LEXICAL BUNDLES IN L1 AND L2 ACADEMIC WRITING

Yu-Hua Chen and Paul Baker
Lancaster University

This paper adopts an automated frequency-driven approach to identify frequently-used word combinations (i.e., *lexical bundles*) in academic writing. Lexical bundles retrieved from one corpus of published academic texts and two corpora of student academic writing (one L1, the other L2), were investigated both quantitatively and qualitatively. Published academic writing was found to exhibit the widest range of lexical bundles whereas L2 student writing showed the smallest range. Furthermore, some high-frequency expressions in published texts, such as *in the context of,* were underused in both student corpora, while the L2 student writers overused certain expressions (e.g., *all over the world*) which native academics rarely used. The findings drawn from structural and functional analyses of lexical bundles also have some pedagogical implications.

INTRODUCTION

"Phraseology" (Granger & Meunier, 2008; Meunier & Granger, 2007) and "formulaic sequences/language" (Schmitt, 2004; Wray, 2002, 2008) are two umbrella terms often used to refer to various types of multi-word units. In recent years, an increasing number of studies have made use of corpus data to add weight to the importance of multi-word units in language. For instance Altenberg (1998), in his exploration of the London-Lund Corpus, estimated that 80% of the words in the corpus formed part of recurrent word combinations. As Wray (2002, p. 9) observes, however, there is a "problem of terminology" when describing word co-occurrence. On the one hand, the same term might be used in different ways by different scholars; on the other hand, various terms are used to refer to similar or even the same notion of word co-occurrence. Some examples of such terms include *clusters* (Hyland, 2008a; Schmitt, Grandage & Adolphs, 2004; also used in the corpus tool *WordSmith*), *recurrent word combinations* (Altenberg, 1998; De Cock, 1998), *phrasicon* (De Cock, Granger, Leech, & McEnery, 1998), *n-grams* (Stubbs, 2007a, 2007b) and *lexical bundles* (e.g., Biber & Barbieri, 2007; Cortes, 2002). These terms—*clusters, phrasicon, n-grams, recurrent word combinations*, *lexical bundles*—actually refer to continuous word sequences retrieved by taking a corpus-driven approach with specified frequency and distribution criteria. The retrieved recurrent sequences are fixed multi-word units that have customary pragmatic and/or discourse functions, used and recognized by the speakers of a language within certain contexts. This methodology is considered to be a frequency-based approach for determining phraseology (see Granger & Paquot, 2008).

From a psycholinguistic viewpoint, formulaic language has been found to have "a processing advantage over creatively generated language" for non-native as well as native speakers (Conklin & Schmitt, 2008, p. 72), although different psycholinguistic studies have used various types of formulaic language, such as idioms (e.g., *take the bull by the horn*) or non-idiomatic phrases (e.g., *as soon as*), as the target forms. A particularly inspirational study was conducted by Jiang and Nekrasova (2007), in which they utilized corpus-derived recurrent word combinations as materials in two online grammaticality-judgment experiments. Their findings provide "prevailing evidence in support of the holistic nature of formula representation and processing in second language speakers" (Jiang & Nekrasova, 2007, p. 433). Schmitt et al. (2004) also investigated the psycholinguistic validity of corpus-derived recurrent clusters and share some similarities with Jiang and Nekrasova (2007).

In a series of lexical bundle studies conducted by Biber and colleagues (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2003, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999), it was found that conversation and academic prose present distinctive distribution patterns of lexical

bundles. For example, most bundles in conversation are clausal, whereas most bundles in academic prose are phrasal. Other studies of bundles have focused primarily on comparisons between expert and non-expert writing. Cortes (2002) investigated bundles in native freshman compositions and found that the bundles used by these novice writers were functionally different from those in published academic prose. In another study, Cortes (2004) compared native student writing with that in academic journals, concluding that students rarely used the lexical bundles identified in the corpus of published writing. Even if they did, the students used these bundles in a different manner. Working with academic writing only, Hyland (2008b) indicated that there was disciplinary variation in the use of lexical bundles. He also investigated the role of lexical bundles in published academic prose and in postgraduate writing and found that postgraduate students tended to employ more formulaic expressions than native academics in order to display their competence (Hyland, 2008a).

To date, only a few studies of L2 written data have performed structural and functional categorization of lexical bundles. Although Hyland, in his two studies (2008a, 2008b), included masters' theses and doctoral dissertations produced by L2 English students in Hong Kong, he did not begin from a perspective of second-language learning. Instead, he treated L2 postgraduate writing as "highly proficient," on the ground that all the data in his corpus of texts had been awarded high passes. Drawing on the previous research, the present study aims to compare the use of recurrent word combinations in native-speaker and non-native speaker academic writing in order to reveal the potential problems in second language learning. Quantitative and qualitative analyses were carried out on three corpora in order to identify similarities and differences in recurrent word combinations at different levels of writing proficiency. One corpus (the L2 or learner corpus) contained writing from L1 Chinese learners of L2 English, while the two other comprised L1 writing: one from academics (whom we term "expert" writers) and the other university students (who are similar in background to the L1 Chinese learners, aside from their first language). *Lexical bundles* is adopted as the primary term throughout this study, as it is used by Biber in a series of studies upon which the theoretical and analytical framework of the current study is based. Another term, *recurrent word combination,* is also used interchangeably, given its transparent literal meaning.

DATA AND METHODOLOGY

Data

Two existing corpora are used in the present study: the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus, and the British Academic Written English (BAWE) corpus. To ensure comparability, only part of each corpus was selected for investigation. The FLOB corpus is a one-million-word corpus of written British English from the early 1990s, comprising fifteen genre categories. For the current study, only the category of academic prose, FLOB-J, was used to represent native expert writing. FLOB-J contains eighty 2,000-word excerpts from published academic texts, retrieved from journals or book sections. With regard to L1 and L2 student academic writing, parts of the BAWE corpus were utilized. The BAWE corpus, released in 2008, contains approximately 3,000 pieces (approx. 6.5m. words) of proficient assessed student writing from British universities. Two subcorpora were selected from the BAWE corpus: BAWE-CH contains essays produced by L1 Chinese students of L2 English, and BAWE-EN is a comparable dataset contributed by peer L1 English students. FLOB-J, BAWE-CH and BAWE-EN cover a wide range of disciplines, including arts and humanities, life sciences, physical sciences and social sciences (for BAWE, see Alsop & Nesi, 2009; for FLOB, see Hundt, Sand & Siemund, 1998). The size of each finalized corpus for investigation is around 150,000 words (see Table 1).

Table 1. *Constituents of the Three Academic Corpora*

Representation	Corpus	Word count	Average length of text	No. of texts
Native expert writing	FLOB-J	164,742	2,059	80
Native peer writing	BAWE-EN	155,781	2,596	60
Learner writing	BAWE-CH	146,872	2,771	53

Operationalization

Several key criteria have been pinpointed in the literature regarding how to generate a list of lexical bundles using automated corpus tools. The first criterion is the cut-off frequency, which determines the number of lexical bundles to be included in the analysis. The normalized frequency threshold for large written corpora generally ranges between 20-40 per million words (e.g., Biber et al., 2004; Hyland, 2008b), while for relatively small spoken corpora, a raw cut-off frequency is often used, ranging from 2-10 (e.g., Altenberg, 1998; De Cock, 1998). The second criterion is the requirement that combinations have to occur in different texts, usually in at least 3-5 texts (e.g., Biber & Barbieri, 2007; Cortes, 2004), or 10% of texts (e.g., Hyland, 2008a), which helps to avoid idiosyncrasies from individual writers/speakers. The last issue concerns the length of word combinations, usually 2-, 3-, 4-, 5-, or 6-word units. Four-word sequences are found to be the most researched length for writing studies, probably because the number of 4-word bundles is often within a manageable size (around 100) for manual categorization and concordance checks. The frequency and dispersion thresholds adopted vary from study to study, and even the sizes of corpora and subcorpora differ drastically, ranging from around 40,000 to over 5 million words.

After repeated experiments with the corpus data under investigation, the frequency and distribution thresholds for determining 4-word lexical bundles were set to 4 times or more (approximately 25 times per million words on average), occurring in at least three texts. This resulted in an "optimum" number of bundles, which was considered sufficiently representative of the corpora being examined. One might argue that an identical standardized threshold, such as 20 or 40 times per million words, should be applied to each of the corpora investigated, as generally reported in the literature. However, when a normalized rate is converted to raw frequencies, it substantially affects the number of generated word combinations when comparing corpora of various sizes. For instance, if we compare an 80,000-word corpus with a 40,000-word corpus with a cut-off standardized frequency set at 40 times per million words, it means that the converted raw-frequency threshold for the larger corpus is 3.2, whereas the converted raw-frequency threshold for the smaller corpus is much lower, at 1.6. Any decimals have to be rounded up or down in order to function as an operational cut-off frequency. Yet rounding down 3.2 to 3 results in a normalized rate of 37.5 whereas rounding up 1.6 to 2 generates a normalized rate of 50, both of which are different from the originally reported frequency threshold of 40 times per million words. Reporting only the standardized frequency criterion could therefore be misleading, because a standardized cut-off frequency would inevitably lose its expected impartiality after being converted into raw frequencies corresponding to different corpus sizes. In this study, it could be argued that both the raw cut-off frequency and corresponding normalized frequency should be reported in order to reflect transparently the threshold adopted. For the sake of comparison, if the frequency threshold is set at 25 times per million words for the present study, the converted raw frequencies for each corpus are 3.7, 3.9 and 4.1 times respectively, which are all rounded up or down to 4 (cf. Table 2 and Table 3).

Table 2. *Raw and Corresponding Normalized Frequency Thresholds Adopted*

Corpus	Set raw frequency threshold	Corresponding normalized frequency (per million words)
FLOB-J	4	24.3
BAWE-EN	4	25.7
BAWE-CH	4	27.2

Table 3. *Normalized and Corresponding Raw Frequency Thresholds for Comparison*

Corpus	Set normalized frequency threshold (per million words)	Corresponding raw frequency
FLOB-J	25	3.7
BAWE-EN	25	3.9
BAWE-CH	25	4.1

After automatic retrieval of 4-word clusters using the corpus tool *WordSmith 4.0* (Scott, 2007), word sequences containing content words that were present in the essay questions (e.g., *financial and non financial*), or any other context-dependent bundles, usually incorporating proper nouns (e.g., *in the UK and*, *the Second World War*), were manually excluded from the extracted bundle lists. It was also found that overlapping word sequences could inflate the results of quantitative analysis. Overlaps were thus checked manually via concordance analyses. Two major types of overlaps are discussed here. One is "complete overlap," referring to two 4-word bundles which are actually derived from a single 5-word combination. For example, *it has been suggested* and *has been suggested that* both occur six times, coming from the longer expression *it has been suggested that*. The other type of overlap is "complete subsumption," referring to a situation where two or more 4-word bundles overlap and the occurrences of one of the bundles subsume those of the other overlapping bundle(s). For example, *as a result of* occurs 17 times, while *a result of the* occurs five times, both of which occur as a subset of the 5-word bundle *as a result of the*. Each case of the above overlapping word sequences (12 cases in total) were combined into one longer unit so as to guard against inflated results.

A further potential problem when comparing bundles across corpora involves what is actually counted (i.e., type/token distinction). Should we count the number of types of bundles (e.g., counting *as a result of* and *it is possible to* each as one type of bundle), or should we count the total occurrence of bundles (e.g., *as a result of* might occur 20 times in one corpus and 50 times in another)? One corpus could exhibit a very narrow range of bundles but have very high frequencies of them, while another might have the opposite pattern. We therefore distinguished between different types of bundles (*types*) and frequencies of bundles (*tokens*).[1] The numbers of bundle types and tokens, before and after data refinement, including removing context-dependent bundles and overlapping ones, are shown in Table 4 below.

Table 4. *Number of Bundles Before and After the Removal of Context-Dependent Bundles and Overlaps*

Corpus	Before refinement		After refinement	
	No. of lexical bundles (types)	No. of lexical bundles (tokens)	No. of lexical bundles (types)	No. of lexical bundles (tokens)
FLOB-J	118	749	108	704
BAWE-EN	120	757	104	667
BAWE-CH	90	554	80	507

ANALYSIS AND RESULTS

Our analyses in the following section are based on the recurrent word combinations retrieved and refined (for the full list, see Appendix). In this section, structural and functional comparisons are made between the three groups of different writing proficiency levels. At the beginning of each sub-section, Structures or Discourse Functions, we begin by illustrating how the lexical bundles are categorized, structurally or functionally. Then we go on to the examples and discuss how usage of these word combinations is different and/or similar in the three groups of writers, in terms of both structures and discourse functions. For functional analysis, we look further at the quantitative comparisons with some statistical analysis.

Structures

The structural classification of lexical bundles in the Longman Grammar of Spoken and Written English (Biber et al., 1999) has been widely used in other studies on recurrent word combinations (Cortes, 2002, 2004; Hyland, 2008a, 2008b). In the Longman Spoken and Written English (LSWE) corpus, fourteen categories of lexical bundles are grouped in conversation and twelve categories in academic prose with some overlap between them. Here, a structural classification, following the LSWE taxonomy, was carried out on the lexical bundles retrieved from FLOB-J, BAWE-EN and BAWE-CH. The results were then compared with the proportions of structural categories in the LSWE corpus. As shown in Table 5, despite the drastic difference in corpus size[2] and different frequency thresholds (ten times per million words for LSWE, and four times as the raw cut-off frequency for the current study), there appears to be a surprising close match between the academic prose component of LSWE and FLOB-J, while the proportions for the two groups of student writing fluctuate to some extent when compared with the academic prose in LSWE. Not only does such comparison lend a good deal of credence to the use of smaller corpora with different frequency cut-offs in the current project, but it also indicates a gap between native expert academic prose and immature student academic writing. This gap might be a result of genre difference between published academic essays and university assignments, but it is more likely that it hinges on writing proficiency.

Three broad structural categories were distinguished: "NP-based," "PP-based," and "VP-based." NP-based bundles include any noun phrases with post-modifier fragments, such as *the role of the* or *the way in which* (i.e., Category (1) in Table 5). PP-based bundles refer to those starting with a preposition plus a noun-phrase fragment, such as *at the end of* or *in relation to the* (i.e., Category (2) in Table 5). With regard to VP-based bundles, any word combinations with a verb component, such as *in order to make* or *was one of the*, is assigned to this category (i.e., Categories (3) to (8) in Table 5).

In Table 5, it can be seen that the use of NP-based bundles differs the most amongst the three groups of writing. We thus grouped the NP-based combinations further into two structural subcategories to see more precisely how these three corpora were distinguished from each other. These two subcategories are noun phrase fragments with *of* (NP + *of*) (e.g., *in the context of*) and any other noun phrase fragments without *of* (NPf) (e.g., *the way in which*). In addition to the relatively low proportion of NP-based bundles when compared with FLOB-J, the Chinese student writing represented in BAWE-CH is notably different from the two groups of native writing in the subcategory of NPf, because there is no NPf bundle in BAWE-CH. In contrast, the NPf bundles present in FLOB-J are mostly used by the British students in BAWE-EN, although there are some slight variations (see Table 6). The NPf combinations found in this investigation are all part of relative clauses, such as *the extent to which*, *the fact that this*, or *the way(s) in which*. It is evident that these L2 students did not use these types of relative clause as frequently as native speakers did.

Table 5. *Proportional Distribution of Lexical Bundles (Types) Across the Structural Categories in LSWE, FLOB-J, BAWE-EN and BAWE-CH (cf. Biber et al., 1999)*

Category		Pattern	ACAD (LSWE)	FLOB-J	BAWE-EN	BAWE-CH	Example
NP-based	(1)	noun phrase with post-modifier fragment	30%	32.5%	15.4%	15%	*the nature of the*
PP-based	(2)	preposition + noun phrase fragment	33%	36%	28.8%	32.5%	*as a result of*
VP-based	(3)	copula *be* + NP/AdjectiveP	2%	2.6%	10.6%	6.3%	*is one of the*
	(4)	VP with active verb	--	0.9%	2.9%	6.3%	*has a number of*
	(5)	anticipatory *it* + VP/adjectiveP + (complement-clause)	9%	8.8%	5.8%	8.8%	*it is possible to*
	(6)	passive verb + PP fragment	6%	7%	10.6%	5%	*is based on the*
	(7)	(VP +) *that*-clause fragment	5%	2.6%	4.8%	6.3%	*should be noted that*
	(8)	(verb/adjective +) *to*-clause fragment	9%	7%	18.3%	15%	*are likely to be*
	(9)	others	6%	2.6%	2.8%	4.8%	*as well as the*
		Total	100%	100%	100%	100%	

Table 6. *Similar NPf Bundles in FLOB-J and BAWE-EN*

Corpus	FLOB-J		BAWE-EN	
Bundle	*the degree to which* (5)* *the extent to which* (6)		*the extent to which* (8)	
	the fact that this (4)		*the fact that the* (8) *the fact that they* (4)	
	the way in which (14) *the ways in which* (4)		*the way in which* (7)	
Total	type	5	type	4
	token	33	token	27

* The raw frequency is indicated in brackets, and this practice is used throughout this paper.

Secondly, a great number of NP + *of* and PP + *of* bundles can be grouped into two productive frames: "*the* + Noun + *of the/a*," and "*in the* + Noun + *of*." The professional writing in FLOB-J manifests a relatively wide range of nouns that collocate with these two frames (Table 7 and Table 8). In this regard, it appears that the patterns emerging from FLOB-J lend support to the finding reported by Biber et al. (2003), who described the same two "fixed frames" (termed "phrase-frame" by Stubbs, 2007a) used for 43 and 17 different lexical bundles respectively in their academic prose as "extremely productive frames" (Biber et al., 2003, p.78). In comparison, neither the British students nor the Chinese students seem to have recognized the importance of these nominal or prepositional expressions in their academic writing.

Table 7. *The Frame for "the + Noun + of the/a"*

	the + Noun + *of the/a*	Total	
		type	token
FLOB-J	**end** (10), creation (4), existence (4), history (7), impact (4), magnitude (4), results (4), **nature** (17), **rest** (11), **role** (5), rules (5), **size** (7), status (4), strength (5), **structure** (4), value (5)	16	100
BAW-EN	development (11), **end** (10), length (6), **nature** (7), quality (4), **rest** (12), **size** (4), **structure** (6), use (9)	9	69
BAWE-CH	development (4), **end** (4), importance (4), **nature** (5), **rest** (8), **role** (4), **size** (4), top (4)	8	37

* The bundles appearing in two or three corpora are indicated in bold. This practice is used throughout this paper.

Table 8. *The Frame "in the + Noun + of"*

	in the + Noun + *of*	Total	
		type	type
FLOB-J	**absence** (7), **case** (19), **context** (19), course (5), face (4), **form** (8), hands (5), light (6), number (6), presence (8)	10	87
BAWE-EN	**absence** (4), **case** (23), **form** (8)	3	35
BAWE-CH	**case** (10), **context** (5), **form** (4)	3	19

As seen in Table 5, both groups of student writing generally had more VP-based bundles than native expert writing, and this tendency is particularly marked in certain subcategories. For example, we found that the student writers in BAWE-EN and BAWE-CH used considerably more "*to*-clause fragments" (see Table 9), showing a preference for the frame "*in order to* + Verb." L1 Chinese students in particular used six different verbs that fit in the slot: *achieve*, *avoid*, *be*, *maintain*, *make* and *understand*, while British students had two such bundles: *in order to make* and *in order to minimise*. For this subcategory, we see more similarity between BAWE-EN and BAWE-CH.

Table 9. *Bundles in the Subcategory of "to-Clause Fragments"*

Corpus	FLOB-J		BAWE-EN		BAWE-CH	
Bundle	**to be able to** (5)		*in order to make* (8)		*in order to achieve* (8)	
			in order to minimise (4)		*in order to avoid* (7)	
			to be able to (8)		*in order to be* (5)	
			to be added to (4)		*in order to maintain* (4)	
			to cope with the (4)		***in order to make*** (4)	
			to enable them to (4)		*in order to understand* (4)	
			to take into account (4)		***to be able to*** (4)	
					to ensure that the (4)	
Total	type	1	type	7	type	8
	token	5	token	36	token	40

Although there is a substantial number of VP-based bundles in BAWE-CH, L1 Chinese students did not use the "Passive verb + prepositional phrases" (PassPP) form as frequently as native speakers did. As can be seen in Table 10, there are seven passive-verb bundles in FLOB-J and eleven in BAWE-EN, both of which make up around 20% of the VP-based bundle types within each individual corpus. In comparison, the four passive-verb bundles in BAWE-CH constitute merely 10% of the total VP-based bundle types. Additionally, none of the four passive bundles were shared by either of the native group of writers.

Table 10. *Bundles in the Subcategory "Passive Verb + Prepositional Phrases"*

Corpus	FLOB-J		BAWE-EN		BAWE-CH	
Bundle	*are shown in fig* (6)		*be seen as a* (5)		*can be divided into* (4)	
	be found in the (5)		*be included in the* (4)		*can be explained by* (7)	
	be seen in the (4)		**be taken into account** (5)		*can be regarded as* (4)	
	be taken into account (5)		*be used in the* (5)		*is illustrated in figure* (4)	
	can be found in (6)		*can be applied to* (7)			
	is concerned with the (4)		**can be found in** (6)			
	was followed by a (4)		*can be seen as* (5)			
			can be seen in (4)			
			can be used for (5)			
			could be seen as (5)			
			should be placed on (4)			
Total	type	7	type	11	type	4
	token	34	token	55	token	19

Discourse Functions

The functional categorization adopted here follows the taxonomy devised by Biber and colleagues (Biber & Barbieri, 2007; Biber et al., 2003, 2004). Three major categories were distinguished: referential bundles, stance bundles, and discourse organizers.

Referential expressions are characterized by the function of attribute specification. The first type, framing bundles, are used to specify a given attribute or condition (e.g., *in terms of the*). Another common type of referential bundles is quantifying expressions (e.g., *per cent of the*), which qualify a proposition with expressions related to anything potentially measurable, such as size, number, amount or extent. The last subcategory of referential expressions includes place/time/text-deictic bundles (e.g., *at the beginning of*).

- **Framing***: in the context of, the nature of the, the existence of a*

- **Quantifying:** *a wide range of, the extent to which, in a number of*

- **Place/time/text-deictic***: are shown in fig, at the same time*

Stance bundles are often used to express a writer's evaluation of a proposition in terms of certainty or uncertainty (epistemic) (e.g., *seems to have been*). They can also convey the writer's attitude about proposition (obligation/directive) (e.g., *it is important to*). If the writer's judgment on the ability to do something is involved, then they are grouped under "ability" (e.g., *will be able to*).

- **Epistemic:** *are more likely to, it can be argued, the fact that the*

- **Obligatory/directive:** *it is necessary to, that need to be, it has to be*

- **Ability:** *it is difficult to, to be able to*

Discourse organizers are used to structure texts. They can introduce a topic (e.g., *essay is going to*), elaborate on the topic (e.g., *be taken into account*), or make inference (e.g., *in the sense that*). In addition, a large number of the discourse organizers discovered here function to identify the focus that the writer is making (e.g., *bear in mind that*).

- **Topic introduction:** *essay is going to, last but not least, in this essay I*

- **Topic elaboration:** *in more detail in, on the other hand, can be used to*

- **Inferential:** *as a result of, in view of the, this is due to*

- **Identification/focusing:** *one of the most, there would be no, we can see that*

As can be seen from Figure 1, FLOB-J contains a higher proportion of referential expressions (60%), whereas they are much less frequent in both BAWE-EN (37%) and BAW-CH (41%). On the other hand, discourse organizers rank as the largest category in both BAWE-EN and BAWE-CH, having very similar proportions at 39% and 42% respectively, while discourse organizers in FLOB-J make up only about half of that (21%). As for stance bundles, BAWE-EN has the highest percentage of use at 24%, but this category is the smallest one in each of the three corpora.
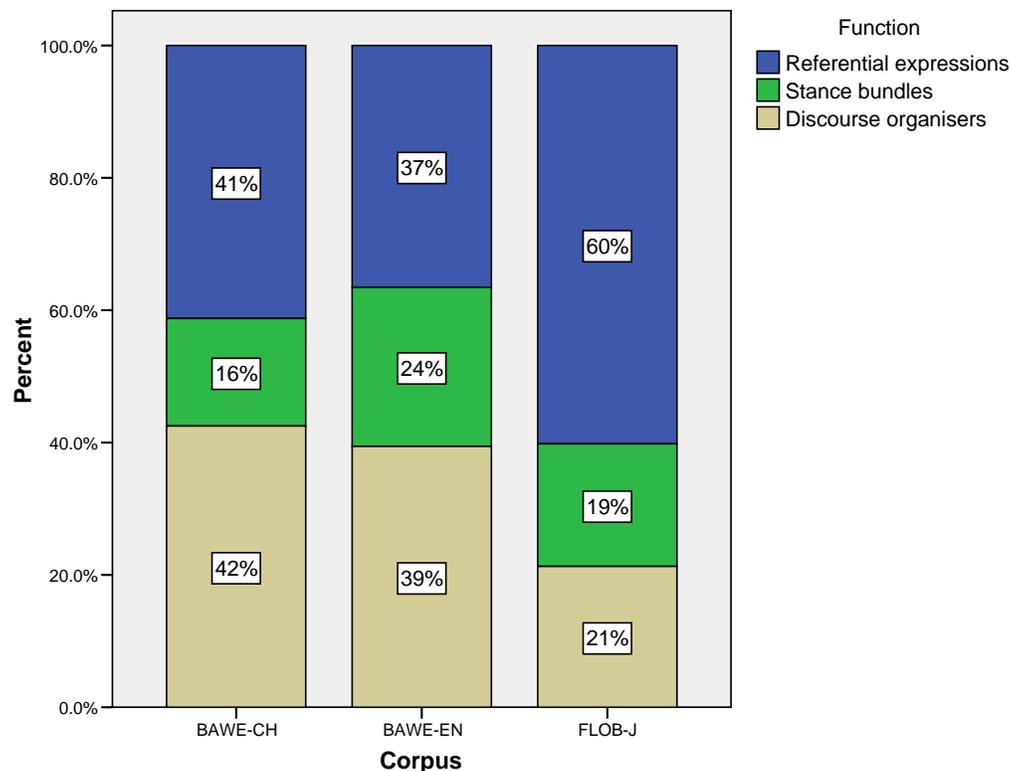

Figure 1. Functional distribution (types).

A chi-square test indicates that there is significant difference, in terms of the functional distribution of bundle types, between BAWE-CH, BAWE-EN and FLOB-J at the 0.05 level ($\chi^2 = 16.4$, *df* = 4, *p* = 0.003, Cramer's *V* = 0.167).[3] The standardized residuals (R), a cell-by-cell comparison of observed and expected counts, were calculated to identify the cells that made a major contribution to the significant difference.

As can be seen from Table 11, only two cells, the referential expressions and discourse organizers in FLOB-J, have an absolute value of R greater than 1.96, which suggests that these two categories in FLOB-J made a statistically significant contribution to the rejection of the null hypothesis. We can use the information from the values of R to conclude that there are significantly more referential expressions and fewer discourse organizers in native academic writing in comparison with academic student writing.

Table 11. *Standardized Residuals in a Chi-Square Contingency Table for Functional Distribution (Types)*

$X^2 = 16.4$, $df = 4$, $p = 0.003$ Cramer's V=0.167		Referential expressions	Stance bundles	Discourse organizers
FLOB-J	Observed Count	65	20	23
	Expected Count	50.3	21.5	36.2
	R	**2.1**	-0.3	**-2.2**
BAWE-EN	Observed Count	38	25	41
	Expected Count	48.4	20.7	34.9
	R	-1.5	0.9	1.0
BAWE-CH	Observed Count	33	13	34
	Expected Count	37.3	15.9	26.8
	R	-0.7	-0.7	1.4

The token distribution of functions among the three corpora is virtually the same as for type distribution. As can be seen in Figure 2, the proportion of referential expressions remains the most marked difference between FLOB-J, BAWE-EN and BAWE-CH, as referential expressions make up almost two thirds of the bundles in FLOB-J. On the other hand, both BAWE-EN and BAWE-CH rely more heavily on discourse organizers, having proportions as high as 39% and 48% respectively.

A chi-square test indicates that there is significant difference, in terms of the functional distribution of bundle tokens, among the three groups at the 0.05 level ($\chi^2 = 148.5$, $df = 4$, $p < 0.0005$, Cramer's $V = 0.199$). The standardized residuals were again calculated. As can be seen from Table 12, apart from the stance bundles in FLOB-J, every cell in this contingency table contributed significantly to the differences. On the basis of the information provided by R, the referential expressions and discourse organizers in FLOB-J are still found to make the most contribution to rejecting the null hypothesis, just like the type distribution. On the whole, there are significantly more referential expressions and fewer discourse organizers in native expert writing, while both groups of student writing contain significantly fewer referential expressions and more discourse organizers. In addition, the British students, represented by BAWE-EN, used more stance bundles than expected, whereas the Chinese students in BAWE-CH used fewer stance bundles.

Drawing on the standardized residuals from functional analysis (Table 11 and Table 12), the FLOB-J corpus appears to represent the group which differs the most from the other two groups of university student writing. Given that the texts retrieved from FLOB-J are published academic texts, written by native academics, and must therefore have been repeatedly edited by experienced editors, it is not too surprising to see that FLOB-J distinguishes itself among the three groups of writing. The similarities between BAWE-EN and BAWE-CH revealed by the standardized residuals also meet with our expectations to a certain extent. The student writing in both BAWE-EN and BAWE-CH was produced by university students, who can be regarded as novice academic writers. In addition, both groups of student writing were originally extracted from the same BAWE corpus, although it should be born in mind that the topics for each piece of assignment varied to a very large degree in these two student subcorpora, covering many disciplines. It should be noted also that the text types and constituents in the FLOB-J and

BAWE subcorpora might have had an impact on the analysis, and this will be discussed further in the section of Discussion.
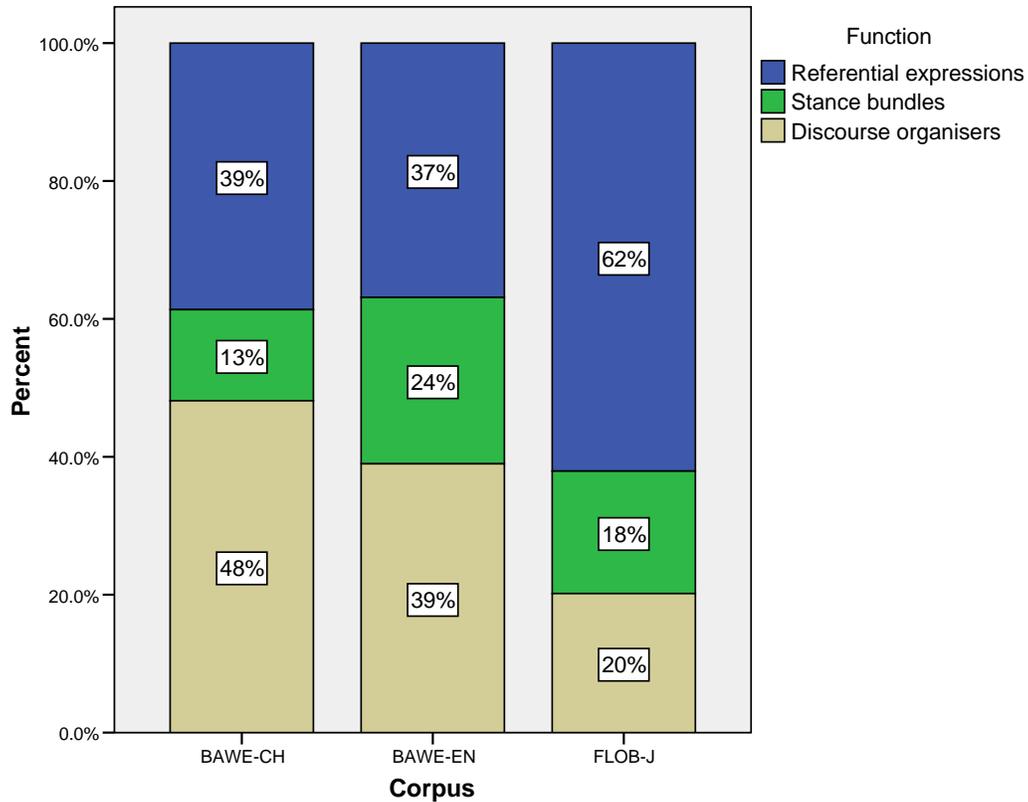

Figure 2. Functional distribution (tokens).

Table 12. *Standardized Residuals in a Chi-Square Contingency Table for Structural Distribution (Tokens)*

$X^2 = 148.5$, $df = 4$, $p < 0.0005$ Cramer's V=0.199		Referential expressions	Stance bundles	Discourse organizers
FLOB-J	Observed Count	437	125	142
	Expected Count	329.5	132.3	242.2
	R	**5.9**	-0.6	**-6.4**
BAWE-EN	Observed Count	246	161	260
	Expected Count	312.2	125.4	229.4
	R	**-3.7**	**3.2**	**2.0**
BAWE-CH	Observed Count	196	67	244
	Expected Count	237.3	95.3	174.4
	R	**-2.7**	**-2.9**	**5.3**

As revealed in the type distribution (see Figure 1 and Table 11) and token distribution (see Figure 2 and Table 12), we already know that referential expressions are highly frequent in expert academic writing, whereas university students do not rely on this discourse function as much. Among the referential expressions, one type of quantifying bundle is noteworthy (i.e., the extent/degree modifiers, which are

present in both groups of native writing, but not in learner writing). There are four such bundles in FLOB-J: *in so far as* (6), *the degree to which* (5), *the extent to which* (6), and *to a large extent* (4), and two in BAWE-EN: *the extent to which* (8), and *to a certain extent* (4). It appears that learners do not use this type of modifier very much, whereas native speakers tend to use them to modify the extent or degree of their proposition as the following examples demonstrate:

- ```
  No matter what the nature of the being, the principle of equality
  requires that its suffering be counted equally with the like suffering -
  in so far as rough comparisons can be made - of any other being. (FLOB-J)
  ```

- ```
  Thus even though when one entertains in commercial setting aspects of
  intimacy can work well to a certain extent. (BAWE-EN)
  ```

On the other hand, Chinese student writers seem to use certain referential deictic expressions, such as *in the long run* (13), *in the recent years* (6), and *all over the world* (6), as exemplified below. These deictic expressions do not appear in the repertoire of word combinations used by professional writers nor British peer students.

- ```
  Almost all economists today agree that monetary policy influences
  unemployment, at least temporarily, and determines inflation, at least in
  the long run. (BAWE-CH)
  ```

- ```
  They are more or less equivalent way of paying out retained earning,
  while stock repurchases indeed have become an important source of payout
  in the recent years. (BAWE-CH)
  ```

- ```
  This strategy is now very popular all over the world, for it maximizes
  the value of limited monetary amount of fringe benefits and gives the
  employees some controls over their own rewards. (BAWE-CH)
  ```

The first word combination, *in the long run*, is an idiomatic expression, occurring 13 times in BAWE-CH but only once in FLOB-J. This idiom, *in the long run,* is actually more characteristic of non-academic text than of academic prose, and is quite frequent in speech, as indicated by the British National Corpus (BNC),[4] albeit not always being identified as an informal expression in dictionaries (e.g., Macmillan Dictionary, Rundell, 2007). The second bundle, *in the recent years*, was generally expressed as *in (more) recent years* and *recently* by native writers in FLOB-J and BAWE-EN. Interestingly, we found 2,344 instances of *in recent years* and only 2 instances of *in the recent years* in the BNC. This suggests that *in the recent year* is therefore a "learner bundle" rather than a "native bundle." The third expression, *all over the world*, might reflect a general tendency of learners to be categorical and to over-generalize as this expression appears to be favored by learners at various proficiency levels (Chen, 2009).

Turning now to stance bundles, it was found that the supposedly least-competent writers, represented by the L2 writers in BAWE-CH, employed the smallest range of epistemic bundles, whereas the most proficient writers in FLOB-J manifested the widest range of epistemic expressions. Further investigation of the epistemic markers used by the native writers shows that both native groups are quite capable of taking advantage of comprehensive measures to hedge their statements. The frame "copula *be* + *likely to*" is frequently used in native writing to mitigate a proposition, with a few variations such as *is likely to be* (7),[5] *are likely to be* (9), *are more likely to* (13). In addition to this frame, native writers are also capable of flexibly employing other hedging devices, including the "Anticipatory *it* + adjective fragment" frame (*it*

is clear that $_{(19)}$, *it is not clear* $_{(4)}$, *it is possible to* $_{(6)}$), modal verbs (*would have to be* $_{(12)}$, *would need to be* $_{(4)}$, *would be difficult to* $_{(5)}$), hedging verbs (*seems to have been* $_{(6)}$, *it has been suggested* $_{(4)}$, *it can/could be argued* $_{(19)}$, *it is estimated that* $_{(4)}$), and hedging nouns (*there is no evidence* $_{(4)}$, *there is evidence that* $_{(5)}$, *the fact that the* $_{(8)}$, etc.) to qualify their propositions.

- This change indicates that two relatively dissimilar clusters have been merged and that the number of clusters prior to this merger **is likely to be** the most appropriate. (FLOB-J)

- If activists actions are justified **it could be argued that** firms should withdraw from the market because they are acting unethically. (BAWE-EN)

By contrast, there are only four bundles in the L2 writing that can be regarded as hedging expressions: *are more likely to* $_{(5)}$, *is considered to be* $_{(4)}$, *it has been suggested that* $_{(6)}$, *it is believed that* $_{(5)}$.

Both British and Chinese students used a relatively high number of discourse organizers in their writing when compared to the academic prose in FLOB-J. In particular, they used more discourse organizers to elaborate and/or clarify a topic, the majority of which are VP-based bundles, such as "Passive verb + prepositional phrase fragment" (*can be regarded as*, *be included in the*, etc.), "Verb + *to*-clause fragment" (*can be used to*, *in order to make*, etc.), and "Subject + verb" (*this means that the*, *that is to say*).

- An example **can be used to** clarify the theory. (BAWE-CH)

- According to the report by Mintel International Group Limited (2004), consumers spending on cars grew in year 2001 excess of 10% due to the drop in prices as a result of pressure from the government. **This means that the** industry do compete on pricing among other things. (BAWE-CH)

- A study on domestic tourism by National Council of Applied Economics Research during 2002-2003 pointed out that nearly two third of all tourists in India traveled for social purpose (Social-cultural Drivers); **that is to say**, traveling for social purposes, overall, stands the largest percentage of trips across the country… (BAWE-CH)

An impression of the instances above is that they all seem to be rather verbose. The most noticeable example of tautology might be the last one, which repeatedly refers to travelling for social purposes in India, using various paraphrases. The contrast with *that is to say* in the professional academic writing below demonstrates one of the major differences between L1 expert writing and learner writing. In the following example, by use of the expression *that is to say*, the native academic does not simply paraphrase what has already been written as learners do, but instead progresses further, using other means (e.g., giving a specific example) to illustrate the previous proposition.

- It is now accepted on all sides that Britain needs more of its workforce to be vocationally trained to intermediate levels; **that is to say**, to craft or technician standards as represented, for example, by City and Guilds examinations (at part 2) or BTEC National Certificates and Diplomas. (FLOB-J)

DISCUSSION

The analysis in the previous sections set out to compare the use of recurrent word combinations, in terms of their structures and functions, in native expert writing, native student writing and L2 student writing. A deeper investigation, however, suggested that the quantitative analysis needed to be complemented and supported by qualitative analyses which considered an examination of expanded concordance lines. By utilizing such a hybrid methodology, a number of distinctive features, which vary according to level of writing proficiency, have been unveiled.

L2 academic writing has been found to be stylistically more verbose (cf. Lorenz, 1998, 1999) and to show less control of cautious language (cf. Hyland, 1994; Hyland & Milton, 1997). Consider the use of hedging in cautious language for example. L1 Chinese learners of L2 English in the current study are found to show some control of this feature in their academic writing, but do not demonstrate it as diversely and robustly as native writers do. Indeed, Hyland and Milton (1997) compared expressions for qualification and certainty in the writing of L1 and L2 students and found that Chinese students in Hong Kong in particular had some problems in this pragmatic area. They concluded that this could be partly attributed to a lack of introduction of hedging devices in EAP textbooks. Another aspect relating to L2 writers' underuse of hedging devices is their tendency to be categorical and to over-generalize. As Ringbom (1998) discovered, even at advanced level, learner language was still in some respects more, in others less, *vague* than native speaker language, although this was a word-based perspective rather than a phraseological one. Investigating learners' writing development using IELTS candidate scripts across band scores, Kennedy and Thorp (2007) also pointed out that L2 learners at lower proficiency levels tend to express their opinions in a more categorical manner, and that their writing is modified less by hedging. The finding here, therefore, reinforces this distinctive aspect of L2 writing from a phraseological viewpoint. The tendency to hedge less and instead adopt an overstating tone seems to be universal for learners from different L1 backgrounds, as the studies discussed above are not exclusive to L1 Chinese learners of L2 English. What is more, it appears that these features may change with proficiency development, as evidenced by Kennedy and Thorp (2007). Learner writing is likely to improve as proficiency progresses, most likely by edging closer to the norms of native expert writing and showing better control of cautious language.

Another interesting issue is the relationship between the number of recurrent word combinations and writing proficiency. As shown in Table 4, the number of recurrent word combinations increases with advancing writing proficiency, which is the case both for the range of lexical bundles used (types), and the overall occurrence of lexical bundles (tokens). It appears that the use of formulaic expressions grows with writing proficiency. This finding is, nonetheless, contrary to some of the results reported in the literature (De Cock, 2000; Hyland, 2008a). It should be noted that these studies did not remove overlapping bundles or context-dependent ones, while the current research does. Take Hyland's study (2008a) for example. He compared academic clusters among published research articles, PhD dissertations, and Master theses. In his conclusion, Hyland indicated that the *least* confident or proficient students at Master's level relied on formulaic expressions most, while the expert writers used the fewest clusters. Comparisons across studies like these, however, need to exercise extreme caution. Firstly, Hyland included all the topic-related clusters occurring in his study (e.g., *in the Hong Kong*), while such context-dependent bundles are excluded in the present paper. Next, our repeated experiments have revealed that the number of recurrent word combinations retrieved might relate to corpus size to a large extent. On the whole, larger corpora will generate fewer recurrent word combinations with the same cut-off normalized frequency, when compared with smaller corpora, because large corpora will elicit higher converted raw frequencies, as discussed in the section on Operationalization. Furthermore, the dispersion requirement (e.g., occurring in at least three texts or 10% of texts) also impacts on the number of recurrent word combinations. It is virtually impossible to find different corpora, of exactly the same size composed of the same number of texts, for direct comparison. For cross-study comparisons, we have to

bear these limitations in mind. As a result, it is still not conclusive as to whether there is a relationship between proficiency and the number of formulaic expressions used, particularly when the student groups are not identical, as in Hyland (2008a) and the current study. Interestingly, the results from the current study are in line with those in De Cock's study (2004), in which she compared recurrent word combinations between native and non-native speech. She found that, after discarding the repeats (e.g., *I I* or *the the*) and the hesitation items (e.g., *er* or *erm*), native data actually contain more recurrent word combinations of different lengths than non-native speech does.

It has to be acknowledged that the use of FLOB-J to represent native expert academic writing might have had some impact on the word combinations derived. First of all, a large proportion of the texts included in FLOB-J are hard-science based. This is probably why we found bundles such as *a function of the*, *the magnitude of the*, *the structure of the*, and *a high level of* in FLOB-J, which appear to be strongly concerned with the disciplines of hard science. Meanwhile, the journal papers or book sections selected in FLOB-J are all 2000-word long excerpts, rather than the complete texts included in the BAWE student-writing corpus. It is probable that there are more occasions in BAWE student writing to use discourse organizers, as student essays are mostly structured as Introduction, Body and Conclusion. However, it should also be noted that when examining concordance lines, we found very few discourse organizers which could be attributed to the differences between excerpts and full texts (i.e., topic-introduction bundles such as *in this essay I* or *last but not least*). In addition to the use of FLOB-J, clearly there have been other constraints on the present study. For one, this corpus-driven approach cannot cater for discontinuous word combinations, and thus certain information might be missing. For another, it is notoriously difficult to obtain large quantities of quality learner data, and the learner writing investigated in this paper is not error-tagged. We cannot know for sure if there are any learner errors which might have affected the generation of word combinations, although these assignments have been assessed as being good university essays.

CONCLUSION

This comparative study has revealed the fundamental differences and similarities between native and learner academic writing. Through structural and functional comparisons, it has been found that the use of lexical bundles in non-native and native student essays is surprisingly similar. They both contain many more VP-based bundles and discourse organizers than native expert writing does, which appears to be a sign of immature writing. On the other hand, native professional writers exhibit a wider range of NP-based bundles and referential markers. A further qualitative examination revealed, however, that native student writing actually shares a few features distinctive in academic writing, such as the control of cautious language in native professional writing. Non-native writing, however, demonstrates a tendency that seems to be exclusive to L2 writing (e.g., over-generalizing and favoring certain idiomatic expressions and connectors).

With the development of corpus techniques, the importance of corpus-extracted word combinations as building blocks in constructing discourse has been increasingly recognized. However, the growing interest in identifying phraseology with corpus tools during the past decade does not appear to have encouraged ELT publishers or practitioners to put more emphasis on computer-retrieved formulaic language in the curriculum and/or materials. In the current study, through investigation of three groups of academic writing, it was found that there was a gap, in terms of the use of lexical bundles, between native expert academic writing and university student writing (native and non-native alike). We argue that, after careful selection and editing, the frequency-driven formulaic expressions found in native expert writing can be of great help to learner writers to achieve a more native-like style of academic writing, and should thus be integrated into ESL/EFL curricula.

NOTES

1. All the frequencies of bundles indicated in this study are raw frequencies rather than normalized ones.

2. In LSWE, the data of academic prose is as large as 5.3 million words.

3. The statistical package used is SPSS 17.0 (2008).

4. In the BNC, for academic writing, the frequency per million words of *in the long run* is 6.72. This figure is 8.27 for non-academic prose and 4.23 for speech.

5. For reasons of space, the frequencies in brackets are the sum of both FLOB-J and BAWE-EN.

ACKNOWLEDGEMENTS

ABOUT THE AUTHORS

Yu-Hua Chen completed her PhD at Lancaster University in 2009, and this paper is part of her doctoral research. Her research interests include corpus linguistics, language testing, phraseology, second language writing, and computer-assisted language learning.

E-mail: cyuhua@ntu.edu.tw

Paul Baker is a senior lecturer in the Department of Linguistics and English Language at Lancaster University. He is the commissioning editor of the journal Corpora. His recent books include Using Corpora for Discourse Analysis (2006), Sexed Texts (2008), and Sociolinguistics and Corpus Linguistics (2010). He has recently built the BE06—a one-million word reference corpus of written British English sampled circa 2006, using the same model as the Brown corpus.

E-mail: j.p.baker@lancaster.ac.uk

REFERENCES

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora, 4*(1), 71–83.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 101–122). Oxford: Oxford University Press.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263–286.

Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversations and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* (pp. 71–93). Frankfurt: Peter Lang.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Chen, Y.-H. (2009). *Lexical Bundles across Learner Writing Development.* Unpublished doctoral thesis, Lancaster University, Lancaster, UK.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89.

Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). Amsterdam: John Benjamins Publishing Company.

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

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59–80.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 51–68). Amsterdam: Rodopi.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgium Journal of English and Literatures (BELL)*, New Series *2*, 225–246.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67–79). London: Longman.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & Meunier (Eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Hundt, M., Sand, A., & Siemund, R. (1998). *Manual of Information to accompany The Freiburg-LOB Corpus of British English ('FLOB')*. Retrieved from http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purpose, 13*(3), 239–156.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183–205.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*, 433–445.

Kennedy, C., & Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor & P. Falvey (Eds.), *IELTS collected paper: research in speaking and writing assessment* (pp. 316–378). Cambridge: Cambridge University Press.

Lorenz, G. (1998). Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on Computer* (pp. 53–66). New York: Addison Wesley Longman Limited.

Lorenz, G. (1999). *Adjective intensification--Learners versus native speakers. A corpus study of argumentative writing*. Amsterdam: Radopi.

Meunier, F., & Granger, S. (Eds.). (2007). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins Publishing.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English. In S. Granger (Ed.), *Learner English on Computer* (pp. 41–52). London: Addison Wesley Longman Limited.

Rundell, M. (Ed.). (2007). *Macmillan English Dictionary For Advanced Learners (Second Edition)*. Oxford: Macmillan Education.

Schmitt, N. (2004). *Formulaic sequences: acquisition, processing, and use*. Amsterdam: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127–152). Amsterdam: John Benjamins Publishing.

Scott, M. (2007). Oxford WordSmith Tools (Version 4.0) [Computer software]. Oxford: Oxford University Press.

SPSS for Windows (2008), Rel. 17.0.0. [Computer software]. Chicago: SPSS Inc.

Stubbs, M. (2007a). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 89–105). Amsterdam: Radopi.

Stubbs, M. (2007b). Quantitative data on multi-word sequences in English: The case of word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, Discourse and Corpora: Theory and Analysis* (pp. 163–189). London: Continuum.

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

Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

APPENDIX. Lexical Bundles in Frequency Order

BAWE-CH		BAWE-EN		FLOB-J	
on the other hand	36	in the case of	23	in the context of + (the/a)	19
(and) + at the same time	24	can be used to	19	in the case of	19
that there is a/an	17	as a result of + (the)	17	on the other hand	19
as well as the	16	it is important to	17	the nature of the	17
in the long run	13	it could be argued that + (the)	14	as a function of	15
as a result of	12	(at) + the end of the	13	on the basis of	14
one of the most	11	in terms of the	13	in terms of the	14
can be used to	10	(for) + the development of the	13	it is necessary to	14
in the case of	10	it is possible to	13	the way in which + (the)	14
is one of the	9	is one of the	12	(at) + the end of the	13
it is difficult to	9	the rest of the	12	it is clear that	11
as one of the	9	it can be seen + (that)	12	the rest of the	11
the rest of the	8	one of the main	11	one of the most	10
in order to achieve	8	as well as the	10	at the same time	10
at the end of + (the)	8	the use of the	9	by the fact that	9
it is necessary to	8	in the same way	9	a wide range of	9
in order to avoid	7	to be able to	8	as a result of	9
in the end of + (this)	7	(due) + to the fact that	8	that there is a	9
we can see that	7	an example of this + (is)	8	per cent of the	9
can be explained by	7	in the form of	8	in the form of	8
as long as the	6	it is clear that	8	on the one hand	8
at the beginning of	6	the fact that the	8	the extent to which	8
a number of factors	6	in order to make	8	in the sense that	8
all over the world	6	the nature of the	7	would have to be	8
it is easy to	6	one of the most + (important)	7	as we have seen	8
to the development of	6	and as a result	7	in the presence of	8
at the expense of	6	it is necessary to	7	in the absence of	7
in the recent years	6	the way in which	7	is likely to be	7
it has been suggested + that	6	can be applied to + (the)	7	the size of the	7
that is to say	6	than that of the	7	are more likely to	7
are more likely to	5	and the use of	6	as we shall see	7
as part of a	5	are more likely to	6	that there is no	7
on the basis of	5	not be able to	6	the history of the	7
bear in mind that	5	the extent to which	6	the turn of the century	7
in order to be	5	was one of the	6	can be found in	6
in terms of the	5	could be used to	6	in the first place	6
last but not least	5	is an example of + (a)	6	it is difficult to	6
the nature of the	5	it is difficult to	6	it is possible to	6
(played) + an important role in	5	the structure of the	6	of a number of	6
as a part of	5	can be found in	6	on the part of	6
in the context of	5	is the fact that	6	in so far as	6
it is believed that	5	the length of the	6	in the number of	6
a wide range of	4	with respect to the	6	it is important to	6
as part of the	4	would have to be	6	seems to have been	6
can be divided into	4	can be seen as +	5	with respect to the	6
can be regarded as	4	+ be seen as a	5	are shown in fig	6
for the development of	4	at the same time	5	can be used to	6
in addition to the	4	be taken into account	5	in the light of	6
in order to maintain	4	be used in the	5	a function of time	6
in order to make	4	could be seen as	5	as shown in fig	6
in order to understand	4	it would have been	5	be taken into account	5
is illustrated in figure	4	this is due to + (the)	5	for each of the	5
is not only a	4	with the development of	5	in a number of	5
it can be seen	4	would be able to	5	it can be seen + that	5
it is important to	4	are likely to be	5	to the fact that	5
the development of the	4	can also be used + (to)	5	would be difficult to	5
the importance of the	4	in relation to the	5	at the time of	5
the size of the	4	in terms of its	5	be found in the	5
the top of the	4	it can be argued + (that)	5	in the hands of	5
this means that the	4	of the number of	5	the degree to which	5
to be able to	4	through the use of	5	the role of the	5
to ensure that the	4	to a lack of	5	the rules of the	5
to the fact that	4	can be used for	5	the strength of the	5
will focus on the	4	for each of the	5	the value of the	5
with respect to the	4	there would be no	5	to be able to	5
with the introduction of	4	a great deal of	4	a function of the	5

a high level of	4	an integral part of	4	in the course of	5
a large number of	4	as part of the	4	there is evidence that	5
as soon as the	4	by the presence of	4	whether or not to	5
essay is going to	4	in an attempt to	4	a large number of	4
in the form of	4	is by no means	4	an example of this	4
is considered to be	4	it is estimated that	4	be seen in the	4
it has to be	4	it should be noted	4	by a variety of	4
it is easy for	4	of some of the	4	in contrast to the	4
meet the requirement of	4	on the other hand	4	in more detail in	4
must be able to	4	should be able to	4	in relation to the	4
of the number of	4	the fact that they	4	in terms of a	4
pay more attention to	4	there is no evidence	4	in view of the	4
the role of the	4	this means that the	4	it has been suggested	4
this is due to	4	to a certain extent	4	it has not been	4
		to be added to	4	it is not always	4
		to enable them to	4	on a number of	4
		to take into account	4	the fact that this	4
		would need to be	4	the right hand side	4
		as a way of	4	the status of the	4
		at the heart of	4	the structure of the	4
		be included in the	4	the ways in which	4
		because it is not	4	to a large extent	4
		can be seen in	4	a high level of	4
		for the use of	4	are likely to be	4
		in order to minimise	4	at each end of	4
		in the absence of	4	at the beginning of	4
		in this essay I	4	end of the spectrum	4
		should be placed on	4	has a number of	4
		taking into account the	4	in the face of	4
		that is to say	4	is concerned with the	4
		that need to be	4	it is not clear	4
		the quality of the	4	the creation of a	4
		the size of the	4	the existence of a	4
		this may be due + to	4	the impact of the	4
		to cope with the	4	the magnitude of the	4
		will be able to	4	the point of view	4
		will be used to	4	the results of the	4
		with the addition of	4	the second half of	4
				to that of the	4
				was followed by a	4
				was not so much	4
				was one of the	4