Exploring conventionalized phraseology in advanced academic writing of Tunisian linguistics students: a lexical bundles analysis

Arab Journal of Applied Linguistics
e-ISSN 2490-4198
Vol. 06, No. 02, 2021, 1-30

http://www.arjals.com

Almontassar Bellah Taieb[1], *Faculty of Humanities and Social Sciences, University of Tunis, Tunisia*

Naouel Toumi, The Higher Institute of Languages of Moknine, University of Monastir, Tunisia

Abstract

Multi-word combinations are crucial for the production of various discourse types as well as for the construction of a credible and persuasive academic persona. This study investigates the frequency, structure, and function of the most frequently occurring 4-lexical bundles in writings from the field of linguistics. Following Biber et al.'s (1999) corpus driven-approach, a total of 87 and 113 of lexical bundles were identified in expert and student writing, respectively. Findings revealed substantial differences in the frequency distribution of lexical bundles across the structural categories. More precisely, the study indicated that the student cohort relied heavily on clausal bundles. Expert writers, on the other hand, showed increased use of phrasal bundles, which are considered indicative of proficient academic writing. In terms of function, the analysis showed a preferential attachment to certain lexical bundle functions, with text-oriented bundles occurring significantly in student writing as opposed to research-oriented bundles which were predominantly used in expert writing. These functional differences also suggest diachronic changes in experts' use of these word sequences due to the discipline's discursive demands.
Keywords: Lexical bundles, academic writing, corpus-driven analysis, structural categories, functional categories

[1] Corresponding author.

Tel.: +216 52214650
Email address: almontassar.taieb@gmail.com

1. Introduction

Over the past few decades, EAP research has focussed on the description of various linguistic features indicative of writing proficiency and balanced argumentation. Most of these studies have employed bottom-up corpus-driven approaches to analyse varied word-sets that occur with relatively high frequency in discourse (e.g., Biber, *et al*., 1999; Byrd & Coxhead, 2010; Simpson & Mendis, 2003; Wray, 2000) while others have devised complex psycholinguistic processing metrics to identify perceptually salient sequences made available for teaching purposes (e.g., Ellis, Simpson-Vlach & Maynard, 2008; Robinson & Ellis, 2008; Simpson-Vlach & Ellis, 2010). These studies highlight the linguistic resources that L2 students are lacking or the writing skills that the L2 population must retain to achieve native-like proficiency. They are also part of a growing tradition in studying various types of lexico-grammatical features, many of which are central to human communication (see the review of research in Conklin & Schmitt, 2008; Ellis, 1996; Granger & Meunier, 2008; Wray, 2000; Weinert 1995).

Research in this area has provided important information about domain-specific vocabulary and its phraseology across the full spectrum of scholarly disciplines and genres. More recently, however, there has been more form-focused lexicogrammatical analysis on the phrasal pattern of English often called *formulaic language*. This term is used to encompass knowledge of different multi-word combinations (MWCs), from the semantically opaque idioms (e.g., *bury the hatched*) and proverbs (e.g., *Birds of a feather flock together*) to the relatively transparent collocations (e.g., *strong tea*) and speech formulas (e.g., *what's up*). The finding that MWCs holds a central position in language learning owes much to Sinclair's (1991) 'idiom principle' which undermined the traditional boundaries between lexis and grammar considering that chunks are the building blocks of language and that they have important implications on "the way linguistic knowledge is learnt and subsequently represented in the lexicon of a native speaker and an L2 learner" (Siyanova-Chanturia & Martinez, 2014, p. 3). So if verbal communication makes considerable use of formulaic expression retrieved as wholes,

then it becomes apparent that the unit of language is "the phrase, the whole phrase, and nothing but the phrase" (Sinclair, 2008, p. 407), hence should occupy a central position in L2 learning.

One strand of lexico-grammatical research, however, has focused on the description of recurrent expressions that are comprehended as quickly and accurately as literal speech even if presented out of context. This line of research has grown tremendously following Altenberg's (1998) pioneering work which employed a frequency-based approach to identify recurrent lexical phrases in English (see also Alternberg & Eeg-Olofsson, 1990). The most comprehensive version of this work was conducted by Biber et al. (1999) whereby their grammatical survey focused, though not exclusively, on examining some types of extended collocations known as 'lexical bundles'. Overall, an approach that aims to isolate these lexico-grammatical features typically draws on a combination of frequency and range criteria to distinguish bundles from other types of lexical strings.

For example, lexical bundles are distinct from idioms on a number of levels: idioms (e.g., *walk around in hot porridge*) have non-literal meaning and are rarely attested in discourse; bundles (e.g., *it is possible to*), on the other hand, are pervasive and cannot be substituted by a single word. The same applies to collocations where some words tend to co-occur in pairs. Unlike idioms, collocations (e.g.; *take a risk*, *take action*) derive their meaning from the literal meaning of the words; and for that reason, they are described as "statistical associations rather than relatively fixed expressions" (Biber et al., 1999, p. 988). On the contrary, a lexical bundle is a sequence of words that contain no less than three-word units. In conversation, according to Biber et al (1999), 15% of lexical bundles have a complete structural unit (e.g., *I don't think so*). Other bundle types could be mere open slots with no underlying meaning (e.g., *I don't know what*); they permit the addition of other words to offer grammatical and semantic completion.

Moreover, it should be noted that the literature on lexical bundles points to "a problem of terminology" when describing the occurrence of multi-word combinations (Wray, 2002, p. 9). Scholars have, thus, proposed different labels to explain these strong co-occurrences in discourse. These co-occurrences include: 'clusters' (e.g., Hyland,

2008a), 'recurrent word combinations' (e.g., De Cock, 1998) and 'n-grams' (e.g., Biber & Barbieri, 2007). Still, it is important to point out that the way these researchers have operationalized the strong co-occurrence of MWCs is contingent on their knowledge and understanding of the phenomenon. Therefore, it is not surprising that studies have devised "a huge set of descriptional and definitional terms" as Wray (2000, p. 464) suggests.

Overall, studies on languages such as English (Biber et al., 1999; Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004), Turkish (Öztürk & Köse, 2016), Spanish (Pérez-Llantada, 2014), and Persian (Esfandiari & Barbary, 2017) have shown dependency on these word sets considering that they are crucial for the construction of discourse in all academic genres. These studies have generally aimed to identify L1/ L2 distinction in lexical bundle use (e.g., Ädel & Erman, 2012; Appel & Murray, 2020; Bychkovska & Lee, 2017; Chen & Baker, 2010; Esfandiari & Barbary, 2017; Pan et al., 2016), or registers/genres- and discipline-determined bundles (e.g., Biber et al., 1999; 2004; Cortes, 2004; Durrant, 2015; Hyland, 2008a, 2008b; Öztürk & Köse, 2016; Pan et al., 2016; Wang, 2017; Wei & Lei, 2010) and the relationship between proficiency levels and bundle use in student texts (Chen & Baker, 2014).

Biber (2006), for instance, compares bundle distribution in four university registers and concludes that classroom interactions combine "the functional priorities of involved spoken discourse […] with the priorities of informational written discourse" (p. 147) by using three major functional categories: (a) stance bundles (*I want you to*), (b) discourse organizers (*it has to do with*), and (c) referential bundles (*a little bit of*). Similar findings were corroborated by Csomay (2013) who explored bundle type distribution in classroom interaction while providing insights into their distributional patterns across three macro-phases in the discourse structure. With respect to bundles in academic writing, studies have shown that "lexical bundles […] are building blocks for extended noun phrases or prepositional phrases" (Biber et al., 1999, p. 992). In characterising the development of language across various writers' groups, Pan et al. (2016) conclude that "phrasal-clausal differences exist between L1 expert academic writers and all other groups [including] L1 novice writers, L2 novice writers, and L2 expert academic

writers" (p. 69). Therefore, it seems convenient that all writers (L1 and L2 alike) must demonstrate appropriate use of lexical bundles to help qualify language users as legitimate members within a particular academic milieu (Coxhead & Byrd, 2007) and project shared disciplinary ethos (Hyland, 2005).

With regard to the relationship between disciplinarity and lexical bundle use, it was found that variation in bundles is "a reflection of disciplinarity, rather than a by-product of differences in text types" (Durrant, 2015, p. 14). For example, Cortes (2004) showed considerable differences between published writing and student writing in biology and history. Hyland (2008b) equally observed large variations in lexical bundle usage in natural sciences (e.g., engineering and biology) and social sciences (e.g., business studies and applied linguistics). Consistent with previous research, his frequency analysis exhibited a significant amount of nouns and prepositional phrase fragments such as *on the other hand*, *at the same time*, and *in the case of*.

Although studies have revealed that novice writers (both L1 and L2) would produce less varied lexical bundles than expert writers, more complex and nuanced accounts have been well documented in the literature (e.g., Hyland, 2008b; Wei & Lei, 2011). There is a need for more focused studies which closely examine the production tendencies / differences of L1 and L2 academic writers. More importantly, there seems to be a lack of empirical evidence on the way these word-sets are used in the Tunisian context, with the majority of research work falling into studying a few types of phraseological sequences (*cf*. Almaktary, 2017; Massaabi, 2014; Toumi, 2011) thus inadvertently excluding other important expressions that are central in writing high-stakes academic texts. We believe that understanding the way these sequences are used by Tunisian EFL majors would consolidate previous findings as well as offer useful discussion points for tracking changes in experts' use of lexical bundles.

With this in mind, the present study aims to compare 4-lexical bundle use in expert and graduate students-produced linguistics writing. The comparison focuses on examining the degree to which Tunisian EFL student writing is formulaic using expert norms to isolate bundles characteristic of proficient writing. The choice of bundle length was informed by the fact that 4-word sequences are much more common than other

MWCs and many of them include shorter sequences which do not tend to be very interesting in EAP research. In this way, target bundles are derived quantitatively by means of frequency and relative semantic opacity. The present research is guided by the following research questions.

RQ1. What are the most frequent 4-lexical bundles used by Tunisian students and expert writers in the field of linguistics?

RQ2. Are there quantitative differences between target bundles used by Tunisian students and expert writers?

RQ3. Are there any structural differences between Tunisian students and expert writers in the use of lexical bundles?

RQ4. Are there any functional differences between Tunisian students and expert writers in the use of lexical bundles?

2. Methodology

2.1. Corpus design

The present study is based on an electronic corpus of 1,499,419 words representing a collection of academic texts from the field of linguistics. The corpus comprises two sets of texts sharing a number of *textual* and *contextual* attributes, such as the IMRD (Introduction/ Method/ Results and Discussion) structure, mode (written academic discourse), situational variety (formal), and communicative purposes (displaying and sharing results with the audience). The first set of texts is a Postgraduate Corpus (PC) which includes Master's dissertations written by Tunisian English language and linguistics majors between 2014 and 2017. The dissertations were collected from three universities after obtaining formal permission to avoid copyright infringements. The other set of texts is an Expert Writers Corpus (EWC); it contains research articles (RAs) published between 2014 and 2017 in three peer-reviewed English medium journals (Journal of Applied Linguistics, Journal of English for Specific Purposes, and TESOL Quarterly). This procedure resulted in a total of 180 texts: 30 dissertations containing 595,292 words in the PC and 150 published RAs amounting to a total of

904,127 words in the EWC. However, we did not attempt to balance the size of the two corpora because we considered it to be more useful to cover the maximum number of these text types. Both corpora were annotated to minimize the amount of noise and error that could result from extraneous data. This clean-up procedure eliminated components such as tables, charts, reference lists, direct quotations, acknowledgements, and notes. Each text sample was saved in plain-word format (.txt) and given a reference code such as RA1 or TH1. These codes indicate the genre category of the text file (TH = thesis / RA = research article) whereas the last digit is a numerical system that shows the order of the file among the remainder text samples.

2.2. Identification of lexical bundles

In line with previous lexical bundles research, the study took a radical corpus-driven approach and made "minimal apriori [*sic*] assumptions regarding the linguistic constructs that should be employed for the analysis" (Biber, 2009, p. 276). To address the first research question, the extraction of lexical bundles has followed a three-step procedure. First, a 4-word scope was set as the default length to avoid writers' idiosyncrasies. This decision took into account the fact that "4-word strings are far more common than 5-word strings and therefore offer a clearer range of structures and functions than 3-word bundles" (Hyland, 2008b, p. 8). Second, we selected a frequency threshold of 20 times per million words (pmw) or higher. Frequency cut-off points, however, are considered "somewhat arbitrary" (Biber & Barbieri, 2007, p. 267), ranging between 10 to 40 times pmw, or even at a much wider range depending on the corpus size. This restricted frequency cut-off point 20 pmw, therefore, is conceived of as a part of creating a representative list of MWCs which can be used for pedagogical purposes. Finally, target bundles should occur across five text samples at least. Although research on student writing has often employed a lower distribution criterion (e.g., Ädel & Erman, 2012; Chen & Baker, 2010), in the present study, it has shown to be possible to adopt similar dispersion criteria of five texts for the analysis of two different sized corpora. This ensured that the extracted bundles are "multi-word formulaic sequences"

(Biber, 2009, p. 277). It should also be noted that "there is no 'correct' list of the important lexical bundles in a register because the identification process is highly influenced by the corpus design/composition and by the identification procedures" (Pan et al., 2016, p. 64).

Following these criteria, lexical bundles were automatically extracted by AntConc (3.5.6). The *clusters/N-Grams* key function was used to scan and generate a complete listing of the most frequently occurring 4-word combinations for each group data set. Due to the absence of "neat form-and-meaning mappings" to lexical bundles (Grabowski & Juknevičienė, 2016, p. 58), a manual reading of the concordance lines was performed to filter out context-dependent bundles. Following Chen & Baker (2010) and Ädel & Erman (2012), bundles incorporating proper nouns (e.g., *in the United States*) as well as topic- and discipline-specific bundles (e.g., *English as a foreign*, *humanities and social science, native speakers of English*) were excluded.

Overall, the identification and filtering procedure resulted in a list of 87 different bundles in the PC and 113 different bundles from the EWC. All frequency counts were normalized by one million words due to the proportionally unrelated size of both corpora. In the next stage, we compared specific lists of the frequency output to identify bundles that are characteristic of the novice and expert writing. This would provide a clear indication of salient 4-word sequences in each writer group.

2.3. Classification of lexical bundles

Target bundles were first grouped into structural categories drawing on Biber et al.'s (1999) taxonomy. Three main structural categories were distinguished: NP-based, PP-based, and VP-based bundles. Each main category was broken down into twelve sub-categories to examine recurrent structural patterns in both corpora. For instance, NP-based (e.g., *the result of the*, *the role of the*) and PP-based (e.g., *on the other hand*, *at the level of*) bundles include nominal and prepositional phrases (see *p-frame* in Chen & Baker, 2010) whereas VP-based bundles (e.g., *is based on the*, *to be aware of*) refer to any word combination with a verb component. Although we adopted Chen & Baker's (2010)

classification, we also considered it more appropriate to promote the subcategory of *others* to a major heading following the methodology proposed by Hyland (2008a). This is because "bundles like *as well as the* or *as well as their* do not fit in the VP-based category" (Pan et al., 2016, p. 66).

The final step of the analysis was to identify the discourse functions of 4-lexical bundles using the concordance listings and expanded *co-text*. It is necessary to note, however, that Hyland's (2008a) functional framework was chosen over Biber et al.'s (2004) mainly because the former "collects bundles into three broad foci of research, text, and participants, and introduces sub-categories which specifically reflect the concerns of research writing" (Hyland, 2008a, p. 13). Four minor adjustments were then made to accommodate multifunctional bundles after examining their primary functions in the concordance lines: (a) Biber et al.'s (2004) subcategory of *intangible framing attributes* (e.g., *the nature of the*, *the fact that the*) was added to the category of *research-oriented bundles*; (b) a new subcategory was added to account for bundles expressing *objective* (e.g., *for the purpose of*, *to ensure that the*) as proposed by Lu & Deng (2019); (c) *stance* markers were divided into *impersonal* (e.g., *a better understanding of*, *it is important to*), *epistemic* (e.g., *can be seen as*, *that there is a*) and *modality* (e.g., *it is likely to*, *to be able to*) bundles following Biber et al.'s (2004) taxonomy; (d) Hyland's *topic* bundles were not included since there were no examples found in the corpus.

To obtain a clearer picture of bundle use in expert and novice writing, a parametric analysis was performed to compare differences in the structural and functional distribution of bundles across the two corpora using the log-likelihood (LL) statistic– also known as G2. Higher G2 values indicate significant differences between corpora data sets which, in this case, suggests major differences in the occurrence frequency of bundles. Following previous research, the term *overuse* and *underuse* were used when discussing the distribution of target bundles in both groups (Guilquin et al., 2007).

3. Results and discussion

The identification procedure generated a sufficiently large data set of 113 bundle types in the EWC and 87 bundle types in the PC. Similar to Hyland (2008a), the study aggregates the 30-top ranked lexical bundles to enable the comparison of the most frequent 4-lexical bundles in each group. As shown in Table 1, frequency analysis reveals an interesting contrast between experts and novice writers. A sizeable portion of target bundles in the PC is far more common than other bundle types, occurring at a normalized rate of 200 times pmw, thus, revealing the extent to which a stock of prefabricated chunks is "mentally primed [or perhaps] primed through explicit instruction or/and incidental learning" (Pérez-Llantada, 2014, p. 89). This reinforces the view that the formulaic of RAs is chiefly built around the use of restricted sets of ready-made language to help confer meaning in the most appropriate and conventionalized manner. The limited number of bundles in student writing, however, reflects the psycholinguistic realities of their mental lexicon which contains representations of few word-sets influencing their linguistic production.

In terms of frequency, the analysis identifies several bundles featuring over 100 times pmw. The top-ranked bundle is the sequence *on the other hand*, although it occurred with a relatively higher frequency in the PC than in the EWC (262.06 and 155.95, respectively). More importantly, the PC contains more bundles occurring over 100 times (e.g., *in the use of, in the Tunisian context, as well as the, the total number of, the result of the, in the field of, the finding of the, in the present study, the end of the, is one of the, the use of the*) as opposed to the EWC which shows only two bundle types featuring at such a higher normalized rate (e.g., *at the same time, in the context of the*). Additionally, Table 1 lists a variety of sequences which were less favourable by expert writers (e.g., *to the fact that* and *of the use of*) who chose instead to substitute these word sets with more salient ones (e.g., *in the use of, the fact that the*). However, the fact that a given lexical bundle did not make it onto either list simply means that "the frequency and dispersion criteria were not met in the group's material" (Ädel & Erman, 2012, p. 85). In connection with our first research question, frequency analysis provided empirical evidence for

quantitatively significant use of 4-word combinations in expert writing. To complement our initial findings, we explore the structural properties of lexical bundles since they are more generalizable than a list of specific lexical bundles.

Table 1

The 30 most frequent 4-word combinations across corpora.

Target Bundles in the EWC	No.	Target Bundles in the PC	No.
on the other hand	155.95	*on the other hand*	262.06
at the same time	129.41	*in the use of*	139.43
in the context of	116.13	*in the Tunisian context*	127.67
the extent to which	96.23	*as well as the*	122.63
in terms of the	94.01	*the results of the*	117.59
it is important to	87.38	*the total number of*	117.59
in the form of	85.17	*in the field of*	110.87
as well as the	78.53	*the findings of the*	110.87
in the case of	78.53	*in the present study*	109.19
on the basis of	76.32	*the end of the*	107.51
the use of the	71.89	*is one of the*	105.83
at the end of	70.79	*the use of the*	102.47
in the field of	68.57	*of the present study*	97.43
in the use of	68.57	*at the same time*	90.71
in relation to the	65.26	*in other words the*	87.35
a wide range of	63.04	*as a matter of*	72.23
the nature of the	63.04	*the role of the*	72.23
in the present study	61.94	*when it comes to*	72.23
as a result of	60.83	*in an attempt to*	70.55
the end of the	59.73	*that there is a*	70.55
the results of the	55.30	*at the end of*	68.87
at the beginning of	54.20	*of the use of*	68.87
the ways in which	54.20	*on the basis of*	68.87
per cent of the	54.20	*the fact that the*	68.87
on the one hand	50.88	*be explained by the*	67.19
the fact that the	50.88	*that is to say*	67.19
the total number of	50.88	*the majority of the*	65.51
that the use of	47.56	*to the fact that*	65.51
in the current study	46.45	*the nature of the*	63.83
in other words the	45.35	*as a result of*	60.47

Note. Shared target bundles are shaded.

3.1. Comparison of the structural categories of lexical bundles

For the analysis, target bundles are grouped according to their grammatical correlates using Biber et al.'s (1999) structural classification. We identify three main

categories: (a) NP-based; (b) PP-based; and (c) VP-based. For instance, NP-based and PP-based bundles include nominal and prepositional phrases whereas VP-based bundles refer to any word combination with a verb component (Chen & Baker, 2010). These structural categories are further divided into meaningful subcategories to facilitate the analysis of different bundle types. The Log-likelihood (LL) test is used to identify significant differences between corpora data sets.

3.1.1. Comparison of the distribution of structural categories

Table 2 presents the type and token distribution of lexical bundles. As shown below, both corpora rely on different grammatical types to communicate knowledge in their respective discipline. LL values show that NP-based bundle tokens and PP-based bundle tokens are somewhat comparable across the two data sets. A cursory examination also indicates that the subcategory of NP with other post-modifier fragments and PP with embedded *of*-phrase fragments are the most frequently used types in the EWC. With respect to VP-based bundles, unlike the expert writers, novice writers use significantly more VP-based bundles in their texts, with the exception of adverbial clauses which are relatively consistently used in the EWC.

Table 2

Structural distribution of lexical bundles.

Category	Structural subcategories	Types		Tokens		LL
		EWC	PC	EWC	PC	
NP-based	NP with of *of*-phrase fragment	24	18	829	717	28.33***
	NP with other post-modifier fragments	8	6	322	183	2.56
PP-based	PP with embedded *of*-phrase	31	19	1263	650	26.68***
	PP with *to*-clause fragment	7	6	212	188	8.75**
	PP with other post-modifier fragments	24	18	873	776	36.65***
VP-based	Copula *be* + noun phrase/adjective phrase	3	4	89	136	39.29***
	Anticipatory *it* + verb phrase/adjective phrase	6	5	196	140	0.54

	Adverbial clause fragments	2	-	74	-	74.8***
	Passive verb + prepositional phrase fragment	5	4	131	132	11.82***
	(VP)+ *that*-clause fragment	1	3	28	105	85.42***
	(Verb/adjective) + *to*-clause fragment	1	2	26	52	23.08***
Others		1	1	71	73	7.11**
Total		**113**	**86**	**4114**	**3152**	**40.68**

Note. *= significant at p < .05 level; **= significant at p < .01 level; ***= significant at p < .001

From these data, it seems that there is a clear distinction between two structural patterns of lexical bundles: the first is related to the use of phrasal bundles incorporating nominal and prepositional fragments in RAs, and the second to the use of clausal bundles mainly consisting of finite dependent clauses in dissertations. This shows that compared to the Tunisian postgraduates, the expert writers' use of phrasal bundles reflects the structural norms of academic prose grammatically complex; therefore, the linguistic patterning of VP-based bundles in the student writing is similar to Chen & Baker's (2010) finding which qualified the clausal style of L2 students as "a sign of immature writing". This finding is consistent with previous studies which found that bundles in academic prose are phrasal rather than clausal (Biber et al., 1999, Cortes, 2004; Pan et al., 2016).

To better understand the structural distribution that has arisen from the present analysis, a systematic comparison of bundles percentage across the main categories is shown in Table 3 below. As can be seen, phrasal bundles are the most common grammatical types in both corpora, accounting for almost two-thirds of bundles distribution (both types and tokens). In the EWC, for instance, nearly 85% of bundle types are phrasal structures, making a total of 87% of bundle tokens. In contrast, the student writing style is marked by a comparatively heavy-use of VP-based bundles (both types and tokens).

Table 3

Distribution of bundles across the main category.

Categories	Types %		Tokens %	
	EWC	PC	EWC	PC
NP-based	29.2	27.6	30.4	28.2
PP-based	55.7	50.6	56.3	51.8
VP-based	15	21.8	13.1	20
Total	100	100	100	100

With respect to corpus data, both writers use strikingly different grammatical types in writing their academic texts. Consider the text examples from the corpus material in Examples (1) and (2).

(1) From a linguistic perspective, a text is made up of dozens of co-occurring lexico-grammatical features. Although registers can be compared for *the extent to which* they use individual linguistic features, more robust descriptions are possible by considering *the ways in which* features co-occur. (EWC- RA7)

(2) The first finding is *that teachers are not* in the position to cope with the students' special needs because *they frequently lack* the background knowledge about the nature of the deficit. (PC-TH8)

The major difference between Example (1) and (2) is primarily a matter of language choice as each group seems to value a discipline-specific repertoire of MWCs. Both examples show preferences for different grammatical configurations when using chunks to link different text segments. These preferences are particularly pronounced in the phrasal-clausal bundle distinction. Importantly, the higher frequencies of phrasal complexity features in written academic discourse indicate a marked precision at the lexico-syntactic channels and this would result in texts being relatively short and highly patterned. Accordingly, it becomes clear that the assumed proficiency in the EWC is

primarily achieved by a greater reliance on phrasal complexity features rather than mere finite dependent clauses. It is also important to reiterate that many phrasal complex features are still far from being naturally acquired since they are produced in "the more specialized circumstances of formal writing" (Biber, Gray & Poonpon, 2013, p. 29). Because bundles might occur in extremely productive p-frames, the study proceeds to discuss the most frequently used lexical frames in each group following Chen & Baker's (2010) research practice.

3.1.2. Comparison of frequent frames in the main categories

Similar to previous studies, we focus on studying bundles belonging to the most productive frames in each structural category using frequency information of the type/token distribution. As discussed earlier, the EWC and the PC slightly share a large number of NP-based and PP-based bundles (both types and tokens) and are found to be collocating with two major p-frames: (a) *the + Noun + of the/a* (e.g. *the result of the*); and (b) *in the + Noun + of* (e.g. *in the use of*). The structural distribution of bundle types directly supports Biber et al's (2004) finding which qualifies the same "two fixed frames as extremely productive" (p. 78). The study also shows a third productive frame pattern associated with *anticipatory it + verb phrase/adjective phrase* in the VP-based category (e.g. *it is clear that*). Table 4 groups NP combinations and presents the results of the type and token distribution in each corpus.

Table 4

The frame for "the + noun + of the/a".

	The + Noun + of the/a	Total	
		Types	**Tokens**
EWC	**use** (65), wide range (57), **nature** (57), **end** (54), **results** (50), per cent (49), **total number** (46), use (43), **beginning** (38), **majority** (38), **content** (32), **analysis** (28), meaning (28), quality (27), large number (24), level (24), **findings** (23), better understanding (22), results (22), overview (21), findings (21), wider range (20), **importance** (20), start (20)	25	829
PC	**results** (70), **total number** (70), findings (66), **end** (64), **use** (61), role (43), matter (41) **majority** (39), **nature** (38), **analysis** (36), **findings** (30), **content** (27), **importance** (27), purpose (25), validity (25), effect (24), rest (24), majority (21), **beginning** (27)	19	758

Note. Shared NP-combinations are in bold. Bundle tokens are presented in parentheses.

As can be seen in Table 4, both corpora share a total of 22% of NP-frames. The results, however, do not presuppose that these nominal expressions were used equally frequently for similar purposes. Previous research pointed out that this particular p-frame is often underused by L2 novice writers (Pan et al., 2016). Overall, the current results indicate that expert writers strategically use a wide range of nouns that collocate with this p-frame whereas Tunisian postgraduates draw almost on a completely different stock of such filler types. These noun phrase fragments, however, are primarily used to highlight research-related aspects, such as physical descriptions (e.g., *the content of the*), size (e.g., *the size of the*) and qualities (e.g., *the quality of the*).

Table 5 presents The type/token distribution of the second p-frame. Frequency data shows that PP-based bundles are extremely productive in the subcategory *of-fragments* with the RAs scoring the highest proportion of these filler types using almost twice as many types and tokens as in the dissertations. Examples of these include bundles incorporating NP fragments like *number*, *beginning*, and *use*. These expressions serve to elaborate logical (particularly temporal), [and] textual connections between elements of an argument" (Hyland, 2008b, p. 52), in addition to describing abstract representations (Biber et al., 1999). Thus, it comes as no surprise that the majority of these NP fragments seem to reflect aspects of research processes and data mining (e.g.

field, *use*, and *number*). However, they remain relatively infrequent in the postgraduate students' writing.

Table 5

The frame for "in the + noun + of".

	in the + Noun + of	**Total**	
		Types	**Tokens**
EWC	context (105), terms (85), **form** (77), **case** (71), end (64) **field** (62), **use** (62), beginning (49), **process** (39), variety (21), number (20), terms (27)	12	682
PC	**use** (83), **field** (66), **process** (35), **form** (27), light (24), light (21), **case** (21)	7	277

Note. Shared PP-combinations are in bold. Tokens for each bundle are in parentheses.

The last productive frame pertains to the structural subcategory of anticipatory-*it* patterns. These clausal bundles are signalled, to a great extent, through the frequent use of *to*-clause constructions across both corpora. The most striking result is the near-total correspondence between the type (6 and 5) and the token distribution (196 and 140). Table 6 groups *it*-clauses per corpus.

Table 6

The frame for "anticipatory it + verb phrase/adjective phrase".

	Anticipatory it + verb phrase/adjective phrase	**Total**	
		Types	**Tokens**
EWC	**It is important to** (79), **it is necessary to** (24), it is likely that (21) it is possible that (21), it is possible to (21), **it should be noted** (30)	6	194
PC	When it comes to (43), **it should be noted** (24), **it is important to** (30), **it is necessary to** (22), it is clear that (21)	5	140

Note Shared PP-combinations are in bold. Tokens for each bundle are in parentheses.

As shown in Table 6, these filler types are primarily controlled either by: (a) predicative adjectives usually functioning as complements in *to*-clauses fragments (e.g. *important to*, *necessary to*, and *possible to*); or (b) by passivized verbs (e.g. *be noted*)

followed by *that*-clause constructions (Biber et al., 1999). More specifically, predicative adjectives seem to occupy a sizeable proportion across most of the extraposed clauses whereas passive verbs are nearly absent. Previous research often considers predicative adjectives as useful devices for reporting writers' detachment from the subject matter (Ädel & Erman, 2012). On other hand, they can play a major role in conveying writers' evaluative stance in academic prose (Biber, 2006). With respect to extraposed clauses with embedded passive verbs, they seem to be formed for two main reasons: (a) to add anonymity to the source (Hyland, 2008b); or (b) to disguise the authorial interpretations (Hyland. 2008a).

Overall, the above comparisons unequivocally support the significant differences between novice and expert writers' linguistic production and hence provides satisfactory confirmation to the broader picture of the interlanguage use of ready-made chunks in novice and expert writing. It is also important to note that bundle usage in the PC does not necessarily mean that the Tunisian students are completely incognizant of formulaic language; rather, readers should be mindful of the fact that this group did not draw on a variety of structures as equally frequently as the expert writers.

3.2. The Discourse functions of lexical bundles

As a final step in the analysis, target bundles are classified functionally using Hyland's (2008a) taxonomy. Table 7 shows significant differences in the proportions of the functional distribution of bundle tokens. Text-oriented bundles are significantly overused by the Tunisian students while research-oriented bundles rank second as the most frequently occurring features in the expert writing. For both groups, participant-oriented bundles are the least frequently used, a finding which is consistent with previous research (Chen & Baker, 2010; Hyland, 2000a; Lu & Deng, 2019).

Table 7

Functional distribution of lexical bundles across corpora.

	Category	Types		Tokens		LL
		EWC	PC	EWC	PC	
Research-oriented	Location	9	1	382	76	115.25***
	Procedure	13	8	447	343	4.51*
	Quantification	14	7	489	274	0.85
	Description	2	2	109	54	24.61***
	Intangible framing attributes	7	6	242	188	2.87
	Subtotal			1669	935	15.86***
Text-oriented	Transitional signals	9	9	525	511	39.25***
	Resultative signals	9	10	296	376	72.11***
	Structuring signals	7	5	191	157	4.20*
	Framing signals	22	19	853	646	7.15**
	Objective	4	5	101	146	37.76***
	Subtotal			1966	1836	144.94***
Participant-oriented	Stance impersonal	4	2	153	77	3.80
	Stance epistemic	5	8	116	228	98.86***
	Stance modality	4	2	99	52	1.78
	Engagement features	4	2	111	54	3.43
	Subtotal			479	381	7.52**
Total		**113**	**86**	**4114**	**3152**	**40.68*** **

Note. * = significant at p < .05 level; ** = significant at p < .01 level; *** = significant at p < .001.

Table 8 presents a comparison of the type-token distribution of the functional categories in both corpora. As can be seen, both groups display similar proportions of type-token distribution in the three main functional categories. Text-oriented bundles (both types and tokens) rank as the largest category in both corpora whereas participant-oriented bundles constituted the smallest proportion and thus the type-token percentage of such features produced negligible effects.

Table 8

Proportional distribution of bundle functions (types and tokens).

Categories	Types %		Tokens %	
	EWC	PC	EWC	PC
Research	**39,8**	27,9	**40,5**	29,6
Text	45,1	**55,8**	47,8	**58,2**
Participant	15,0	16,3	11,6	12,2
Total	100	100	100	100

Note. Higher percentages are in bold.

3.2.1. Research-oriented bundles

Among the most important demarcation points between the two groups is the greater concentration of research-oriented bundles in the expert writing. Bundles in this category are particularly helpful in describing the research context, its objectives and recounting methodological procedures. As shown in Table 7, the expert writers use significantly more *location*, *procedure*, and *description* bundles than the Tunisian students. Tokens of *quantification* and *intangible framing attributes*, however, appear to be somehow equally distributed, although they are signalled in the EWC through a variety of nouns collocating with the p-frame *in the + Noun + of* (e.g., *range*, *variety*, *degree*, *number, exception*, *part*) than in the PC (e.g., *majority*, *rest*, *one*). This finding corroborates previous lexical bundles research indicating that L2 writers are likely to produce fewer bundle types collocating with this p-frame (Pan et al., 2016).

More importantly, the study indicates some changes in lexical bundle use which may be attributed to the shifting landscape of knowledge production in soft-science linguistics. It has been shown that research-oriented bundles are extremely common in science and technology writing and that writing in social sciences disciplines requires expert writers to place more focus on interpreting findings and organizing discourse rather than recounting research-related aspects. In this way, it is possible to argue that lexical bundles are not entirely discipline-specific and are expected to show dynamic changes in response to the demands of the global knowledge economy.

3.2.2. Text oriented bundles

Text-oriented bundles are significantly overused across all the functional subcategories in the PC data set than in the EWC. A key role in supporting the use of such sequences is related to their capacity to organize the discourse using a variety of textual relations. The first of these functions is *resultative* bundles and *transition* bundles

(see Table 7). A close examination of the data reveals that these two subcategories are primarily signalled through NP and PP fragments, with the exception of bundles in the form of *copula be +noun /adjective phrase* (e.g., *is due to the*), *that*-clause (e.g., *that is to say*) and passive constructions (e.g., *was found to be*). In addition, the study indicates an area of learners' infelicity in the use of the *transitional* bundle *as well as the*, a finding which aligns with previous lexical bundle research showing that L2 writers often mistake such conjunction to be a synonym to *and*, as shown in Example (3).

(3) The third chapter describes the research setting, subjects, instruments, *as well as the* procedure of the investigation. (PC-TH16)

Objective bundles are used significantly more frequently by the student writers than by their expert counterparts. More precisely, the Tunisian students rely more heavily on the tri-gram *in order to* which is internally merged with a verb component as in the bundle *in order to be*, *in order to achieve*, and *in order to make*. This finding remains consistently congruent with Chen & Baker (2010) and Lu & Deng's analysis which revealed the extensive use of the tri-gram *in order to* in L2 student writing.

With respect to *framing* bundles, they are typically formed using the syntax of phrasal modification. Although the Tunisian students "rehash" a small number of such bundle types, the expert writers employ more varied sequences which in this case suggests greater reliance on such conventionalized lexical phrases in writing high-stakes genres. The vast majority of these sequences are made up of PP-based bundles with embedded *of*-phrase, all of which help readers focus on aspects of the research (*in the present study*), mark key relationships (*in relation to the*) and elaborate different viewpoints (*in terms of the*). Other bundles types can be also found, mainly containing NP fragments and occurring with higher frequency across both corpora (*the extent to which*, *the end of the*, *the beginning of the way/ways in which*).

Finally, tokens of *structuring* are the least frequently used functions of the five subcategories whereby they function as a useful tool for weaving data, as in tables and figures, to construe a clear position towards the material in addition to directing readers'

attention to certain findings (Hyland, 2008b). Nevertheless, it seems that their functional distribution contradicts previous studies which showed the paucity of such sequences in student academic writing (Hyland, 2008b; Lu & Deng, 2019; Pan et al., 2016).

3.2.3. Participant-oriented bundles

Participant-oriented bundles are more frequently used by the Tunisian students than the expert writers (Bychkovska & Lee, 2017; Lu & Deng, 2019; Pan et al., 2016). In particular, the analysis shows that the Tunisian students significantly overuse *epistemic* bundles in the form of existential structures (*that there is a*), anticipatory-*it* patterns (*it is necessary to*) and *can*-modal verbs (*can be explained by*). Here, the students seem to allocate considerable attention to epistemic judgement through which they seek to produce a discourse reflecting their understanding of disciplinary matters. The greater occurrence of these expressions, however, may detract from impartiality in academic prose especially given the relatively high frequency of existential structures which are likely to result in a superfluous writing style. This finding is consistent with Chen & Baker (2014) who concluded that bundles occurring with copula *be* can render writing "both simplistic and verbose" (p. 870).

No significant differences are found in the remaining three subcategories (i.e., *impersonal* bundles, *modality* bundles, and *engagement* features). Nevertheless, it is important to emphasize that the only subcategory closer to the level of statistical significance is *engagement* bundles (see Table 8). Readers should be also mindful of the fact that the Tunisian students are still incognizant of many bundle types characteristic of their respective discipline (e.g., *to be able to*, *it is important to*, *the importance of the*).

4. Conclusion

The present study has offered supporting evidence of lexical bundle use by Tunisian linguistics students. The reported results are beneficial for the production of more proficient academic texts by these novice writers. With regard to how corpus data were useful, frequency analysis showed significant differences in the proportional

distribution of bundles. Compared to the Tunisian students, expert writers used a variety of bundle types indicative of their writing proficiency, thus reflecting not only the psycholinguistic realities of their mental lexicon but also highlighting which MWCs had salient use in their respective discipline. Findings from the structural distribution also showed a near-total absence of association between the two groups. The co-occurrence of NP-based and PP-based bundles indicated that the expert writers conformed well to the normative features of English written academic discourse, whereas the student writing style was edging closer towards the characteristics of spoken discourse. The study, however, cautioned that the students were still showing signs of maturity in their rhetorical practices given the presence of a fair amount of PP fragments. The study examined group-specific use of lexical bundles. It found that research-oriented bundles were significantly overused in the EWC whereas text-oriented bundles had a higher proportion in the PC. In addition, the study emphasized that lexical bundle use is changing due to the way knowledge is communicated in soft-science disciplines. This can be explained by the increasing demands for publication and the wide range of linguistic resources available for constructing meaning in academic texts. With respect to participant-oriented bundles, they were significantly overused by the Tunisian students, especially across the subcategory of *epistemic* bundles. Still, it was noted that students' propensity to use such bundle types is likely to give rise to a verbose writing style.

On the basis of these results, we recommend a genre-based pedagogical instruction to help Tunisian linguistics students become aware of target bundles in their discipline. It is generally assumed that a vocabulary-focused approach to academic formulas has shown to have a positive impact at the level of "recognition, cued output, and spontaneous use" (Peters & Pauwels, 2015, p. 37). An initial suggestion would be to assist students to observe the concordance listings of embedded NPs used by expert writers to fill slot structures like *the + N + of the* or *in the + N + of*. This might take the form of: (a) a presentation stage in which students' attention is drawn to a particular contextualized lexical item, (b) a practice stage by requiring students to complete several activities such as fill in the gap exercises and guided rewriting tasks; (c) and eventually a

production stage to assess students' writing performance in using target items. Corpus-informed discussion about the discoursal functions of *transitional* bundles would be also needed to point out areas of infelicitous use.

References

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes 31*(5), 81–92.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Eds.), *Phraseology*: *Theory*, *Analysis and Application* (pp. 101–122). Oxford University Press.

Altenberg, B., & Eeg-Olofsson, M. (1990). On the phraseology of spoken English. In J. Aarts and W. Mejis (Eds), *Theory and Practice in Corpus Linguistics* (pp. 1-26). Rodopi.

Anthony, L. (2018). *AntConc* (Version 3.5.6) [Computer Software]. Waseda University. http://www.laurenceanthony.net/software

Appel, R., & Murray, L. (2020). L1 differences in L2 English academic writing: A lexical bundles analysis. *Journal of English for Academic Purposes*, *46*, 1-15.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*, 97-116.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*, 275–311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

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

Biber, D., Staples, S., & Gray, B. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*, 5, 639–668.

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

Bychkovska, T., & Lee, J.-J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. University of Sydney Papers in TESOL.

Chen, Y.H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Chen, Y.H., & Baker, P. (2014). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849-880.

Conklin, k., & Schmit, 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. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*, 129–147.

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

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). Rodopi.

Durrant, P. (2015). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 35*, 328–56.

Ellis, N. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition, 18*(1), 91-126.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–396.

Esfandiari, R., & Barbary, F. (2017). A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes, 29*, 21-42.

Guilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English of Academic Purposes*, *6*(4), 319-335.

Grabowski, Ł., & Juknevičienė, R. (2016). Towards a Refined Inventory of Lexical Bundles: An experiment in the formulex method. *Studies About Languages*, 29, 58-53.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173–192.

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

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

Jalali S., & Moini, R. (2014). Structure of lexical bundles in introduction section of medical research articles. *Social and Behavioral Sciences*, *98*, 719-726.

Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, *39*, 21-36.

Meunier, F., & Granger, S. (2008). *Phraseology in Foreign Language Learning and Teaching*. John Benjamins.

Öztürk, Y., & Köse, G.D. (2016). Turkish and native English academic writers' use of lexical bundles. *Journal of Language and Linguistic Studies, 12*(1), 149-165.

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60-71.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, *14*, 84-94.

Peters, E., & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes, 20*, 28–39.

Robinson, P. & N. C. Ellis (2008). *A Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge.

Simpson, R., & Mendis, D. (2003). A Corpus-based study of idioms in academic speech. *TESOL Quarterly*, *37*(3), 419-441.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics, 31*, 487–512.

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

Sinclair, J. (2008). The phrase, the whole phrase and nothing but the phrase. In S. Granger, & F. Meunier (Eds.), *Phraseology*: *An Interdisciplinary Perspective* (pp. 407–410). John Benjamins.

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549–569.

Wang, Y. (2017). Lexical bundles in spoken academic ELF: Genre and disciplinary variation. *International Journal of Corpus Linguistics*, *22*(2), 187-211.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics, 16*(2), 180-205.

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal*, *42*(2), 155-166.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463–89.

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