carleton University, University of Toronto and University of Waterloo, I have developed a list of 4 textbooks for EE introductory courses. I would appreciate your evaluation of these textbooks since you are the expert in the field. This questionnaire will take approximately half an hour to complete.


Are you familiar with this book?

Have you used it before?

The table of contents has been photocopied. Can you please underline the sections that you think must be covered in an introductory EE course?

Does the book cover the basic concepts in Electrical Engineering?

How representative is this textbook?

<table>
<thead>
<tr>
<th>not representative at all</th>
<th>very representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>


Are you familiar with this book?
Have you used it before?

The table of contents has been photocopied. Can you please underline the sections that you think must be covered in an introductory EE course?

Does the book cover the basic concepts in Electrical Engineering?

How representative is this textbook?

<table>
<thead>
<tr>
<th>not representative at all</th>
<th>very representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6</td>
<td></td>
</tr>
</tbody>
</table>


Are you using this book now?

The table of contents has been photocopied. Can you please underline the sections that you think must be covered in an introductory EE course?

Does the book cover the basic concepts in Electrical Engineering?

How representative is this textbook?

<table>
<thead>
<tr>
<th>not representative at all</th>
<th>very representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6</td>
<td></td>
</tr>
</tbody>
</table>


Are you familiar with this book?

Have you used it before?
The table of contents has been photocopied. Can you please underline the sections that you think must be covered in an introductory EE course?

Does this book cover issues in electricity and electromagnetism?

How representative is this textbook?

<table>
<thead>
<tr>
<th>not representative at all</th>
<th>very representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  2  3  4  5  6</td>
<td></td>
</tr>
</tbody>
</table>

Besides the four textbooks listed above, do you have some other good introductory textbooks to recommend? If yes, please list those books.
Appendix B: Breakdown of the Electrical Engineering Introductory Textbook Corpus (EEITC) and the ESP Textbook Corpus (ESPTC)

1. Breakdown of the Electrical Engineering Introductory Textbook Corpus (EEITC)

<table>
<thead>
<tr>
<th>Topics</th>
<th>File names</th>
<th>Number of running words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic concepts in circuit analysis</td>
<td>cachapter01.txt</td>
<td>4,157</td>
</tr>
<tr>
<td>Resistive circuits</td>
<td>cachapter02.txt</td>
<td>10,457</td>
</tr>
<tr>
<td>Nodal and loop analysis techniques</td>
<td>cachapter03.txt</td>
<td>9,645</td>
</tr>
<tr>
<td>Additional analysis techniques</td>
<td>cachapter04.txt</td>
<td>9,479</td>
</tr>
<tr>
<td>Capacitance and inductance</td>
<td>cachapter05.txt</td>
<td>3,825</td>
</tr>
<tr>
<td>First- and second-order transient circuits</td>
<td>cachapter06.txt</td>
<td>11,672</td>
</tr>
<tr>
<td>AC steady-state analysis</td>
<td>cachapter07.txt</td>
<td>9,636</td>
</tr>
<tr>
<td>Magnetically coupled networks</td>
<td>cachapter08.txt</td>
<td>4,736</td>
</tr>
<tr>
<td>Steady-state power analysis</td>
<td>cachapter09.txt</td>
<td>9,503</td>
</tr>
<tr>
<td>Ployphase circuits</td>
<td>cachapter10.txt</td>
<td>4,937</td>
</tr>
<tr>
<td>Variable frequency network performance</td>
<td>cachapter11.txt</td>
<td>11,808</td>
</tr>
<tr>
<td>The Laplace transform</td>
<td>cachapter12.txt</td>
<td>3,222</td>
</tr>
<tr>
<td>Application of the laplace transform to circuit analysis</td>
<td>cachapter13.txt</td>
<td>6,987</td>
</tr>
<tr>
<td>Fourier analysis techniques</td>
<td>cachapter14.txt</td>
<td>5,534</td>
</tr>
<tr>
<td>Two-port networks</td>
<td>cachapter15.txt</td>
<td>3280</td>
</tr>
<tr>
<td>Basic semiconductor electronic circuits</td>
<td>cachapter16.txt</td>
<td>4,604</td>
</tr>
<tr>
<td>Lab manual</td>
<td>lab_manual.txt</td>
<td>7,257</td>
</tr>
<tr>
<td>Introduction to electronics</td>
<td>mechapter1.txt</td>
<td>14,674</td>
</tr>
<tr>
<td>Operational amplifiers</td>
<td>mechapter2.txt</td>
<td>18,284</td>
</tr>
<tr>
<td>Diodes</td>
<td>mechapter3.txt</td>
<td>24,251</td>
</tr>
<tr>
<td>MOS field-effect transistors</td>
<td>mechapter4.txt</td>
<td>32,518</td>
</tr>
<tr>
<td>Bipolar junction transistors</td>
<td>mechapter5.txt</td>
<td>36,880</td>
</tr>
<tr>
<td><strong>Total running words</strong></td>
<td></td>
<td><strong>247,346</strong></td>
</tr>
</tbody>
</table>

2. Breakdown of the ESP Textbook Corpus (ESPTC)

<table>
<thead>
<tr>
<th>Textbooks</th>
<th>File names</th>
<th>Number of running words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English for electrical and mechanical engineering</td>
<td>eeme.txt</td>
<td>46,190</td>
</tr>
<tr>
<td>(Glendinning &amp; Glendinning, 1995)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English for electronics</td>
<td>ee.txt</td>
<td>53,584</td>
</tr>
<tr>
<td>(Glendinning &amp; McEwan, 1993)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total running words</strong></td>
<td></td>
<td><strong>99,774</strong></td>
</tr>
</tbody>
</table>
Appendix C: Encoding the functional taxonomy (Biber et al., 2004a)

The functional taxonomy (Biber et al., 2004a) can be encoded at four levels:

1) At the first level, functions of lexical bundles are categorized into three types: stance expression bundles, discourse organizers, and referential bundles.

2) At the second level, the three types of functions can be further divided into subcategories. For instance, stance bundles consist of two sub-categories: 1) epistemic stance, and 2) attitudinal/modality stance. Discourse organizing bundles have two sub-categories: topic introduction/focus and topic elaboration/clarification. Referential bundles have four sub-categories, identification/focus, imprecision, specification of attributes and time/place/text reference.

3) At the third level, the subcategories can be further divided into sub-sub-categories. For example, the attitudinal/modality stance bundles are further divided into stances that express 1) desire, 2) obligation/directive, 3) intention/prediction, 4) ability.

4) At the forth level, lexical bundles are distinguished for their personal/impersonal form.

According to the stratified structure of the functional taxonomy as described above, I propose a coding format for this taxonomy, which consists of four slots, A, B, C, and D: ABCD. Those slots, A, B, C, and D, respectively represent the first-, second-, third-, and forth-level of the function taxonomy. Table C.1 lists the values of the slots.

Table C.1 the coding values for the functional taxonomy of lexical bundles

<table>
<thead>
<tr>
<th>Slot position</th>
<th>Value setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0-the default value;</td>
</tr>
<tr>
<td></td>
<td>1-stance expressions;</td>
</tr>
<tr>
<td></td>
<td>2-discourse organizers;</td>
</tr>
<tr>
<td></td>
<td>3-referential expressions.</td>
</tr>
<tr>
<td>B</td>
<td>0-the default value;</td>
</tr>
<tr>
<td></td>
<td>1-epistemic stance;</td>
</tr>
<tr>
<td></td>
<td>2-attitudinal/modality stance;</td>
</tr>
<tr>
<td></td>
<td>3-topic introduction/focus;</td>
</tr>
<tr>
<td></td>
<td>4-topic elaboration/clarification;</td>
</tr>
<tr>
<td></td>
<td>5-identification/focus;</td>
</tr>
<tr>
<td></td>
<td>6-imprecision;</td>
</tr>
<tr>
<td></td>
<td>7-specification of attributes</td>
</tr>
<tr>
<td></td>
<td>8-time/place/text reference</td>
</tr>
<tr>
<td>C</td>
<td>00-the default value;</td>
</tr>
<tr>
<td>D</td>
<td>01-stance expression bundles</td>
</tr>
<tr>
<td></td>
<td>02-discourse organizers</td>
</tr>
<tr>
<td></td>
<td>03-referential expressions</td>
</tr>
<tr>
<td></td>
<td>11-epistemic stance</td>
</tr>
<tr>
<td></td>
<td>12-attitudinal/modality stance</td>
</tr>
<tr>
<td></td>
<td>13-topic introduction/focus</td>
</tr>
<tr>
<td></td>
<td>14-topic elaboration/clarification</td>
</tr>
<tr>
<td></td>
<td>15-identification/focus</td>
</tr>
<tr>
<td></td>
<td>16-imprecision</td>
</tr>
<tr>
<td></td>
<td>17-specification of attributes</td>
</tr>
<tr>
<td></td>
<td>18-time/place/text reference</td>
</tr>
<tr>
<td>Slot position</td>
<td>Value setting</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>01-desire;</td>
</tr>
<tr>
<td></td>
<td>02-obligation/directive;</td>
</tr>
<tr>
<td></td>
<td>03-intention/prediction;</td>
</tr>
<tr>
<td></td>
<td>04- ability;</td>
</tr>
<tr>
<td></td>
<td>05- quantity specification</td>
</tr>
<tr>
<td></td>
<td>06- tangible framing attributes</td>
</tr>
<tr>
<td></td>
<td>07- intangible framing attributes</td>
</tr>
<tr>
<td></td>
<td>08- place reference</td>
</tr>
<tr>
<td></td>
<td>09-time reference</td>
</tr>
<tr>
<td></td>
<td>10-text deixis</td>
</tr>
<tr>
<td></td>
<td>11-multi-functional reference</td>
</tr>
</tbody>
</table>

D

0-the default value
1-personal
2-impersonal

Note: not all subcategories at the second level have a third- or fourth-level sub-division. In this case, the value for slots C and D are set as the default value, 00 or 0.

For instance, according to Biber et al. (2004a), the lexical bundle *at the same time* is a time reference bundle, accordingly, the code for this bundle is 38090. In this way, the newly identified lexical bundles can be organized in a convenient way for further quantitative analysis.
References


Jarvis, H. (2004). Investigating the classroom applications of computers on EFL courses
at higher education institutions in UK. *Journal of English for Academic Purposes*, 2, 111-137.


Noguchi, J., Orr, T., & Tono, Y. (2006). Using a dedicated corpus to identify features of professional English usage: what do "we" do in science journal articles. In A. Wilson, D. Archer, & P. Rayson (Eds.), *Corpus linguistics around the world* (pp. 155-166). Amsterdam: Rodopi, B. V.


Williams, G. (2002). In search of representativity in specialised corpora. *International

