How General is the Vocabulary in a General English Language Textbook?

Hedy McGarrell*
Brock University, Canada

Nga Tuiet Zyong Nguien
Brock University, Canada

Abstract

This study reports on the analysis of a widely used "General English" textbook to explore the relationship between lexical bundles included in the text and lexical bundles identified in relevant corpora to determine the appropriateness of the text's vocabulary in relation to its stated objective. Appropriateness is examined through the analysis of usefulness and functions, and the relationship between the two, by comparing the usefulness scores of various functions. The results show a relatively low level of usefulness of the lexical bundles in the textbook, meaning low frequency and small range of usage for the analysed items. The function analysis showed that textbook includes all the functions. The most common function was referential, followed by stance, special conversational, and discourse organizing functions. The current study offers an initial step for future research of lexical bundles and their functions, and usefulness in language teaching and teaching materials development; specifically, it suggests a possible methodology to be used in such research. Moreover, the results of this study provide insights into the value of lexical bundles in teaching and the development of teaching materials.

Keywords: multiword constructions, corpus research, English textbooks, textbook design

Introduction

Textbooks in second or foreign language learning programs are typically the main or even sole source of vocabulary input for learners in classroom contexts and thus have a major impact on the vocabulary learners encounter (McDonough, Shaw & Mashuhara, 2012; Neary-Sundquist, 2015). However, researchers, teachers and their learners have repeatedly questioned whether the language included in these textbooks reflects the language used in real life situations (Biber & Reppen, 2002). Increasingly, studies show that language from language in use, as captured in corpora, and language teaching materials are often at odds (Gabrielatos, 2006; Koprowski, 2005; Meunier & Gouverneur, 2009; Shortall, 2007). The availability of suitable techniques for analyses may have prevented more extensive research in the past, but increasingly corpus linguistics, with its large data banks of naturally occurring text, provides a promising way of investigating such questions. One such technique involves the analysis of multi-word combinations that co-occur repeatedly within the same register in native speaker usage but are not typically fixed nor structurally or semantically complete (Csomay, 2013). Conrad & Biber (2004) show that approximately 20 percent of the words (tokens) in written academic

**Tel. 1 905 688 5550, ext. 3757. Email: hmcgarrell@BrockU.CA, Department of Applied Linguistics, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, ON, Canada, L2S 3A1*

texts occur within three or four word groups of such multiword combinations, which makes them an important focus for further investigation as they have the potential to support subsequent language learning (L2). The question then is whether textbooks include the multiword combinations typical for the stated purpose of a given L2 textbook, in this study specifically in an English as a Subsequent Language (ESL) for 'general English' textbook.

Researchers interested in the relationship between actual language use and the language presented in textbooks and other teaching materials have pointed out how corpora can be used to answer questions about variations in language across registers, lexico-grammatical associations, discourse variables, language acquisition. McCarten (2010) argues that corpora provide sources for textbook developers to compile systematic lexico-grammatical syllabi based on authentic texts. Research that focuses on the study of frequently occurring multiword combinations (Biber, Johansson, Leech, Conrad, & Finegan, 1999: Cortes, 2004; Sinclair, 1991), often referred to as lexical bundles in recent work, encountered in the texts represented in corpora is particularly relevant for the current study. These vocabulary focused studies have investigated multiword combinations and their structural and functional characteristics in various disciplines and registers such as academic prose, conversation, classroom discourse, demonstrating their importance in diverse naturally occurring, which in turn makes them an important component for learners' vocabulary development. Increasingly, research findings support arguments in favour of including lexical bundles in textbooks and other pedagogic materials.

Considering the discrepancies pointed out in recent research between the vocabulary included in textbooks and its occurrence in authentic language illustrated in corpora, the current study presents an analysis of a widely used General English textbook, *English File intermediate (student's book)* by Latham-Koenig and Oxenden (2013). The focus of analysis is on lexical bundles and seeks to determine whether the lexical bundles included in the textbook represent broader usage as indicated in corpora. Given the research that demonstrates the frequency of lexical bundle in a broad range of naturally occurring texts, with different bundles and functions depending on register, such bundles have been shown to be present in textbooks Biber, 2006; Hyland, 2005), but with important disciplinary differences. The underlying question that motivates the current study is whether the lexical bundles in language learning texts reflect those bundles researchers have identified as particularly frequent in language situations relevant to the stated purpose of such a textbook.

The literature review below serves to define key terms used in the study and to provide background from recent, directly related, studies. The role of corpora in materials development is discussed first. It is followed by a definition of lexical bundles and a discussion of their role in natural language. The section concludes with a definition of function in relation to lexical bundles.

Corpora and Materials Development

Corpora represent typically large, principled collections of language use in different naturally occurring spoken or written contexts and registers. They provide information on various aspects of language including distribution, frequency, concordances, collocations and aspects of grammar that are often omitted in learner grammars (Carter & McCarthy, 1995). Such a resource would seem ideal for L2 learning and teaching purposes and has had a strong impact on dictionaries (McCarthy, 2008) and is beginning to influence the development of reference and learner grammars (Lee & McGarrell, 2011; Meunier & Gouverneur, 2009). Corpora have yet to be widely exploited for the creation of learning and teaching materials, but researchers are increasingly focusing on potential areas that could support L2 learning.

Past research has compared corpus information and teaching materials in terms of various language learning aspects, including vocabulary and grammar, in different contexts, such as written and spoken English for general communication as well as professional and academic purposes. For example, Biber and Reppen (2002) analysed the relationship between the information presented in ESL-EFL text books and findings from corpus linguistics. The researchers surveyed six widely used traditional grammar texts and analysed them based on three specific areas: grammar features included (types of adjectives), order of grammar topics (simple and progressive aspect), and vocabulary used to present these areas. Their analyses showed that the relevant materials in the selected textbooks did not reflect the frequency data in corpora, that the sequence of grammar points presented was not grounded in actual use and that there was little consistency in selecting vocabulary. Biber and Reppen concluded that the textbooks analysed were developed based on instinct rather than language in use. Considering their findings, they argued that frequency information should be a key factor in materials development choices, as frequently occurring vocabulary and grammar features are likely more useful for learners. A replication of Biber and Reppen's study, with more recent editions of either the same or comparable grammar texts (Lee & McGarrell, 2011), suggests increasing awareness of the existing gap between materials and language in use. These more recent editions were either corpus-based or corpus-informed (McCarthy, 2008), thus were expected to reflect a more authentic description of the specific areas being analysed. Lee & McGarrell's analysis showed that the more recent texts tended to represent corpus findings more closely, but still left considerable room for improvement in terms of reflecting actual language use. Similarly, Cheng and Warren (2007) examined 15 EFL textbooks endorsed by the Hong Kong Education and Manpower Bureau and compared them to the findings generated from the Hong Kong Corpus of Spoken English (HKCSE). Analyses showed that the vocabulary and language forms introduced in the textbooks were low-frequency items associated primarily with academic registers, thus more complex and explicit than the forms found in the HKCSE. Finally, two studies investigated specifically the use of multiword combinations. Koprowski (2005) investigated the usefulness of lexical phrases, in terms of their frequency and range, in contemporary textbooks compared to corpus data. The analysis involved 822 items based on their usefulness scores generated from the frequency and range data in the COBUILD Bank of English, a computerized corpus containing 17 different British and American native-speaker subcorpora (e.g., newspapers, magazines, books, radio, informal conversations). Findings showed that one third of lexical phrases used in the textbooks analysed were low-frequency items, thus unlikely useful in most real communication. Koprowski questioned the validity of lexical selections, suggesting that they were likely again based on the textbook writers' intuitions and experience rather than real language.

The studies discussed in this section show that despite the availability of corpora and corpus research findings, materials writers rely heavily on intuition. The paucity of textbooks that incorporate insights from corpora may be attributed to the fact that early corpora tended to be designed for linguists and were difficult to access for materials designers and teachers. In his investigation into the attitudes of text book writers towards corpus materials, Burton (2012) discovered that many of these authors share a lack of knowledge of corpora, in terms of their existence, benefits and exploitation. Considering these findings, he agrees with McCarthy (2008) in his conclusion that to effect change, teachers and their students will need to request that publishers produce materials that reflect the most accurate portrayals of language. This, in turn, underlines the need for language teacher education programs to include readings on corpus linguistics and encourage student teachers to become familiar with the exploitation of corpus materials for learning and teaching language. Timmis (2013) stresses the value of viewing corpora as contributors to course materials rather than arbiters of lexical-grammatical choices. He points out that such a view allows for corpus frequency information to be reconsidered to accommodate e.g., developmental sequences, local need, intuition, cultural and pedagogic considerations and concludes that corpora do not inform practitioners what or how to teach, they do, however, provide valuable information on the nature of language and language production for consideration in materials design.

Definition and Importance of Lexical Bundles

Several researchers have commented on the repetitiveness of language, especially multiword combinations, across registers (e.g., spoken discourse, academic prose, fiction) but refer to these combinations with different terms. Conrad and Biber (2004) list six characteristics of multiword items: fixedness; idiomaticity; frequency; length of sequence; completeness in syntax, semantics, or pragmatics; and intuitive recognition by fluent speakers of a language community (p. 57). For example, the multiword combination *at the drop of a hat* has such components as fixedness, idiomaticity, completeness, and intuitive recognition by native speakers of English. The combination is an idiom that has a fixed form and is recognized by native speakers as one unit (one is unlikely to hear native speakers use a variation such as *at the hat's drop*), with low transparency in meaning. Multiword combinations thus differ from idiomatic expressions and collocations in both form and scope. For a review of full discussion see e.g., Wray (2002). Nattinger and DeCarrico (2001) refer to multiword combinations as *lexical phrases,* stressing the importance of fixedness and pragmatic completeness, while Bahns and Eldaw (1993) use the term *word combination,* which for their purposes does not include the fixedness component. Building on their own and earlier work, Conrad and Biber refer to multiword items as *lexical bundles,* and point out two main criteria: frequency and register. The first criterion, frequency, relates to cut-points, meaning the number of times a lexical bundle occurs in a corpus, in relation to the size of the overall corpus and the research goals. The second criterion relates to multitext occurrences (i.e., dispersion), typically at least five texts in any one register, but again dependent on the corpus and research goals. This criterion is intended to avoid personal preference by individual writers of text in their use of lexical bundles. While Conrad and Biber (2004) recognize that other features are involved in defining multiword combinations, they have identified frequency and multitext occurrences as the most important. They argue that such lexical bundles represent "the most frequent recurring fixed lexical sequences in a register" (p. 59).Researchers have identified lexical bundles of varying lengths but increasingly focus their analyses on 4-word bundles. The structure of 4-word bundles tends to contain 3-word bundles (Cortes, 2003; Hyland, 2008; Wood, 2013) but exclude most non-standard or meaningless bundles of two or three words (Hyland, 2008). Wood (2013) points out that 5- and 6-word bundles are relatively less common than 4-word bundles, thus the longer bundles would provide more limited frequency data.

Research shows that the use of lexical bundles is connected to improved fluency in learners' spoken and written discourse (Fan, 2009; Nation, 2001; Wood, 2010; Wood & Appel). From a psycholinguistic perspective, there is an underlying assumption that such lexical bundles are stored as one unit, making their recognition and retrieval easier, faster, and requiring less attention to complete a task, thereby freeing up processing capacity for greater fluency (Conrad & Biber, 2004; Wood, 2010). To examine the relationship between ESL learners' use of lexical bundles in academic writing and their English language ability, Appel (2016) analysed argumentative essays the learners wrote for the Canadian Academic English Language (CAEL) test. The resulting corpus of essays was divided into three subcorpora: the Lower Level Corpus (LLC), which included essays that the examiners had judged to be at a beginner level, the Medium Level Corpus (MLC), texts produced by intermediate level writers, and the High Level Corpus (HLC) from upper-intermediate and advanced level writers. The lexical bundles in each subcorpus were then examined in terms of their frequency similarity, and length. The findings showed that high-level writers tended to use more lexical bundles than low-level writers. In addition, HLC writers typically used shorter bundles with less repetition of usage. Appel's study thus provides support for the notion that lexical bundle use is correlated to ability level in ESL learners.

Functions of Lexical Bundles

Research into multiword combinations or lexical bundles shows that certain types of lexical bundles are frequent in different texts, often due to their functional characteristics, and lexical bundles with specific functions are associated with specific registers and discourses (Schmitt & Carter, 2004). Biber et al. (2004) identified three categories of functions for lexical bundles: *stance, discourse*, and *referential*. *Stance* bundles express personal attitude or modality towards a proposition; *discourse* bundles indicate the relationship between parts of discourse; while *referential* bundles directly indicate the temporal, spatial, and physical attributes of an object or a subject. Conrad and Biber (2004) investigated the role of lexical bundles in spoken and written discourse in two subcorpora, one that included transcripts of conversations from about 500 participants over the course of one week, while the other included research articles and extracts from academic books. The researchers identified, then analysed 3- and 4-word bundles with a minimum frequency requirement of 10 occurrences per million words in each of the two registers. The lexical bundles had to have been used by more than one speaker in the conversation corpus, and to have occurred in at least five different texts in the academic prose corpus. The researchers compared the resulting 4000 bundles from the conversation sub-corpus and the 3000 bundles from the academic subcorpus based on three criteria: frequency in each register, structural pattern, and function. The frequency analysis showed that the bundles appeared more frequently in conversation (28%) than in academic texts (20%). The structural analysis showed that most of the bundles in conversations included part of a verb phrase while most of the bundles in academic texts included parts of noun phrases and/or prepositional phrases. Finally, the function analysis, which focused only on 4-word bundles as longer bundles are less frequent and typically include 4-word bundles as part of their structures, showed that register resulted in noticeable differences between the function bundles. For example, epistemic stance and discourse organizing bundles were more frequent in conversation, while referential bundles occurred widely in academic texts. An additional category of function identified, special conversational bundles, covered such functions as politeness routines (*thank you very much*), simple inquiry (*what are you doing?*), and reporting clauses (*I said to him*), were identified in conversational discourse only. Further qualitative analysis showed that epistemic stance bundles in conversation were widely used to express personal uncertainty, opinions, desires, and intentions, while stance bundles in academic prose reflected personal certainty. Discourse organizing bundles in conversation were used to introduce or focus on a topic or as clarification, while the same type of bundles in academic texts was used to convey explicit contrast. Conrad and Biber concluded that while lexical bundle use is frequent in both conversations and written academic texts, the type of bundle used depends on register, its context, and purpose. Their findings show that lexical bundle use is not accidental but reflects common patterns and types of bundles that vary depending on register, context, and purposes. One conclusion that is suggested in these findings is that language learners would likely benefit from some explicit instruction in some of the most common patterns and bundles relevant to their learning goals.

In related research, Wood and Appel (2013) IN their analyses of the lexical bundles from the business and engineering textbooks, showed that referential bundles were the most frequently occurring (62%), followed by discourse organizing (24%), and stance bundles (14%). The researchers attribute the large number of referential bundles to the fact that textbooks typically point out and explain subject matter. Wood and Appel suggest that awareness of high-frequency lexical bundles used in different disciplines is likely to assist teachers and materials developers in selecting the most appropriate items to include in textbooks of various disciplines. The inclusion of lexical bundles in language teaching thus should serve to benefit learners' awareness and linguistic ability.

An investigation of lexical bundles and their functions in relation to discourse structure is also the focus in Csomay (2013), who examines classroom discourse. A corpus based on selected data from the TOEFL 2000 Spoken and Written Academic Language corpus and the first six units of 196 university classroom sessions in the Michigan Corpus of Academic Spoken English was analysed. The 84 4-word bundles identified were analysed for their functions. The findings show that stance bundles were used more frequently in the opening phase of classroom session, while referential bundles were used more frequently in the instructional phase of the classroom discourse. Stance bundles were typically used to convey personal obligation (e.g., *I don't know; do you think so)* , while directive (e.g., *it is necessary to; you don't have to)* and referential bundles (e.g., *at the same time; one of the most)* were used to express time, place, and the specification of attributes. Discourse organizing bundles (e.g., *what do you think; on the other hand*) were the least frequent in classroom discourse. Similar to previous studies, Csomay concluded that the use of different types of lexical bundles varied according to the communicative context and purpose and also suggested that the inclusion of different types of lexical bundles in pedagogy would likely enhance students' understanding of these lexical items in academic settings.

The above studies support the notion that various registers are associated with different types of lexical bundles, based on the context and purposes of a register. Further research will likely clarify and confirm the various associations. In the meantime, the authors of the above studies tend to agree that findings should be reflected in textbooks and other classroom materials. While textbooks might be expected to reflect frequently occurring lexical bundles, studies exploring the relationship between textbooks and relevant corpora are not yet readily available. Yet, the underlying assumption is that explicit explanations and illustrations in appropriate text selections will have beneficial effects on learners' language development. To address this perceived gap in the literature, the study described in the following was designed to determine the extent to which a textbook used for intermediate level English learners incorporates both relevant examples of lexical bundles and their functions. Findings serve to shed light on the relationship between textbook language and language use as reflected in corpora.

This Study

The general English textbook *English File: Intermediate Student's Book* (Latham-Koenig & Oxenden, (2013) was selected as it is a widely used, with much of it available online. The text is intended to focus on spoken English for general purposes, consists of 10 units divided into sections including grammar, vocabulary, pronunciation and *practical English episodes*.

Methods

The data for the current study consist of an electronic version of the textbook under investigation. The 41,752-word corpus created includes the reading texts, dialogues, and listening transcripts from all parts of all units, but excludes grammar and vocabulary exercises, which include tasks such as matching, fill-in-the-blank, answering questions, and instructional language. Similarly, items with names of people, nicknames, names of countries, states and websites and social media were also excluded from analysis to avoid coincidences related to the textbook itself.

Four-word bundles were generated through use of *kfNgram* concordancing software (Fletcher, 2012), a free tool that extracts lexical bundles and provides frequency numbers. To generate the bundles, *kfNgram* was set to extract 4-word bundles that had at least three occurrences, which reflects the frequency cut-off of 40-99 times per million words identified in Biber et al. (2004). This frequency requirement resulted in a total of 222 4-word bundles. These 4-word bundles were analysed to identify sequences that were true 4-word bundles rather than 3-word bundles with variable slots (Wood, 2013). The procedure entails the separation of each 4-word bundle into two 3-word bundles. If frequency counts indicate that the 3-word bundle is more frequent than the 4-word bundle, the 3-word bundle is considered to be the base structure. For example, the 4-word bundle *in other words the* can be separated into *in other words* and *other words the.* As the frequency of the former is greater than that of the latter, *in other words* is considered the base structure and the article *the* is considered a variable slot and placed in parentheses. Once all 4-word bundles generated by *kfNgram* had been analysed, a list of 169 4-word bundles with between three to 10 occurrences resulted, as illustrated in Table 1.

Table 1
Frequency of 4-Word Bundles

Frequency	Number of items	Percentage	Example
10	2	1.2	I don't know; I don't think
8	2	1.2	I don't want; do you think you
7	6	3.6	at the end of; don't want to
6	5	2.9	as soon as I; if you don't
5	6	3.6	do you think you; I was going to
4	22	13	and there is a; do you have a
3	127	75.1	about going to the; can you pass the
Total	169	100%	

Analyses

Three stages of analysis served to address the research question. The first stage assessed the 4-word bundles based on their usefulness score. The second stage identified the various functions of the 4-word bundles in the textbook corpus. A quantitative and qualitative analysis of the usefulness scores and functions was carried out in the third stage of analysis. Each stage is described in the following.

Research referred to in the above has shown that the importance of a given lexical item is reflected in its frequency in naturally occurring texts from different but relevant sources. Koprowski (2005) suggested a procedure to assign usefulness scores, i.e., a value that captures the frequency of lexical items in terms of occurrences per million words in specific corpora in addition to information about range, which refers to the number of registers or text types in which a given lexical item can be found. Following Koprowski, usefulness scores were assigned to the 4-word bundles in the textbook under investigation by comparing analysed items with the COBUILD concordance, to determine their frequency data in five sub-corpora of different text types, where the analysed items were most commonly found. For the first stage of analyses, averaged frequency scores from the five individual frequency scores provided the usefulness score for each 4-word bundle and reflect their frequency and range across five text types.

A second stage of analyses involved identifying the various functions of all 4-word bundles in the textbook corpus. Whilst the purpose of functions varies depending on register, Conrad and Biber (2004) identified three types of functions of 4-word bundles: *stance expressions, discourse organizers*, and *referential expressions*. Table 2 shows that the function *stance expressions* includes bundles that reflect personal or impersonal attitudes towards an action or event in a text and is sub-divided into *epistemic* bundles and *attitudinal/modality* bundles, a group that is further divided *into desire, obligation/directive, intention/prediction* and *ability*. The function *discourse organizers* is divided into the sub-categories topic *introduction/focus* and *topic elaboration/clarification* bundles. The former introduces new topics or directs attention toward specific topics, the latter provides additional information or clarification to a topic. The third function includes *referential* bundles, which indicate specific features of physical or abstract entities. *Referential bundles* are divided into the four sub-categories *identification/focus, imprecision, specification of attributes*, and *multi-functional* bundles. In turn, these sub-categories serve to stress the importance of an object, reflect imprecision or uncertainty about an object, focus on selected aspects of an object and may include quantity, physical or abstract attributes and, the fourth sub-category, to refer to various time-related aspects. Conrad and Biber also identified a specifically *conversational* function, which includes categories such as politeness routines, simple inquiry and reporting clauses. Bundles from this last function appear in their conversation sub-corpus only. A summary of these functions and their sub-categories is offered in Table 2.

Conrad and Biber's four functions and their sub-categories served to classify the 4-word bundles from the textbook corpus. Bundles that did not clearly fit into any of these functional categories were placed into a *no-function* category for further analysis.

The third stage of analysis entailed the quantitative and qualitative analysis of the usefulness scores and functions of the extracted 4-word bundles. Each function and its subcategories was allocated the overall usefulness scores achieved by averaging the usefulness scores of the items under the functions and their subcategories. The purpose of the analyses was to determine the relationship between the functions of 4-word bundles and their usefulness and represents the final stage in the analyses carried out to answer the research questions. These stages served to answer three specific research questions:

1. What is the relationship between the 4-word lexical bundles identified in the textbook under investigation and corpus-research findings in terms their frequency and range?
2. How do the 4-word lexical bundles presented in the textbook reflect corpus-research findings in terms of their functions?
3. What is the relationship between the usefulness and functions of the 4-word lexical bundles in the textbook?

A key assumption underlying these questions is that textbooks intended for general language purposes reflect frequently occurring 4-word lexical bundles in corpora collected from naturally occurring language.

Table 2
Functions of Bundles According to Conrad and Biber (2004)

Function	Stance expressions	Epistemic	

		Attitudinal/modality	desire;
			obligation/directive;
			intention/prediction;
			ability

	Discourse organizers	Topic Introduction/focus	

	Referential expressions	Identification/focus;
		imprecision;
		specification of attributes;
		multifunctional;	

	Special conversational	Politeness routines;
		simple inquiry;
		reporting clause.	

Findings

The findings from the analyses described above will be presented to respond to each of the three research questions. The first question *What is the relationship between the 4-word lexical bundles identified in the textbook under investigation and corpus-research findings in terms their frequency and range?* is addressed through the usefulness score. This score, representing frequency and range, was determined based on information from COCA and BNC and shows that the 169 4-word bundles vary in usefulness between a high of 93.78 and a low of 0, with an average usefulness score for all the items of 4.4. Nineteen of the 169 lexical bundles identified in the textbook, 11.2% of the total number of 4-word bundles, reach a usefulness score over 10, as shown in Table 3.

A total of 20 (11.8%) of the 169 4-word bundles in General English have usefulness scores of zero, indicating that they did not occur in either COCA or BNC, while another 88 (52%) 4-word bundles in General English have usefulness scores between 0.005 and 0.995. In addition, 13 (7.7%) of the 4-word bundles have a raw frequency of one to four occurrences in both corpora, or one to four occurrences in one corpus and zero occurrences in the other. For example, the bundle *it is considered bad* has zero occurrences in COCA and two in BNC. The limited number of 4-word bundles with high usefulness scores, the low average usefulness score and large percentage of items with zero usefulness scores suggest that the 4-word bundles included in the textbook have comparatively low range and frequency in everyday language as reflected in COCA and BNC.

Table 3
Four-word Bundles with Usefulness Scores of over 10

Item	Frequency per million words in COCA	Frequency per million words in BNC	Usefulness score
the end of the	83.24	104.32	93.78
at the end of	68.87	91.68	80.275
for the first time	63.18	53.4	58.29
on the other hand	48.38	52.62	50.5
one of the most	54.18	40.49	47.335
in the middle of	48.51	28.07	38.29
the middle of the	31.58	22.11	26.845
was one of the	26.36	23.05	24.705
what do you think	30.9	12.4	21.65
the back of the	21.36	20.82	21.09
I'd like to	17.22	15.75	16.485
I was going to	18.39	10.8	14.595
do you want to	14.58	11.25	12.915
in one of the	13.52	12.1	12.81
from time to time	9.28	16.32	12.8
a member of the	23.7	0	11.85
a bit of a	7.63	15.74	11.685
what do you mean	11.8	9.72	10.76
a lot of money	13.02	7.87	10.445

To answer the second question asked in this study, *How do the 4-word lexical bundles presented in the textbook reflect corpus-research findings in terms of their functions?,* the 169 4-word bundles identified in the textbook examined were analysed and sorted according to the different functions identified in Conrad and Biber (2004). The findings show that 55 (32.5%) of the 4-word bundles reflect identifiable functions, of which referential ones were the most frequent, followed by stance, special conversational and, the least frequent, discourse organizer functions, as summarized in Table 4.

The most frequent of the functions identified were referential expressions with 21 (12%) items, of which 6 (3.6%) items fall into the subcategory of identification/focus (e.g., *in one of the*, *one of my best)*, 12 (7.1%) items under specification of attributes (e.g., a bit of a, as soon as I), 3 (1.8%) items under multi-functional (e.g., *the end of the, and in the end*), while no items were identified as part of imprecision subcategories. The second most frequently occurring function in the textbook is that of stance expressions with 16 (9.5%) items, of which 9 (5.3%) are epistemic bundles (e.g., *do you know if, I don't know*) and 7 (4.1%) attitudinal bundles (e.g., *do you want to, I was going to*). The function for special conversational expressions included 14 (8.3%) items, 2 (1.2%) of which are part of the subcategory of politeness routines (e.g., *no thanks I'm*), 12 (7.1%) items of simple inquiries (e.g., *can you tell me*), none of reporting clauses. The least frequently identified category of functions, discourse organizers, includes 4 (2.4%) items that belong to the topic elaboration/clarification (e.g., *on the other hand*) and none for introduction/focus subcategory.

Table 4
Summary of functions in General English textbook

Functions		Number of occurrences	Percentage
Referential	Total	21	12.0
	Identification/focus	6	3.6
	Specification of attributes	12	7.1
	Multi-functional	3	1.8
	Imprecision	0	0
Stance	Total	16	9.5
	Epistemic	9	5.3
	Attitudinal	7	4.1
Special conversational	Total	14	8.3
	Politeness routines	2	1.2
	Simple inquiries	12	7.1
	Reporting clause	0	0
Discourse organizer	Total	4	2.4
	Topic elaboration/clarification	4	2.4
	Topic introduction/focus	0	0
No-function	Total	144	66.9
	Collocational phrases	14	8.3
	Context specific	28	16.6
	No subcategory	71	42.0

The no-function category includes 144 (66.9%) 4-word bundles, a large enough category to suggest further analysis. This analysis shows that 14 (8.3%) of these bundles belong to the collocational phrases (Conrad & Biber, 2004) subcategory (e.g., *had a great time, o'clock in the morning*), while an additional 28 (16.6%) no-function bundles belong to the context specific subcategory (e.g., *lawyer of the defence, the docklands light railway*). The remaining 71 (42%) bundles could not be attributed to any subcategories reflected in the literature (e.g., *and there is a, I usually have a*). Figure 1 summarizes these findings, reflecting that all the functions and most of the subcategories identified in Conrad and Biber were also identified in the textbook under analysis, but more than half of the 4-word bundles in the textbook could not be attributed to any of the function categories identified.

The third stage of analyses was designed to answer the question *What is the relationship between the usefulness and functions of 4-word lexical bundles in the textbook?* The analysis of the 4-word bundles that reflect one of the functions shows that their overall usefulness score is 11.9. A breakdown of the different function categories identified is presented in Figure 2.

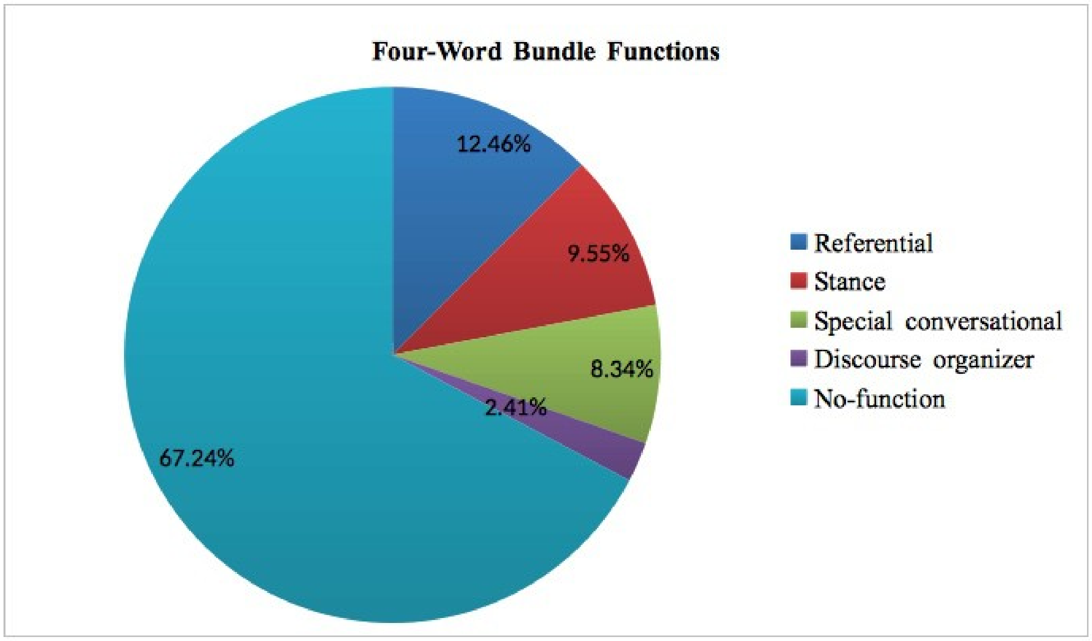

Figure 1. Overall Summary of Functions Identified in Textbook

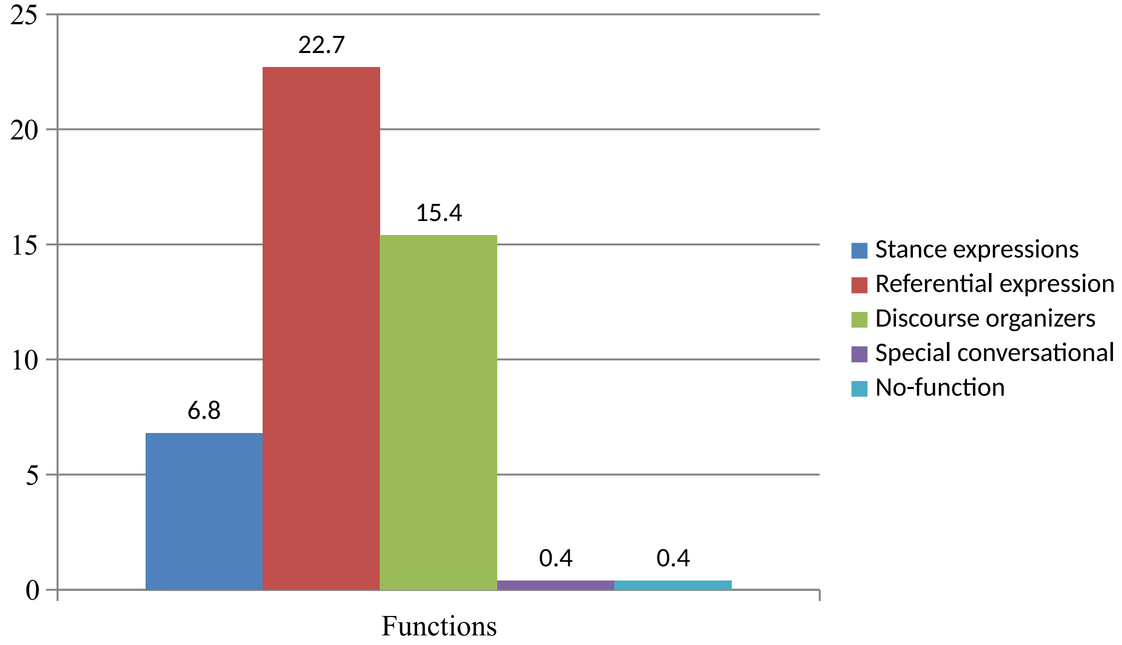

Figure 2. Usefulness Score of Functions

Figure 2 shows that the highest usefulness score, 22.7, was achieved by referential expressions, followed by discourse organizers at 15.4, stance expressions at 6.8 and, special conversational expressions at 0.4. The usefulness score of the no function expressions attained a usefulness score of 6.8, while the This overall usefulness score includes the no-function expressions, which was calculated as 0.4.

The usefulness scores for each function's subcategories were also calculated and are reflected in Figure 3.

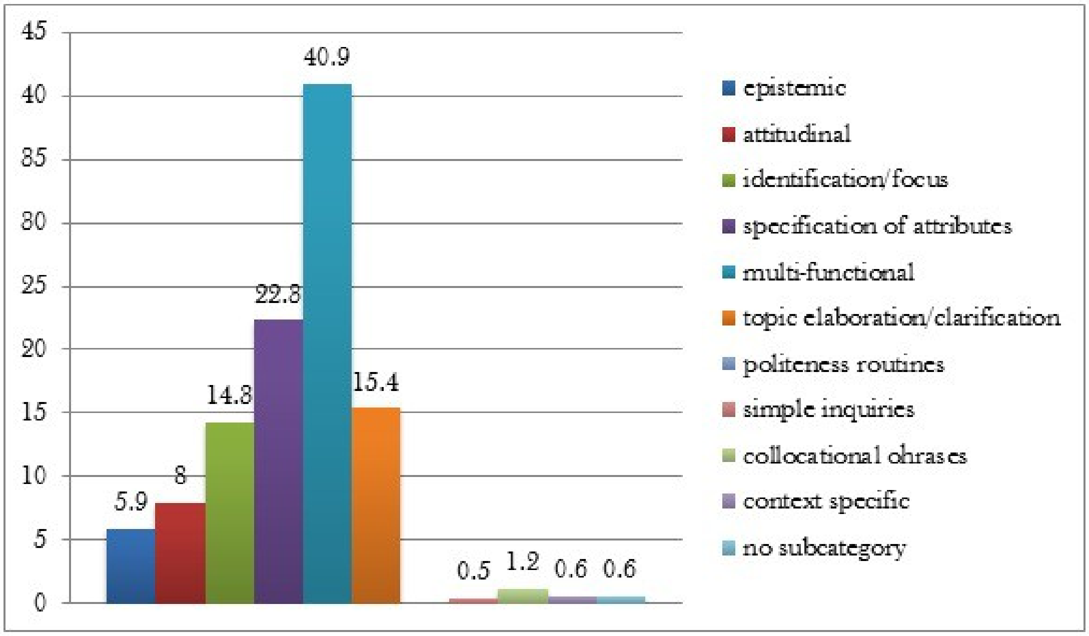

Figure 3. Usefulness Score of Subcategories

It shows that the sub-categories for the referential function achieved usefulness scores of 14.8 for identification/focus, 22.3 for attributes, specification of attributes 22.3 40.9 for multi-functional expressions. Within stance expressions, epistemic expressions reached a usefulness score of 5.9, while attitudinal expressions' usefulness score reached 4.1. Discourse organizers include items from only the topic elaboration/clarification subcategory, which obtained an overall usefulness score of 15.4. Finally, the subcategories for the special conversational function, politeness routines and simple inquiries, achieved usefulness scores of 0 and 0.5 respectively. The no-function subcategory of collocational phrases received a usefulness score of 1.2, the context specific subcategory 0.6 and the uncategorized group 0.6. The above findings show that lexical bundles with functions tend to have higher usefulness scores than those without functions. The most useful items identified in the textbook corpus are part of referential expressions, followed by discourse organizers, stance expressions, with special conversational functions showing the lowest usefulness scores. The most useful subcategory is that of multi-functional expressions, the least useful the one covering politeness routines.

Discussion

Key findings from this exploratory study of the usefulness and function of lexical bundles identified in a textbook for general English language learners are discussed in the order of the specific research questions raised. The first research question explored the level of usefulness of the 4-word bundles generated from the textbook. *Usefulness* was determined through numeric scores developed in Koprowski (2005), scores comprised of frequency and range data from COCA and the BNC. The findings show a comparatively low level of usefulness of the analysed items, determined by their low frequency of usage in various registers and text types reflected in corpora from general language use. The findings in this study are consistent with the findings reported in Koprowski and in

Cheng and Warren (2007), whose work also found low-frequency items and inconsistencies between the vocabulary items included in teaching materials and those found in actual language use reflected in corpus data. The findings also reflect the observation other studies on teaching materials (Biber & Reppen, 2002; Gabrielatos, 2006; Koprowski, 2005; Lee & McGarrell, 2011; Meunier & Gouverneur, 2009; Shortall, 2007) that the language presented in these materials do not closely match the language from naturally occurring language reflected in corpora. Although the stated purpose of the textbook analysed for the current study is to improve students' general English abilities, the findings suggest that most of the lexical bundles included have highly limited usage in general communication contexts. This lack of convergence between textbook and corpus material suggests that the textbook developers may have relied on intuition in the selection of material, as discussed in Biber and Reppen (2002) and Lee and McGarrell (2011) rather than actual data sources, or that the selection criteria used were unable to identify material representative of general language use.

The second research question investigated in the current study examined the functions, as defined by Conrad and Biber (2004), of the 4-word lexical bundles identified in the textbook. The findings show that over 65% of the lexical bundles do not fall within any identifiable function. This may, in part, be due to the low frequency of lexical bundles with at least three occurrences in the textbook, suggesting that the textbook lacks the kind of repetition typically needed for language development. The most frequently identified function of the textbook bundles was referential, followed by stance, special conversational and, least frequent, discourse organizing. The finding that referential bundles are the most frequently occurring in the textbook under discussion reflects an academic influence in its language focus, as shown in previous research. Wood (2013), in an analysis of business and engineering textbooks, showed that the referential function was the most frequently occurring in those academically oriented textbooks. Wood's study shows discourse organizing and stance bundles as the second, respectively third most frequently used functions. Similarly, Conrad and Biber showed that referential bundles are more common in academic prose compared to other language uses. The objectives of the textbook and the focus on academic language again suggests a misalignment between the two. The second most frequent function of lexical bundles in the current study, stance, shows that the textbook also focuses on conversation and speaking registers, but to a lesser extent. These functions are associated with more informal language use, as indicated in Conrad and Biber. The latter's investigation of bundles from conversation and academic prose discovered that stance and special conversational bundles are more frequent in conversation. In light of these findings, the textbook under discussion thus presents a mix of academic and conversational registers. This mix, combined with the relatively low recurrence of bundles, may prevent learners from encountering relevant functions in sufficient numbers for each register to internalise them successfully. In turn, this may impede register appropriate production as the information available to them lacks clear distinctions of function use in different registers.

The third research question investigated the relationship between usefulness and functions of the 4-word bundles identified in the textbook. To address this question, each function and its subcategories were given an overall usefulness score by averaging the usefulness scores of the items under the functions and their subcategories. The findings show that the lexical bundles with functions have a higher usefulness score compared to those that cannot be attributed to any function. This finding is linked directly to past studies that have stressed the importance of referring to frequency information on actual language use in teaching and materials development (Biber & Reppen, 2002; Cheng & Warren, 2007; Koprowski, 2005, Lee & McGarrell, 2011). A detailed analysis also shows that the referential function, which is typically associated with more formal and academic language, has the highest usefulness score in the textbook under discussion. In addition, the second stage of the analysis shows that referential bundles are also the most common type of bundles identified in the current study. They include the items with the highest usefulness scores, such as *the end of the* (93.7), *at the end of* (80.2), *for the first time* (58.2). This finding suggests that the inclusion of referential bundles in teaching syllabi and textbooks working on academic registers may be particularly valuable in support of language learners' ability to acquire native-like multiword expressions. The second most useful type of lexical bundle belongs to the discourse organizing function. For example, the bundle *on the other* hand (50.5) was also shown as frequently occurring in academic prose in Conrad and Biber (2004). Although the discourse organizing function is considered the least common type of function, its high usefulness score suggests that it is a valuable item for inclusion in teaching materials. The stance function, with noticeably lower scores, is third in terms of usefulness in the current study. As the detailed presentation in the results section shows, the stance bundles in the textbook have low usefulness scores, with a few exceptions such as *what do you think* (21.6). A careful analysis of corpus data may help materials developers identify selections that are useful in terms of broad actual language use. The fourth function, the special conversational function, has the lowest usefulness score, even though it was found to be more frequent than the discourse organizing function. One explanation for this may be the range criteria imposed in determining usefulness scores, i.e., the criteria that ensures that lexical items learned are useful in varied contexts. As the special conversational function is expected to be used in the conversational register only, the range of bundles in this category are, by definition, limited. Usefulness scores are comprised of both frequency and range data, thus such types of bundles will not yield high scores, even if the bundles are frequent in their register. A textbook concentrating on conversational English might reasonably be expected to have many lexical bundles that fall within the conversational function. Again, careful matching of corpus data in light of the purposes of a given textbook would seem to be a key objective for materials designers. Finally, the lexical bundles with the lowest usefulness scores, even though they account for over half the bundles identified in the textbook under discussion, were those that fell within the no-function category. One potential explanation for the large number of no-function bundles may be the subject matter around which the textbook presents language items, subject matter that may be guided more by introspection and intuition rather than an analysis of general language needs in relation to data of actual language use reflected in corpus materials. Koprowski (2005:328) noticed a similar outcome in his analysis of the usefulness of lexical phrases in contemporary textbooks and attributes such low usefulness to "an unprincipled and careless selection process" by textbook developers. A selection process that, Koprowski adds, is likely focussed around the selection of themes and topics rather than the usefulness of lexical phrases.

Summary and Conclusions

The current study draws on a corpus created from a widely used textbook intended for learners of general English and function analysis to determine the appropriateness, defined as usefulness and functions, of the lexical bundles presented in the text. The three specific research questions examined whether the 4-word lexical bundles included in the textbook reflect frequency, range and functions of the data reflected in corpora of language in use. The first question explored whether the 4-word bundles in the text were of a frequency and range, according to corpus data, to suggest that they would be useful in general language use. The usefulness score (Koprowski, 2005) assigned to each bundle reflects both frequency and range based on COCA and the BNC. The findings show a comparatively low level of usefulness of the analysed items and reflect similar findings in Conrad and Biber (2004) and Korpowski. Of 169 lexical bundles identified, only 19 reached a useful score of over 10, with just four of these lexical bundles scoring over 50. To probe the usefulness of the 4-word lexical bundles further, the second research question examined the functions of these bundles. The analysis shows that most of the 169 lexical bundles in the textbook do not have identifiable functions and that such no-function bundles have low usefulness scores. By contrast, 4-word bundles with specific functions had relatively high usefulness scores, which suggests that language learning materials developers and teachers might usefully focus on including function bundles when selecting language items. Jones and Haywood (2004) and Byrd and Coxhead (2010), for example, suggest using a list of frequent lexical bundles identified for a specific discipline or register and needs analyses to meet the needs of language learners. Depending on such needs, less frequent 4-word bundles might nevertheless be highly relevant to learners and require teachability/learnability strategies for teachers and their learners (Nation, 2001).

This study provides insight into the 4-word lexical bundles included in one specific textbook in relation to their occurrence in corpora. Whilst its findings cannot be generalized, they raise several questions about the identification and selection of lexical items for a textbook. As Timmis (2013) points out, corpora are less appropriate as arbiters of what to teach and how to teach, but they are valuable in reflecting details about the nature of language and language use. In the case of general English, corpora suggest a considerable difference between corpora and the lexical bundles and functions presented in the textbook. A broader question is the relationship between corpora, which typically include long passages of texts, with typical textbooks and their short, often unconnected texts representing different genres. This question has been addressed in part by studies that highlight differences in the use of lexical bundles depending on genre (e.g., Biber, 2006; Hyland, 2008). In the case of a textbook, one question is whether such learning and teaching materials might be developed to include relevant, engaging topics that serve to illustrate language that is truly *general* and widely used. The inclusion of frequently recurring lexical bundles is particularly important as research shows that even advanced learners of ESL have difficulties in producing texts that reflect native speaker usage (Grami & Alkazemi, 2016). Yet pedagogical materials rarely include activities or instructions on which words go together (Alali & Schmitt, 2012). Increased attention to careful selection of lexical strings that reflect actual language use reflected in relevant corpora can only support the challenging task of developing vocabulary skills, which include appropriate use of lexical bundles. Combined with explorations into ways in which the acquisition of lexical strings might be facilitated in ESL classes, as illustrated in Jones and Haywood (2004) and more recently AlHassan (2016) and AlHassan & Wood (2016), promise to further support the very challenging task of supporting vocabulary development in subsequent language learning.

References

Alali, F., & Schmitt, N. (2012). Teaching formulaic sequences: The same or different from teaching single words? *TESOL Journal, 3*(2), 153–180.

AlHassan, L. (2016). Learning all the parts of the puzzle: Focused instruction of formulaic sequences through the lens of activity theory. In H.M. McGarrell & D. Wood (eds.). *Contact - Refereed Proceedings of TESL Ontario Research Symposium*, *42*(2), 44-65. Available at: http://www.teslontario.net/publication/research-symposium

AlHassan, L. & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes, 17*, 51-62.

Appel, R. (2016). Lexical bundles in L2 English academic writing: Proficiency level differences. In H.M. McGarrell & D. Wood (eds.). *Contact - Refereed Proceedings of TESL Ontario Research Symposium*, *42*(2), 66-81. Available at: http://www.teslontario.net/publication/research-symposium

Ari, O. (2006). Review of three software programs designed to identify lexical bundles. *Language Learning & Technology, 10*(1), 30-37.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: Benjamin.

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

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition, 24*, 199–208.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 2*(3), 371-405.

Burton, G. (2012). Corpora and coursebooks: destined to be strangers forever? *Corpora, 7*(1), 91-108.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*, 31-64. Available at: http://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume05/Article02.pdf

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*(2), 141-158.

Cheng, W., & Warren, M. (2007). Checking understandings: Comparing textbooks and a corpus of spoken English in Hong Kong. *Language Awareness, 16*(3), 190-207.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71.

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

Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics, 34*(3), 369-388.

Davies, M. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at http://corpus.byu.edu/

Fletcher, W. (2012). *kfNgram: Information and help*. Available at: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

Gabrielatos, C. (2006). Corpus-based evaluation of pedagogical materials: *If*-conditionals in ELT coursebooks and the BNC. In: 7th Teaching and Language Corpora Conference. Available online at http://eprints.lancs.ac.uk/882/

Grami, G., & Alkazemi, B.Y. (2016). Improving ESL writing using an online formulaic sequence word-combination checker. *Journal of Computer Assisted Learning, 82*, 95–104.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sentences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 269-291). Amsterdam, Netherlands: John Benjamins.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*, 322–332.

Latham-Koenig, C., & Oxenden, C. (2013). *English file: Intermediate student's book.* Oxford, UK: Oxford University Press.

Lee, D., & McGarrell, H. (2011). Corpus-based/corpus-informed English language learner grammar textbooks: An example of how research informs pedagogy. In H.M. McGarrell & D. Wood (Eds.). Contact - *Refereed Proceedings of TESL Ontario Research Symposium, 37*(2), 78–100. Available at: http://www.teslontario.net/publication/research-symposium

McCarten, J. (2010). Corpus-informed course book design. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 413–427). London, UK: Routledge.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching, 41*(4), 563-574.

McDonough, J., Shaw, C., & Mashuhara, H. (2012). *Materials and methods in ELT : a teacher's guide*. Malden, MA : Blackwell.

Meunier, F., & Gouverneur C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Corpora and Language Teaching,* (pp. 179-201). Amsterdam & Philadelphia: John Benjamins.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Neary-Sundquist, C.A. (2015). Aspects of Vocabulary Knowledge in German Textbooks. *Foreign Language Annals, 48*(1), 68–81.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1–22). Amsterdam, Netherlands: John Benjamins.

Shortall, T. (2007). The L2 syllabus: Corpus or contrivance? *Corpora, 2*(2), 157-185.

Timmis, I. (2013). Corpora and materials: Towards a working relationship. In B. Tomlinson (Ed.), *Developing materials for language teaching* (2nd ed.) (pp. 461-474). London, UK: Bloomsbury Academic.

Wood, D., & Appel, R. (2013). Formulaic sequences in first year university business and engineering textbooks: A resource for EAP. *In H.M. McGarrell & D. Wood (eds.). Contact - Refereed Proceedings of TESL Ontario Research Symposium, 39*(2), 92-102. Available at: http://www.teslontario.net/publication/research-symposium

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. Bloomsbury Publishing.

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

About the Authors

Hedy McGarrell is a faculty member in Applied Linguistics at Brock University where she teaches undergraduate and graduate courses.

Nga Tuiet Zyong Nguien is an ESL teacher with a special interest in corpus applications to ESL teaching. She has an MA Applied Linguistics/TESL from Brock University.