

A Corpus Study of Structural Types of Lexical Bundles in MUET Reading Texts

CHRISTINA ONG SOOK BENG
Universiti Tunku Abdul Rahman
Malaysia
ongsb@utar.edu.my

YUEN CHEE KEONG
School of Language Studies and Linguistics
FSSK, Universiti Kebangsaan Malaysia
Malaysia

ABSTRACT

In view of the recurrent issues concerning university students' inability to comprehend reading passages in their studies, texts used in a Malaysian examination for tertiary education were analysed. This study investigates the use of lexical bundles (LBs) in the reading passages of Malaysian University English Test (MUET). A specialised corpus of MUET test papers comprised of only the reading passages grouped into two main disciplines namely arts and science was built. Besides identifying commonly used LBs, this study aims to compare and contrast structural types of LBs found in arts and science-based texts. Using WordSmith Tools version 5, the lists of LBs of the identified disciplines were generated. They were analysed qualitatively based on Biber, Conrad & Cortes' (2004) Structural Taxonomy. Findings revealed that the number of LBs in both disciplines differs significantly but many similar LBs are employed. It was also evident that science-based texts tend to employ more NP-based and VP-based LBs while the most commonly used structure in arts-based texts is dependent clause. In general, PP-based LBs are very significant in both science and arts-based texts. The extent to which LBs are specific to particular disciplines is therefore confirmed and an overview of how LBs in texts construct information is also obtained. Pedagogically, teachers should consider incorporating corpora-based material, to exploit consciousness raising tasks and not to emphasise too much on grammatical items so that LBs do not go unnoticed.

Keywords: lexical bundles; structural types; reading text; MUET; reading test

INTRODUCTION

Reading, according to Venezky (Lee 2004), is a psycholinguistic process combined with a socially and culturally embedded practice. Text comprehension has always been the main concern of reading. Pardo (2004) defined reading comprehension as “a process in which readers construct meaning by interacting with text through the combination of prior knowledge and previous experience, information in the text, and the stance the reader takes in relationship to the text” (p.272). In university, undergraduates have to comprehend texts beyond the printed words. Understanding the relationship between form and function, locutionary expression and illocutionary force as well as language and culture (Kern 2000) is vital because one of the measures of their academic ability is their maturity in comprehending texts presented to them. However, many undergraduates in Malaysia, as claimed by Noorizah (2006) and Nambiar (2007), were found to be unprepared for the reading demands expected of them.

This can be linked to the presence of unfamiliar words and expressions in academic texts which impedes second language learners' reading (Hyland & Tse 2009). The importance of lexical bundle (LB) – recurrent sequences of words – as a major component in

coherent linguistic production was demonstrated by Hyland (2008) and it is, according to Dontcheva-Navratilova (2012), an essential aspect of the shared knowledge of a professional discourse community. Without the knowledge at hand, LBs could not be utilised by students to aid their comprehension of any given text.

LBs are “sequences of two or more words frequently occurring in a particular setting” (Biber, Conrad & Cortes 2004, p.376). Neely & Cortes (2009) reiterated that LBs are ‘chunks’ of language at varying lengths. Thus, they can either be complete or incomplete grammatical units. Number of occurrences is another factor that distinguished LBs from other phrases (Dontcheva-Navratilova 2012, Laane 2011, Li 2008). Biber et. al. (2004) who was regarded as the proponents of LBs created a taxonomy of both structural and functional types. The structural classification especially has been widely relied on in the studies of LBs including this present one.

Insufficient knowledge of structural forms of LB which appears to be an obstacle in speaking tests has been proven in a few past studies. Read and Nation (2002) found that Band 8 achievers of the International English Language Testing Systems (IELTS) use substantially more LBs in the speaking test than those in the lower bands. One possible reason for this limited use of LBs could be that they have not been explicitly taught in class despite the claim that LBs are agents that determine the familiarity of language users in a particular discourse (Hyland 2008, Neely & Cortes 2009, Hyland & Tse 2009). Similarly in Malaysia, Nazira and Kamaruzaman (2009) revealed that exposing and routinizing authentic phrases have greatly benefited a group of diploma students undergoing a preparatory course for the speaking test of the Malaysian University English Test (MUET).

To investigate whether a similar trend is happening in reading tests, this study is conducted by analysing reading passages of MUET, a criterion-referenced test that gauges the overall English Language proficiency of candidates in the cumulative score of the four language skills (listening, speaking, reading and writing) in a single Band Score. As the focus of this study is to analyse reading texts adopted in MUET, the characteristics of the texts used must be identified. Malaysian Examination Council (2008) has summarised the features of the passages used in MUET test papers and they are as follows:

- i) Basic criteria for text selection:
Length (200 – 700 words), level of complexity (content and language), text type
- ii) Possible genres:
Articles from journals, newspapers and magazines, academic texts, electronic texts
- iii) Rhetorical style:
Analytical, descriptive, persuasive, argumentative, narrative

However, according to Zuraidah & David (2005), MUET may not contain all the varieties of genres in academic studies because the texts in MUET are sourced from a variety of reading materials while highly specialised texts are avoided.

Despite the fact that MUET may not contain all genres in academic studies, MUET is selected as the corpus for the current study for two reasons. Firstly, reading test in MUET compared to the combined tests of the other three skills; writing, speaking and listening in MUET, carries 40% of the total marks which means it is a significantly more important skill than that of the rest. Secondly, test papers such as MUET which had been validated are certified instruments used to draw conclusion about learners’ ability to read (Alderson 1990). Due to the recurrent issue on reading difficulty among university students which can be

attributed to the lack of knowledge on LBs and the emphasis past studies placed on LBs in spoken form, this study intends to:

- i) investigate LBs that are commonly used in MUET,
- ii) identify the structural categories of LBs in MUET, and
- iii) compare and contrast structural types of LBs in arts and science- based texts

As such, the concern of how undergraduates who have been found to be unprepared for reading demands of them can be addressed. However, before exploring further, the use of LBs in academic discourse and in specific disciplines has to be discussed.

LEXICAL BUNDLES IN ACADEMIC DISCOURSE

Multi-word expressions in academic prose according to Biber et al. (2004), often served to bridge two phrases, using for example a prepositional phrase (*on top of the*) or a noun phrase (*in the case of*). In other words, they function like scaffolding for new information (Biber & Barbieri 2007). The abundance of LBs in academic genres is undoubtedly obvious as Hyland (2008) revealed the occurrences of 3-word bundles over 60, 000 times and 4-word bundles over 5000 times per million words in academic prose. Biber et al. (2004) pioneering work in establishing taxonomy for the structural and functional categorization of LBs has led to many studies in academic discourse.

Studies on LBs in the written form are reviewed below as they are parallel with the present study which deals with reading passages. The use of LBs in academic writing has been abundant. According to Laane (2011) problems arose among the non-native writers as they may face the challenge of not finding the right words to express their ideas. She revealed the unsuitability of second personal pronoun in the academic context in the following examples (Laane 2011, p.72):

- a) *What you need to* know is that qZSIs suit for different renewable power applications.
- b) Instead, an impersonal lexical bundle is appropriate:
It is important to know that qZSIs suit for different renewable power applications.
- c) Another alternative can be:
As we have seen, qZSIs suit for different renewable power applications.

Dontcheva-Navratilova (2012) investigated diploma theses written by Czech students of English and found structural inaccuracy of bundles was rare but the distribution of functional categories differed considerably from the conventions of expert academic writing. Chen and Baker (2010), on the other hand, compared LBs retrieved from one corpus of published academic texts and two corpora of student academic writing (one L1, the other L2). They found that published academic texts exhibited a wider range of LBs while the use of LBs in L1 and L2 students' essays were similar. Pang (2010) demonstrated the essential role of LBs in academic writing to help L2 students to expand their repertoire of academic rhetorical features. Both studies by Chen and Baker (2010) and Pang (2010) have inevitably contributed to English language teaching pedagogy. The latter shared five pedagogical options namely, i) text analysis; ii) disciplinary ethnographies; iii) concept or semantic maps;

iv) writing sentences and; v) comparing registers, which she deemed effective in raising students' awareness of LBs (Pang 2010).

Despite the abundance of research on bundles in the written forms, to date, not many studies appear to have been undertaken in which the focus is on reading texts or shorter texts. Thus, research such as the present study is clearly needed.

DISCIPLINARY SPECIFIC BUNDLES

Analysing LBs in specific disciplines could provide the learners with a better understanding of the ways writers employed the resources of English in different academic contexts (Hyland 2008). Many past studies which demonstrated the degree of certain language patterns are specific to particular disciplines. To substantiate the fact that LBs are register-bound, three studies looking at occurrences of LBs in different disciplines are reviewed.

Using academic writing sub-corpora from two large corpora, Corpus of Contemporary American English (COCA) and The British National Corpus (BNC), Liu (2012), aimed to identify the most frequently-used multi-word constructions of various types (for instance, idioms, LBs, and phrasal verbs) and to examine their use patterns. The divisions of academic sub-corpus of COCA and BNC, made up of eight and six disciplines respectively were revealed before analysis took place. She discovered some differences in academic written multi-word constructions between American and British English (Liu 2012). For instance, in some related pairs (*as long as* vs. *as far as* and *in general* vs. *as a whole*), one is preferred by Americans and the other by the British. Bal (2010) compiled a corpus of published research articles produced by Turkish scholars in six different academic disciplines namely Economics, Education, History, Medicine, Psychology and Sociology prior to identifying frequently occurring four-word LBs. To prove that LBs functioned differently across disciplines, Hyland and Tse (2009) analysed research articles from Biology, Electrical Engineering, Applied Linguistics and Business studies. He detected that more research-oriented bundles were discovered in science and engineering texts whereas the Applied Linguistics and Business studies texts were dominated by text-oriented bundles.

All the above-mentioned studies corresponded with Hyland's (2008) assertion that bundles occurred and behaved differently in diverse disciplinary environments. Analysing LBs in a specific discipline is deemed more effective as it could yield more accurate results. Thus, this technique has been employed as the foundation of this study.

METHODOLOGY

To analyse commonly used LBs in MUET reading passages, all MUET reading test papers since its commencement in 1999 which amounted to 22 reading tests had been purchased for this study. Except cloze test which was part of 1999 to mid-2008 MUET reading test, all the other passages were used. A specialised corpus of MUET test papers made up of only the reading passages categorised into two main disciplines namely arts and science was built. Building such a small corpus has been proven significant by many (Green, Unaldi & Weir 2010, Romer 2004) because it can be used to systematically analyse the kind of language adopted in a particular context, in this case, an examination setting. It can also quantitatively study frequently occurring language patterns in reading texts which appeared to be insufficient as highlighted by Biber et.al (2004) as cited in Green et al. (2010).

An electronic corpus of MUET reading test papers, like any other computerized corpus as stated by Romer (2004), can be used to calculate frequencies of occurrence of single lexical items or cluster words (also known as LB in this study). The frequency lists are valuable as they identified words and multi-word items that are used commonly in MUET.

This was substantiated by Biber et al. (2004) who stressed on the usefulness of patterns marked by their frequency and the necessity to explain them which otherwise will go unnoticed. The creation of this corpus database underwent the 3-stage process adapted from Bahiyah, Mohd. Subakir, Kesumawati, Yuen and Azhar's (2008) research design. The process is as follows:

1. *Digitisation stage.* The purchased MUET test papers in the form of books were first transformed into digital form by scanning.
2. *Format conversion stage.* A .jpeg format of the books was produced after the scanning. This format was then converted to word document file and later into text files. The word document files contained both pictures and text data while the text files contained only text data. Clearing of the raw data was done to remove unnecessary items like graphics, instructions and multiple choice questions. The text files were then manually adjusted again to ensure accuracy and consistency.
3. *Merging stage.* During the scanning process, each section, for example, 1 MUET examination paper was split to a different file where 4 different passages with 3 pages were split to 4 text files. These different files were catalogued and merged; only then the data was ready for analysis.

The size of the finalised corpus for investigation is as follows. Although its size is relatively small, it exhibits the characteristics of a specialised corpus where it represents specific language use, it captures language features with small amount of data and it makes detailed qualitative data analysis more manageable (Lin 2008).

TABLE 1. Constituents of the MUET reading test corpus

Corpus	Word Count	Number of Texts
MUET	15,606	111

Using WordSmith Tools (WST) version 5, the numbers of LBs alongside their frequencies in the MUET reading test corpus were generated within the respective disciplines. It is a type of software which allows three types of analysis; generating wordlist, displaying the concordance of selected words and identifying keywords. However, the last function was not used in this study. The Wordlist tool can generate a list of all the words and word-clusters (two or more words) in a text, set out in alphabetical or frequency order (Scott, 2011). The parameters of LBs within which this study operated on were: i) length – two to six words; ii) frequency – 4 times in at least 3 texts. The LBs were then classified based on Biber et. al.'s (2004) Structural Taxonomy which is made up of four categories namely NP, VP, PP and Dependent Clause which are further divided into several sub-categories each. Although slight differences in the categorisation of LBs were noted in other studies, for instance, Biber and Barbieri (2007) and Hyland (2008), the presence of similar concepts in structural and functional categories was apparent. Biber et. al.'s (2004) taxonomy was adhered to instead of other taxonomies because it was regarded as among the earliest models dealing with LBs. Finally, analysis of LBs in relation to their structures was conducted.

FINDINGS

A total of 15,863 words were generated first followed by the word cluster list ranked in descending order with a total of 1,359 bundles. However, only 730 LBs consisting of 2, 3 and 4-word bundles were analysed because the remaining LBs especially 5-word bundles did not meet the predetermined cut-off frequency, hence, they were eliminated. Findings of this study are presented in three different parts based on the objectives of this study.

In answering research question one, a ranking of ten most frequently occurring LBs in the MUET reading test corpus extracted from the Wordlist generated by WordSmith Tools 5 is revealed as follows.

TABLE 2. Most frequent 2-word, 3-word and 4-word bundles in MUET reading text corpus

2-Word	Freq.	3-Word	Freq.	4-Word	Freq.
of the	388	one of the	22	in the United States	13
in the	296	the united states	20	at the University of	10
to the	138	per cent of	18	the end of the	10
it is	130	part of the	17	on the other hand	9
to be	113	be able to	16	will be able to	8
on the	102	more likely to	16	are more likely to	7
and the	99	in the united	15	as well as the	5
for the	91	it is a	15	for the first time	5
of a	84	in the world	13	a wide range of	4
in a	82	some of the	12	by the end of	4

Table 2 depicts 10 most frequent 2, 3 and 4-word bundles in general. Two word-bundles, *of the* top the list with 388 times and was found in 100 texts, followed by *in the* and *to the* appearing 296 and 138 times in 97 and 78 texts respectively. Occurring 22 times in 19 different texts was three word-bundles *one of the* which was also the most frequent three word-bundles. The second and third most frequent three word-bundles were *the United States* and *per cent of* appearing 20 and 18 times in 16 and 13 texts respectively. The four word-bundles were the least popular with only 13 occurrences of *in the United States*, 20 occurrences equally distributed between *the end of the* and *at the University of* appearing in 11, 8 and 6 texts respectively.

The second research question involved answering the question quantitatively where structural categorisation was made based on careful matching of all LBs identified with the sub-category terms used by Biber et al. (2004). LBs in which the structures did not correspond with meaning of any categories in the taxonomy were automatically disregarded. The numbers of occurrences for each sub-category alongside examples of structural LBs are shown in the following table.

TABLE 3. Number and example of structural categories of LBs in MUET reading passages

Structures	Sub-Categories	Examples
NP	NP + of (N=32)	number of, species of, per cent of, amount of, part of
	NP + post modifier (N=56)	the world, the disease, the government, the first time
	NP + be (N=12)	they are, there is, he was, we are, researchers are
VP	Passive + PP (N=31)	have been, can be, would be, will be, must be, based on

	be +N/AdjP (N=25) it + V/AdjP (N=11)	is not, are not, is also, be able to, was the, are more, it is, it has, it can, it is a, it is also, it could, it must be
PP	PP + of phrase (N=0) PP + NP (N=88)	- of the, in the, as the, of a, at the, according to, from the
Dependent Clause	VP + that clause (N=17) VP/Adj + to clause (N=55) Adverbial clause (N=30)	is that, found that, believe that, that the, that is, that they able to, have to, seem to, to be, more than, to make likely to, for example, not only, because they, in fact

As shown in Table 3, the three most frequently occurring structures of LBs are PP + NP, NP + post modifier and V/Adj + to clause, followed by three structures namely NP + of, Passive + PP and Adverbial clause which share almost similar number of occurrences in different texts. They are qualitatively analysed as supported by the following excerpts and concordance lines extracted from the corpus to substantiate the answers for the second research question. This is done after the frequencies were tabulated.

a) PP + NP

N	Concordance
1	the largest informal settlement in Africa , almost no one has a tap in their (Science)
2	The sight of lightning crackling at the edges of a night sky or flickering (Science)
3	Mr Diamond, a professor of geography at the University of California, (Arts)
4	10-hectare ponds has channels running around the perimeter and across (Science)
5	industrial and laboratory gloves from the non-medical sectors, including (Arts)
6	Over-supply from the 1970s to the mid-80s kept the prices down and this (Arts)
7	producing a more nutritious grain. This form of agriculture is less than (Science)
8	of the attachment system. One of these effects would be Aggression.” (Science)
9	level of intensity could not be stirred by a machine? If those qualities that (Science)
10	careers to those that are consistent with their emerging self-concepts as (Arts)

Bundles starting with *in* and *at* as shown in examples 1, 2 and 3 above signify location and that matches with the function of preposition. Similarly, LBs beginning with *around* and *from* pointed to certain places can also be seen in examples 4 and 5. To describe time, bundle with *to* as the head is used. Two examples namely 7 and 8 have an embedded *of-phrase*. Similar to Hyland and Tse’s (2009) findings, bundles starting with *of-phrase* identified in this study are used to indicate relations between propositional elements. Other bundles starting with *by* and *with* are used for elaboration and clarification.

b) NP + post modifier

N	Concordance
1	be blended into motor oils without the need for costly engine conversions (Science)
2	Seattle, are investigating the possibility that high levels of fat and fructose (Science)
3	remains far from proven. For example, the rise in asthma, because that (Science)
4	technological advances, using animals to work the soil enabled some (Science)
5	Immigrants from countries with the disease are offered screening when (Science)
6	common ground between the reader and the writer. You want the reader to (Arts)
7	it works for others. Even before we or other people put knowledge into (Arts)
8	and more upset, imagining what we will do or say to the perpetrators, (Arts)

Most of the LBs incorporating NP in MUET reading texts take the form of a noun phrase with post modifier fragment as shown in the examples above. Definite article *the*, with 3577 occurrences in all the texts, is regarded as the cause for the abundance of structure of this kind (See examples 1 to 3). This is because the definite article is usually placed in front of a noun. Lin (2008) mentioned that this category of LBs is discipline specific, used to indicate the importance of basic concepts. This is clearly shown in examples 4 and 5 extracted from the science based passages. However, LBs from the arts disciplines rarely allude to content-based words (See examples 6 to 8).

c) *VP/Adj + to clause*
N Concordance

- 1 9 billion by 2050, cannot afford **to treat** the sea as an infinite resource. (Science)
- 2 It was a price that had **to be** paid then, as the economy of the country (Arts)
- 3 proudly to us. "Little Masako here," for the **first time** to my recollection (Arts)
- 4 (Gordon & Follmer 1994). Children are also **more likely to** guess when (Arts)
- 5 pattern continues, man will **need to** colonise another two planets within (Science)
- 6 kill their prey for food, dolphins **seem to** have murderous urges unrelated (Science)
- 7 German psychologist William Stem, children **tend to** be less accurate than (Arts)
- 8 who seems just a little too perfect? We **need to** watch our listening habits (Arts)

A complete V/Adj + to clause is rare however, purely *to* clause is abundant in the corpus. Examples 1 and 2 are instances where *to* clause exist without a verb or adjective phrase. However, as shown in the first two examples, words which come before *to* clause are verbs; they are not taken into account because they have low frequency count. Adjective phrase is also plentiful; especially in arts based-texts which are prone to employ more bundles of this type (See examples 3 and 4). Focusing on the head word of V/Adj + to clause structure, a verb is preferred over an adjective as depicted in examples 5 to 8.

d) *NP + of*
N Concordance

- 1 This is evidenced by the **growing number of** pharmaceutical companies (Science)
- 2 Apart from corals, various **species of** marine invertebrates such as sea (Science)
- 3 chemists at the Coca-Cola company were dealing with **one of** the chemical (Science)
- 4 pool. With an increase in the **number of** Asians qualified for white-collar (Arts)
- 5 The personal data is the first **part of** a resume. Put this information in a (Arts)
- 6 drew ordinary people into imaginative writing. **One of** these was the amateur (Arts)
- 7 Over thinking seems to be a horrible side-**effect of** affluence and, she says, (Arts)
- 8 Indians older than 6, the national mean for **years of** schooling is 3. In (Arts)

All examples of LBs in this sub-category shown above have *of-phrase* as the head of the post modifying component which are typically followed by another noun. The *of-phrases* are used to describe characteristics of the noun. They are usually linked to the issues discussed. For example, 'White Collar Jobs' and 'Writing Resume' are titles of the reading passages where concordance line 4 and 5 are extracted. Moreover, acting as the head of this bundle, NP specifies either quantity or quality of the following phrase.

e) *Passive + PP*
N Concordance

- 1 system with fewer command-and-control cells **known as** regulatory T cells (Science)
- 2 taste **can be** classified into four main categories; sweet, sour, salty, bitter (Science)
- 3 one place at a time. Bits, on the other hand, **can be** copied and presented (Arts)
- 4 variety of ethical criticisms **have been** levelled against advertising. Because (Arts)
- 5 its infancy. Private companies **have been** allowed to film independently only (Arts)

This sub-category of structural type is made up of a passive verb followed by prepositional phrase fragment. Some are used to introduce terms as shown in example 1 while examples 2 and 3 direct readers' attention to the topic of discussion. As for examples 4 and 5, only the passive form – 'have been' meet the cut-off frequency but not the main verb as well as the preposition.

f) *Adverbial clause*
N Concordance

- 1 used in traditional remedies. **For example**, "kacip Fatimah" (Labisia (Science)
- 2 Malaysia's exports to markets **such as** Poland, which grew by 73.3%, Qatar (Arts)

- 3 useful, although not always correct, **as well as** with feelings of comfort (Science)
- 4 style of vacationing. It refers not only to leisurely and environmentally (Arts)
- 5 life's frustrations as adolescents. **In contrast**, those who gave in to their (Arts)
- 6 (JE) are very sensitive to climate changes **because they** are cold-blooded. (Science)

Most LBs in this sub-category are sentence connectors or conjunctions specifically. To exemplify, adverbial clauses like *for example* and *such as* are appropriately used as shown in examples 1 and 2. Bundles in examples 3 and 4 are used to show addition of ideas. The rest of the bundles explicitly mark a logical relationship of comparison/contrast and cause/effect which can be seen in examples 5 and 6 respectively. Unlike past studies conducted by Chen and Baker (2010) and Strunkeyte and Jurkunaite (2008) where bundles of this kind are not significant due to their infrequency, they are however rather apparent in this study.

While the above are the findings of the first and second research questions, results of the third research question which focuses on comparison of structural types of LBs in science and arts-based texts are as follows. It is apparent that the distributions of LBs show slight differences in terms of occurrences. PP-based LBs with approximately 30% of occurrences came in first, followed by dependent clause and NP-based LBs with 26% and 23% respectively. The least is VP-based bundles with less than 20%. Prepositional phrase is the most apparent structure regardless of disciplines while noun phrases and verb phrases are not very prevalent. These results appear to be parallel with previous studies such as Dontcheva-Navratilova (2012), Bal (2010) and Liu (2008) who investigated a variety of registers and reported that the largest part of the LBs was made up of PP. The remaining structures play important roles in forming sentences but their distributions are not as significant as PP + NP. The summary of occurrences of Biber et al (2004) structural components in the two disciplines is shown below:

TABLE 4. Occurrences of structural categories in science and arts-based texts in percentage

		Science (%)	Arts (%)
NP-based	NP + of	8.2	6.6
	NP + post modifier	13.2	10.3
	NP +be	3.0	4.8
VP-based	Passive + PP	9.9	5.9
	be + N/AdjP	7.6	8.5
	it + V/AdjP	3.3	3.7
PP-based	PP + of phrase	0.0	0.0
	PP + NP	28.9	32.7
Dependent Clause	VP + that clause	3.6	4.4
	V/Adj + to clause	13.5	14.3
	Adverbial clause	8.9	8.8
		100.0	100.0

DISCUSSION

In the sub-categories of structural analysis, scientific-based texts tend to employ more NP-based and VP-based LBs, specifically NP + post modifier, NP + of phrase and Passive + PP. There are justifications as to why the three structures mentioned were preferred in scientific texts. Alderson (2005) stated that science articles are generally more argumentative, hence the facts are linked in routinely patterned ways as speculated by Hyland (2008). To highlight the importance of basic concepts in a particular discipline, the gist presented in the form of a

noun phrase with post modifier fragment containing its elaboration is relevant. The NP with of-phrase fragment (*a lot of, growing number of, range of*) widely used to identify quantity is crucial in presenting accurate information. As for the passive bundles followed by a prepositional phrase fragment, it may be due to the emphasis the authors wish to put on the subject matter.

Focusing on arts-based texts, the most frequently used structure is dependent clause despite marginal difference is recorded between arts and science-based texts as shown in Table 4. Two structures namely (VP) + that clause and (V/Adj) + to clause play important roles in making the language more comprehensible, functioning as merging devices. However, most of the time, verb and adjective may not be seen because only the occurrences of the clauses meet the criteria set.

Generally, PP-based LBs are very significant in this study even though PP + of phrase structure is not found. Its occurrence is approximately one third of the total for both science and arts-based texts. This result appeared to be parallel with previous studies such as Dontcheva-Navratilova (2012), Bal (2010) and Liu (2008) who investigated a variety of registers and reported that the largest part of the LBs was made up of PP. The remaining structures namely NP + be and it + VP/AdjP play significant roles in sentence formation but their distributions are not as significant as PP + NP. They belong to the Phrase Structure (PS) Rule where NP and anticipatory-it are in the subject positions while be-verb, VP and AdjP are the predicates. Despite the claim made by many about LBs where they were not grammatically complete units (Biber et al 2004, Cortes 2004, Bal 2010), the three structures reveal that the basic subject verb agreement rule has been adhered to.

Besides that, the frequently occurring structural categories of LBs in science and arts based texts differ moderately. Associating them to teaching and learning could reveal the advantages gained from incorporating corpus into teaching. Incorporating corpora-based materials into designing English language materials is not new. The word cluster lists and concordances could be employed by teachers to encourage students to use frequently occurring LBs in MUET preparation classes. Corpus-informed lists have been proposed by many including Dontcheva-Navrotilova (2012) and Lin (2008) where they justified its suitability with the learning needs of students from different proficiency level. To assist the weaker students, teachers could provide them with examples of LBs frequently used in the respective disciplines of the word lists. For more advanced learners, inductive learning could be adopted; by providing relevant texts and a computer software (WST 5, for instance), these learners would then be able to obtain concordance lines with particular LBs to expand or enhance their understanding of the LBs in original contexts. This method is proven effective because Lin (2008) managed to influence their students to write better by drawing their attentions to the LBs in the corpora. By using these lists, Liu (2012) affirmed that students' awareness of commonly used multi-word units could be raised and the students might be inclined to use them in their writing. When these LBs are correctly used in essay writing, it signifies absolute understanding of their structures and functions.

As suggested by Hyland (2008), "consciousness raising tasks which offer opportunities to retrieve, use and manipulate items can be productive." (p.20) Teachers could develop a range of teaching and learning activities or exercise based on the frequently occurring LBs identified in this study. The following tasks can be used in MUET preparation courses for assistance in recognition, practice and contextualization of LBs: cloze test exercise, error identification and making a personal list of LBs. The suggested tasks could familiarize students with the structure of LBs in reading texts. Most of the LBs resembled the Phrase Structure (PS) Rule except Dependent Clause. This indicates the importance of understanding the structure of LBs which coincides with the PS Rule to ensure specifiers,

heads, and complements occupy the appropriate positions (Celce-Murcia & Larsen-Freeman 1999) in a phrase.

In addition to employing overt teaching of LBs, teachers should not uphold rigidly to the belief that having a good grasp of knowledge in grammar is one of the main criteria in mastering English language. It is a well-known fact that emphasis is given to the teaching of grammatical rules in Malaysian schools. This scenario is proven when many investigations conducted aim to reveal common grammatical errors made by students (for example, Saadiyah & Khor 2009). The claim that “the teaching and learning of English therefore is seen as learning a subject, focusing on the mechanics of the language without making connections to how it is used in real communicative events,” (Normazidah, Koo & Hazita 2012, p.39) reflects Malaysian teachers’ attitude towards grammar. Hence, attention should now be given to LBs because they are principally the foundation of a grammatical sentence. A combination of any two or more items from the four structural categories of LBs namely NP, VP, PP and Dependent Clause could form an intelligible sentence; for example we are (NP) at the university (PP). Thus, equal status must be given to LBs in teaching and learning of English language. They are not less important than subject-verb agreement (SVA), tenses and other grammatical items.

LIMITATIONS

Two limitations arose while carrying out this study. They were mainly caused by the size of data obtained for this study. Despite the number of MUET reading test papers collected throughout the span of ten years, the amount of data collected was rather small. This was because of two main reasons: a) the passages used in MUET reading tests were relatively short and; b) MUET started only in 1999. As a result, the size of the corpus built for this study was smaller compared to past studies which investigated LBs in million word corpora such as Bal (2010) and Hyland (2008).

Firstly, due to the size of the corpus (15,606 running words) instead of focusing on the researchable length of LBs, 4-word bundles (Hyland 2008), this study took into consideration two to six-word bundles. Most of the LBs generated by WST 5 were 2-word bundles, followed by 3-word bundles and very few 4-word bundles; 5 and 6-word bundles were not detected. However, 2-word bundles as mentioned by Hyland (2008) could not offer a clear range of structures compared to 4-word bundles. Hence, the categorisation became difficult due to the presence of 2-word bundles which were rather ambiguous.

Secondly, because of the small amount of data leading to the emphasis put on 2-word bundles, fitting certain LBs into any of the sub-categories proposed using Biber et al. (2004) taxonomy became difficult. For example, *which are*, *when it is*, *and the*, *and it is*, *a new* were among the bundles that could not be fitted into any existing sub-structural category.

CONCLUSION

As the objectives of the study were to determine the types and structural categories of LBs as well as their frequencies in MUET reading tests, the findings were limited to these aspects of frequently occurring LBs in the tests. In short, science-based texts consisted of more NP and VP-based LBs while arts-based texts contained more dependent clauses; PP-based LBs are very significant in both science and arts-based texts. With the completion of this study, an overview of how LBs in texts construct information is obtained.

Despite the small amount of data collected, an overview of LBs commonly used in MUET reading texts has been provided. The results also proved that authors of the reading

passages convey the intended meaning through certain language patterns depending on the context. In other words, the extent to which LBs are specific to particular disciplines was confirmed. This scenario is comparable to studies which looked at textbook authors and other writers in the academic context. To ensure they do not go unnoticed, teachers must provide students with an understanding of the features of passages they may encounter in reading texts. Providing students with an understanding of the features of the discourses they will encounter was regarded by Hyland and Tse (2009) as the best way to prepare them for their studies. The results obtained could also function as a guideline in choosing passages for reading tests of other English examinations in Malaysia. Test designers can use results from this study to adopt suitable passages. Thus, the significance of this study is three-fold where not only teachers and students will benefit but also the test developers.

REFERENCES

- Alderson, J.C. (2005). *Assessing reading*. United Kingdom: Cambridge University Press.
- Alderson, J.C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language*, 6(2), 425- 438. Retrieved from <http://nflrc.hawaii.edu/rfl/PastIssues/rfl62anderson.pdf>
- Bahiyah, A. H., Mohd. S. M. Y., Kesumawati, A. B., Yuen, C.K. & Azhar, J. (2008). Linguistic sexism and gender role stereotyping in Malaysian English language textbooks. *GEMA Online Journal of Language Studies*, 8(2), 45-78.
- Bal, B. (2010). Analysis of four-word lexical bundles in published research articles written by Turkish scholars. *Applied Linguistics and English as a Second Language Theses. Paper 2*. Retrieved from http://digitalarchive.gsu.edu/cgi/viewcontent.cgi?article=1001&context=alesl_theses
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(2007): 263-286.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at.....: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.
- Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The grammar book*. 2nd ed. USA: Heinle & Heinle Publishers
- Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30–49. Retrieved from <http://llt.msu.edu/vol14num2/chenbaker.pdf>
- Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speakers. *Brno Studies in English*, 38(2). Retrieved from http://www.phil.muni.cz/plonedata/wkaa/BSE/BSE_2012/BSE_2012-38_2_-XX_Dontcheva-Navratilova.pdf
- Green, A. Unaldi, A. & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(2), 191–211.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(2008), 4–21
- Hyland, K & Tse, P. (2009). Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies (IJES)*, 9(2), 111-129.
- Kern, R. (2000). *Literacy and language teaching*. Hong Kong: Oxford University Press. Laane, Mare-Anne. (2011). Lexical bundles in engineering research articles. *Proceedings of the 10th International Symposium “Topical Problems in the Field of Electrical and Power Engineering”*, 72-75.
- Lee, K.S. (2004). Exploring connection between the testing of reading and literacy: The case of the MUET. *GEMA Online Journal of Language Studies*, 4(1), 1-12.
- Lin, C. (2008). *An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks* (Unpublished master’s thesis). Canada: Carleton University.
- Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, 31 (2012), 25–35.
- Malaysian Examination Council. (2008). *Malaysian University English Test (MUET)*. Malaysia: Majlis Peperiksaan Malaysia.
- Nambiar, R.M.K. (2007) Enhancing academic literacy among tertiary learners: A Malaysian experience. *3L The Southeast Asian Journal of English Language Studies*, 13, 77-94.
- Nazira, B. O. & Kamaruzaman, J. (2009). Routinizing lexical phrases on spoken discourse. *International Education Studies*, 2(2), 188-191. Retrieved from: <http://ccsenet.org/journal/index.php/ies/article/viewFile/1726/1660>
- Neely, E., & Cortes, C. (2009). A little bit about: Analyzing and teaching lexical bundles in academic lectures. *Language Value*, 1(1), 17-38.

- Noorizah, M. N. (2006). Reading academic text: Awareness and experiences among university ESL learners. *GEMA Online Journal of Language Studies*, 6(2), 65-78.
- Normazidah, C. M, Koo, Y.L., & Hazita, A. (2012). Exploring English language learning and teaching in Malaysia. *GEMA Online Journal of Language Studies*, 12(1), 35-51.
- Pang, W. (2010). Lexical bundles and the construction of an academic voice: A pedagogical perspective. *Asia EFL Journal* 47. Retrieved from [http://www.academia.edu/1428822/Lexical Bundles and the Construction of an Academic Voice A Pedagogical Perspective](http://www.academia.edu/1428822/Lexical_Bundles_and_the_Construction_of_an_Academic_Voice_A_Pedagogical_Perspective)
- Pardo. L. S. (2004). What every teacher needs to know about comprehension. *International Reading Association*, 272–280. Retrieved from: http://www.learner.org/workshops/teachreading35/pdf/teachers_know_comprehension.pdf
- Read, J., & Nation, P. (2002). An investigation of the lexical dimension of the IELTS Speaking Test. *IELTS Research Reports Volume 6*. Retrieved from http://www.ielts.org/pdf/Vol6_Report7.pdf
- Romer, U. 2004. Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Eds.). *Corpora and Language Learners* (pp. 151–68). Amsterdam: John Benjamins.
- Saadiah, D. & Khor, H. C. (2009). Common errors in written English essays of form one Chinese students: A case study. *European Journal of Social Sciences*, 10 (2), 242-253. Retrieved from http://www.eurojournals.com/ejss_10_2_07.pdf
- Strunkyte, G. & Jurkūnaite, E. (2008). *Written academic discourse: Lexical bundles in humanities and natural sciences* (Unpublished bachelor's thesis). Vilnius University, Vilnius.
- Zuraidah, M. D., & Maya, K. D., 2005. The testing of literacy skills in an ESL environment. In Chan Swee Heng & Malachi, Edwin Vethamani (Eds.). *ELT Concerns in Assessment* (pp. 111-126). Malaysia: SASBADI MELTA ELT Series.

