Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
42(3), Summer 2023, pp. 1-27
https://doi.org/10.22099/tesl.2023.46684.3168
https://dorl.net/dor/20.1001.1.20088191.2023.42.3.1.3

Online ISSN: 2717-1604
Print ISSN: 2008-8191
Research Paper

A Corpus-based Study of the Use of Lexical Bundles in EAP Texts by Iranian EFL and ESL Learners

Masoud Azadnia *

Abstract

Research on native vs. non-native formulaic language use in academic texts, despite its wealth in scope and frequency, lacks an inclusive conceptualization of a non-native language learning context. Impressed by such a flawed approach, the bulk (if not all) of studies in the field compared the use of different multi-word strings in the academic discourse of either foreign or second language learners with a native baseline. The current study sought to address the gap, focusing on the structural and functional use of lexical bundles in two culturally parallel corpora developed in two non-native learning context modes: English as a foreign (EFL) and second (ESL) language. To this end, research reports written by Iranian Applied Linguistics MA and Ph.D. learners studying in different universities in Iran and English-speaking provinces of Canada were compared by a structurally similar native corpus, running cross-tabulation, Chi-square, and residual analysis analyses. The results revealed a significant association between language learning context and lexical bundle use on a functional level. The contextual variations yielded significantly different patterns of use concerning several micro-functions underlying text-oriented and research-oriented functions. Compared to functional differences, the between-corpus structural differences were inconspicuous, specifically concerning micro-structures constituting noun, prepositional, and verb phrase-related bundles. The study embraced the notion that EFL writers need to have immense exposure and enhanced language input available in ESL and native learning contexts to foster a native-like formulaic language.

Keywords*:* Academic writing, formulaic language, language learning context, lexical bundles, lexical development

* Received: 30/01/2023 Accepted: 19/06/2023

* Department of Foreign Languages, Islamic Azad University, Naein Branch

How to cite this article:
Azadnia, M. (2023). A Corpus-based Study of the Use of Lexical Bundles in EAP Texts by Iranian EFL and ESL Learners. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills), 42*(3), 1-26. doi: 10.22099/tesl.2023.46684.3168

As a gateway to the international scholarly discourse community, English academic writing has remained an engaging research area under the spotlight over the years. To keep pace with the demand for developing well-organized English for Academic Purposes (EAP) texts, those struggling with field-specific writing need to strike a balance between structural requirements/standards of academic discourse and lexical/grammatical features characterizing the field (Bhatia, 2014; Wood, 2015). Formulaic sequences (FSs), as manifestations of authentic language use (Staples et al., 2013), represent an extensive category of lexical strings (e.g., lexical phrases, fixed expressions, ready-made units, and lexical prefabs) essential in developing formal-register EAP texts qualified as commendable (Biber et al., 2004; Ellis & Simpson-Vlach, 2009; Hyland, 2008a). Among the various subsets of sequences constituting formulaic language (e.g., idioms, collocations, multi-word expressions), lexical bundles (LBs) are the core category used widely and recurrently in logical conversation and academic prose (Biber & Barbieri, 2007). Defined as "sequences of words that commonly go together in natural discourse" (Biber & Conrad, 1999, p. 184), LBs are mainly non-idiomatic in nature and incomplete in structure (e.g., *as a result of*, *with the help of*) (Biber, 2009). Proficient use of LBs in academic texts contributes to meaning construction, fluent linguistic production, text comprehension with the least processing time and effort, and discourse coherence, naturalness, distinctiveness, and predictability (Cortes, 2004; Ellis et al., 2008; Hyland, 2012; Hyland & Jiang, 2018; Kashiha & Chan, 2014).

The contributory role of LBs in promoting writing quality aroused researchers' attention to factors influencing second (L2) or foreign (FL) language writers' access to a native-like pool of these multi-word cohesive devices. Research has shown that the density and diversity of LBs that native and non-native writers use are heavily contingent on various typological features, such as register (Biber, 2006; Huang, 2018), genre (Biber et al., 2004; Gao, 2017), and discipline (Kashiha & Chan, 2014; Liu & Chen, 2020). Not denying the influence of the features enumerated above, Northbrook and Conklin (2019) argued that among the factors enhancing the formulaic advantage of LBs, the frequency and modality of exposure are of utmost importance. There is also a plentitude of evidential data on probabilistic learning (e.g., Glicksohn & Cohen, 2013; Mitchell et al., 2014; Northbrook et al., 2021) that validate the essential role of exposure frequency and multimodal language learning stimuli in fostering various language learning skills, such as writing. The undeniable link between the two factors and the language learning context where academic writers are involved explains why contextual features are substantially significant predictors of lexical and syntactic variations in academic texts (Vercellotti, 2015; Zhang & Kang, 2022; Zhou & Lü, 2022).

Inspired by the theoretical and empirical underpinnings above, a large chain of corpus-based comparative research was launched to explore the lexical variations between non-native and native writings on a morphological (e.g., Azadnia, 2021; Hakansson & Norrby, 2010; Zhang & Kang, 2022) or phraseological (e.g., Amirian et al., 2013; Yakut et al., 2021) level. Nonetheless, research on how contextual variations affect EFL vs. ESL non-native writers' use of LBs has remained a missing link in the chain. The

current study sought to address the gap by comparing the structural and functional patterns of LBs in EAP texts written by EFL and ESL writers with those in a structurally similar native corpus as a benchmark for natural formulaic language use. Compared to research articles, which are mainly fruits of collaborative scholarly attempts, research reports are individual works reflecting their writers' lexical flavor. The study, therefore, focused on MA theses and PhD dissertations on Applied Linguistics developed by EFL and ESL writers of the same cultural background (Iranian). The following questions were the main areas of inquiry in this research study.

1. To what extent does language learning context affect the structural use of LBs in Applied Linguistics MA theses and Ph.D. dissertations?
2. To what extent does language learning context affect the functional use of LBs in Applied Linguistics MA theses and PhD dissertations?

Literature Review

Theoretical Profile of LBs

As attested by the literature on phraseological text analysis (e.g., Chen & Baker, 2010; Kashiha & Chan, 2014), Biber et al.'s (1999) computational attempt to discover word combinations used recurrently in the *Longman Spoken and Written English* (LSWE) corpus provided a springboard for today's in-depth exploration of a specific multi-word variation, called LBs. Labeled variously as N-grams (Allen, 2010), clusters (Schmitt et al., 2004), and non-idiomatic expressions (Hyland, 2008a), LBs refer to "continuous word sequences retrieved by taking a corpus-driven approach with a specified frequency and distribution criteria" (Chen & Baker, 2010, p. 30). According to Kashiha and Chan (2014), LBs are pattern-free sequences of three or more word forms that co-occur more frequently than expected by chance. Given the conceptualization above, frequency and range (distribution) of occurrence are two hallmarks of these continuous-structure lexical sequences. Acknowledging celebrated works in the field (e.g., Biber & Barbieri, 2007; Cortes, 2004, 2006), multi-word sequences used between 10 and 40 times per million words within three to five texts in a natural collection of oral/written language have the potential to qualify as LBs. Aside from frequency and range, LBs could easily be distinguished from other multi-word chains (i.e., collocational and idiomatic word combinations) based on their continuous fixed forms and non-idiomatic nature (Biber et al., 2004; Cortes, 2002). Moreover, LBs are commonly incomplete in structure and transparent in meaning (e.g., *in the use of*) (Conrad & Biber, 2005).

Quite similar to other lexical units, the structure and function of LBs used in a corpus of natural discourse are contingent on a variety of factors, such as discipline (Hyland, 2008a), genre (Biber, 2006), and writers' first language (L1) (Shin, 2019). These structural and functional variations have been the area of focus in scientific works performed following Biber et al.'s (1999) archetype of LBs analysis (e.g., Biber, 2006; Biber et al., 2004; Biber & Barbieri, 2007; Chen & Baker, 2010; Hyland, 2008a, 2008b; Simpson-Vlach & Ellis, 2010). The primary model proposed by Biber et al. (1999)

classified LBs into twelve structural categories. The categories included prepositional phrases plus of, other prepositional phrases, noun phrases plus of, other noun phrases, passive verbs plus prepositional phrase fragment, anticipatory *it* plus verbs/adjectives, *be* plus noun/adjectival phrases, verb phrases plus *that*-clause fragment, verb/adjective plus *to*-clause fragment, adverbial phrases, pronoun/noun phrases plus *be*, and other expressions (Biber et al., 1999). Some researchers (e.g., Biber et al., 2004; Chen & Baker, 2016) later grouped these structures under three main headings: prepositional phrase-based, noun phrase-based, and verb phrase-based LBs. As for functional variations, one of the widely used models of LBs is the one proposed by Biber et al. (2004), which distinguishes between stance, discourse-organizing, and referential bundles. Stance LBs are concerned with writers' feelings and attitudes, discourse-organizing LBs maintain within-discourse links, and referential LBs facilitate direct reference to various entities in discourse and its textual context.

LBs in Academic Discourse

As lexical resources essential in organizing discourse and reflecting writers' stances and experiences, LBs are presumed to be the cornerstone of the language used by an established academic community (Hyland, 2012). These formulaic sequences also offer the formulaic advantage of retrieving integrated chunks rather than individual words (Biber & Barbieri, 2007; Wray & Perkins, 2000), which could facilitate discourse processing and predictability among hearers/readers. Furthermore, the proper use of LBs peculiar to a particular discipline helps writers reveal their field-specific competence and join the target academic community (Pang, 2010). The vital importance of LBs in EAP texts is traceable in the contention made by Hyland and Jiang (20018) that "lexical bundles are pervasive in academic language use and a key component of fluency, marking out novice and expert use in both spoken and written contexts" (p. 385). The contributory role of LBs in developing commendable EAP texts has been a trigger point to explore LBs occurring more frequently in academic registers (e.g., Hyland & Jiang, 2018; Simpson-Vlach & Ellis, 2010) and practical ways to incorporate these formulaic sequences into writing pedagogy (e.g., Kazemi et al., 2014, Rastchi & Ali Mohamadi, 2017).

Research on LBs use in academic discourse shows that, even within a discipline-bounded discourse area, mode (Biber, 2006), genre (Biber & Barbieri, 2007), time (Hyland & Jiang, 2018), and proficiency (Chen & Baker, 2010) variations may account for a significant difference in writers' preferences for LBs. Researchers who explored LBs variations in academic discourse mainly established categorical frameworks subsuming all frequently-used structural and functional patterns of LBs. For instance, Hyland (2008a) proposed a tripartite model, relying upon Biber et al.'s (2004) functional taxonomy. Based on this model, the whole range of LBs in a representative corpus of academic texts fulfill three discourse functions: (a) situating and contextualizing the research, (b) organizing the research discourse, and (c) addressing research readers/writers. Admitting Hyland's (2008a) functional classification, Hyland and Jiang

(2018) proposed a revised classification of LBs structures, believing that the earlier structural model (i.e., Biber et al., 2004; Chen & Baker, 2016) lacked adequate foresight to distinguish between phrasal and clausal structures, which is a real need for academic LBs analysis. A detailed picture of Hyland and Jiang's (2018) framework, which constituted the analytical basis of the current study, is presented in the Method section.

Language Learning Context and Lexical Variations

The EFL/ESL dichotomy, a well-established classification of L2 language learning contexts, refers to two distinct instructional environments surrounding non-native English learners. English is the dominant language widely spoken in an ESL context, but an additional language confined to instructional settings in an EFL one (Brown, 2001). A long, hard look at the peculiarities of the two contexts provides a hypothetical picture of the potential impact of learning context on language acquisition in general and lexical development in particular. As the first distinguishing feature, compared to an EFL context, where English learners receive adaptive, carefully designed learning input, an ESL context immerses learners in floating, authentic language learning stimuli (Zhang & Kang, 2022). In other words, ESL learners' exposure to the target language (English) is immense in both quality and quantity. Thanks to living in an English-speaking community, ESL learners are also exposed to authentic language use in their daily lives. In contrast, learning chances in school settings are scarce cases of input exposure for EFL learners. As another defining feature, the learning stimuli in the ESL context are much more diverse in modality (auditory, visual, textual), register (formal and informal), and interactive mode (Vold, 2022).

The virtues of enhanced input available in ESL learning contexts reinforce the idea that written or spoken output by ESL writers may enjoy higher degrees of structural and lexical richness. Such deductive reasoning is based on a couple of L2 learning hypotheses, including the Critical Mass Hypothesis (Marchman & Bates, 1994), the Comprehensible Input Hypothesis (Krashen, 1985), and the Language Exposure Hypothesis (Ortega, 2014). As postulated by the input hypothesis, language input one level beyond the learners' current level (i + 1 input) is adequately understandable for learners and may result in language development. The authentic input learners encounter while living and studying in an ESL context seems comparable to the "just beyond input" in Krashen's theory and, therefore, has the potential to trigger the meaning-negotiation mechanism required for producing high-quality output. The massive, multi-faceted exposure to real-life language input also signifies the type of exposure characterized by critical mass and language exposure hypotheses, a milestone in lexical and syntactic development (Zhang & Kang, 2022). In contrast, EFL writers who mainly lack frequent multimodal exposure to the English language may find it challenging to produce the output required to be regarded as English-speaking community insiders (Gil & Caro, 2019). The empirical data showing improper use of LBs in texts developed by novice or non-native L2 learners (e.g., Chen & Baker, 2010; Meunier & Granger, 2008) could be an inevitable consequence of flawed exposure to authentic, enhanced input.

Overview of the Earlier Studies

A cursory look at the numerous early (e.g., Bibber et al., 1999, 2004; Cortes, 2004; Hyland, 2008a, 2008b; Wray, 2002) and recent (e.g., Ghorbani et al., 2022; Yakut et al., 2021) evidential data on formulaic language use in oral and written discourse suffices to claim that the researchers and scholars of the field have given the issue saturation coverage. The fact that LBs variations have been explored concentrating on various defining factors, such as writers' L1 (Lu & Deng, 2019), discipline (Hyland, 2008a), register (Biber & Barbieri, 2007), genre (Gholaminejad, 2021), writing expertise (Jalali, 2009), time-based variations (Hyland & Jiang, 2018), and language proficiency (Wei & Lei, 2011) reinforce the wealth of investigation in the field. Despite the differences in scope and methodology, one common consensus among the research studies enumerated above was that the function and structure of LBs in discourse are contingent on a collection of contextual, interpersonal, and textual factors. This conclusion, however, does not negate the existence of minor commonalities in frequently used LBs in structure and function.

There is also a heavy load of research on the use of LBs in EAP texts of various academic genres, including argumentative essays (Karabacak & Qin, 2013), research articles (Chen and Baker, 2010), BA theses (Dontcheva-Navratilova, 2012), textbooks (Liu & Chen, 2020), and MA theses and PhD dissertations (Hyland, 2008a). The bearings of these studies included genre-specific lists of high-frequently LBs occasionally accompanied by comparative structural and functional schemes. Many comparative studies also focused on the similarities and differences between L1 and L2 academic texts in LBs use (e.g., Amirian et al., 2013; Ghorbani et al., 2022; Shahmoradi et al., 2021; Yakut et al., 2021) and different academic genres (e.g., Jalali, 2013; Shirazizadeh & Amirfazlian, 2021). The comparative results implied that the native/non-native dichotomy and genre variations could substantially account for the differences in frequency, structure, and function of LBs used in academic written discourse. Despite the wealth of research on LBs in EAP texts, there is an apparent lack of focus on the structural and functional patterns of LBs in EFL and ESL writings while being compared with a native baseline.

Method

Corpus

The corpus of the study was composed of three main sub-corpora, namely EFL, ESL, and Native. The EFL corpus included MA theses and Ph.D. desertions written by Iranian students majoring in Applied Linguistics in Iranian universities countrywide. The ESL corpus comprised EAP texts of the same genre (theses and dissertations) written in Applied Linguistics by Iranian MA and Ph.D. students studying in universities of English-speaking provinces of Canada. Having an overall configuration quite similar to that of ESL and EFL Sub-corpora, the baseline corpus, entitled Native, comprised texts written by Canadian students whose L1 was English. The EFL corpus was selected from

the *Irandoc* website, the official, integrated database of the Iranian Research Institute for Information Science and Technology. The electronic files of theses and dissertations written by ESL or native students were downloaded from the *Theses Canada Portal*. Since the texts written by Iranian ESL learners were the most limited-size corpus available electronically in the collection (*Theses Canada Portal*), the ESL corpus was the first sub-corpora chosen through purposive sampling. Aside from centering around Applied Linguistics titles, the sampling method entailed selecting the only works written within the ten recent years (2013 to 2023) that feature a detailed author profile and affiliation. The author profile helped to ensure the authenticity of the corpus selection. The ESL texts that met the inclusion criteria included 19 master's theses and 22 Ph.D. dissertations. The three sub-corpora were intended to be similar in text number; accordingly, random sampling was employed to choose the same number of thesis (19) and dissertations (22) among the whole body of EFL and native resources compatible with the inclusion criteria. Given the lengthy nature of theses and dissertations, the texts in three chapters widely shared between written research reports, including Introduction, Methods, and Discussion/Conclusion(s), were extracted and constituted the analytical basis of the study. The Literature Review and Results chapters were excluded since the former is likely to include direct quotations or other plagiarism manifestations, and the latter regularly comprises numerical and tabular data inappropriate for textual analysis. Table 1 provides an outline of the whole study corpus.

Table 1
Corpus Details

Corpus	Text genre	Text Number	Word Count
EFL	MA	19	156215
	PhD	22	298230
	Total	41	454445
ESL	MA	19	177128
	PhD	22	365287
	Total	41	542415
Native	MA	19	168745
	PhD	22	390381
	Total	41	559126
The Whole Corpus		123	1555986

Research Design

The current corpus-based comparative study employed a quantitative design to explore whether or not the structural and functional use of LBs in academic texts is associated with the language learning context. To this end, a purposive sample of MA theses and PhD dissertations written by Iranian EFL and ESL learners were compared with texts of the same genre developed by their native counterparts. The comparative quantitative approach entailed discovering LBs used throughout the whole corpus, classifying them according to the analytical framework, determining various structural

and functional use distributions, and evaluating the association between patterns of LBs use and language learning context. Despite the parallels between the three sub-corpora, such as text number, genre, and construction, the total word count differed. The between-corpus difference in word count was hardly a cause of concern since the analytical procedure relied upon proportional and expected values of various structural and functional macro and micro categories.

Analytical Framework

Based on Biber et al.'s (2004) bipartite framework of LBs, the classification of LBs in the current study focused on structural and functional variations. Among the many frameworks proposed for the structural and functional use of LBs, the study employed the one proposed by Hyland and Jiang (2018). Based on earlier well-established classification models (e.g., Biber et al., 1999, 2004; Chen & Baker, 2016; Hyland, 2008a, 2008b), the model provided a specific categorical scheme for LBs. The chief rationale behind employing the model was its central focus on LB use in academic-genre English texts. Additionally, grounded on a large corpus of academic writings published in high-ranking research journals, the classification model seemed a valid representative of various functional and structural variations of LBs used in EAP texts. Table 2 displays the macro and microstructures and functions defined in the model. The short forms of every macro-and micro-category are provided in parentheses for later reference.

Table 2
Analytical Framework of the Study

LBs Classification	Category	Sub-category	Example
Structural	Verb phrase-related (VR)	Passive Verb (PV)	*can be noted that*
		Copular be (C be)	*is one of the*
		Imperative (Imp.)	*should note that the*
	Clause-related (CR)	Anticipatory it (Ant. It)	*it follows that the*
		Abstract Subject (AS)	*the goal is to*
		Human Subject (HS)	*one should note that*
		as-fragments (as-f)	*as shown in fig.*
		if-fragments (if-f)	*if and only if*
		there-fragments (there-f)	*there seems to be*
		wh-fragments (wh-f)	*which is to be*
		that-fragments (that-f)	*that need to be*
	Noun/preposition-related (NR)	Noun Phrase with of-Phrase Fragment (of-p)	*the nature of the*
		Noun Phrase with Other Post-Modifier Fragments (other-p)	*the extent to which*
		Prepositional Phrase Expressions (PPE)	*in terms of the*
		Comparative Expressions (CE)	*as far as the*
Functional	Research-oriented	Location (Time and Place) (Loc.)	*at the same time*

LBs Classification	Category	Sub-category	Example
		Procedure (Pro.)	*the role of the*
		Quantification (Quan.)	*a wide range of*
		Description (Des.)	*the structure of the*
	Text-oriented	Transition Signals (TS)	*in addition to the*
		Resultative Signals (RS)	*as a result of*
		Structuring Signals (SS)	*in the present study*
		Framing Signals (FS)	*on the basis of*
	Participant-oriented	Stance Features (SF)	*may be due to*
		Engagement Features (EF)	*as can be seen*

LBs Selection Criteria

In line with the vast majority of studies on LBs in EAP texts (e.g., Chen & Baker, 2010; Hyland & Jiang, 2018; Yakut et al., 2021), the current study focused on 4-word clusters as an optimal class of LBs. 4-word LBs are presumed to be much more manageable than 3-word ones and more inclusive than non-frequent 5-word clusters (Hyland & Jiang, 2018). Relying upon the frequency and range thresholds operationalized by Hyland and Jiang (2018), the 4-word strings that occurred more than 20 times per million words across 20% of the texts in every corpus constituted the LBs list of the study.

Computational Tools

AntConc (version 4.1.1), one of the latest versions of a well-established shareware text analysis toolkit released in 2022, was employed to identify the 4-word LBs used in the whole corpus under investigation. Since its official launch in 2002, the toolkit has been released in more than 60 versions and has been employed by a large body of researchers in different contexts (e.g., Jalali & Zarei, 2016; Cao, 2021). Featuring the N-Grams and Key-Word-In-Context (KWIC) tools, the software made it possible to scan each sub-corpora for LBs of the target word lengths and show each result in a concordance format, which allows context-based analysis of LBs' functions. The original corpus included either PDF or Microsoft Word (.docx) files. Hence, a freeware tool called AntFileConverter (version 2.0.2) was used to convert the files into .txt format, a plain text file format recognizable by most applications and operating systems (OSs), such as AntConc.

Data Collection Procedure

The process of LB identification commenced after converting the three sub-corpora to AntConc's executable format (.txt). As the preliminary step, the texts were cleaned, removing oddities such as the odd foreign (non-English) letters, mathematical/numerical symbols, notes, page numbers, pictures, tables, diagrams, direct quotations, running heads, and pre-fabricated titles (e.g., statement of the problem). Every corpus was then scanned for 4-word LBs, setting the cut-off frequency and range values based on the pre-determined, standardized range and frequency thresholds. Two experts in Linguistics,

fully cognizant of the analytical model, were invited to code all the LBs independently to maximize the classification reliability. The inter-rater reliability coefficients (.94 for structural and .81 for functional classification) implied a high level of inter-rater agreement. As the initial step, the analysts examined the LBs list and excluded the context-specific noun phrases (e.g., *English as a foreign language*) and overlapping LBs (e.g., *due to the fact* and *the fact that*, as overlapping clusters constituting *due to the fact that*). The analysts then grouped the LBs under the macro and micro-structures and functions in the analytical framework. They subsequently evaluated the types (i.e., the frequency of various LBs) and tokens (i.e., the total occurrence of LBs across the corpus) of different functional and structural categories and sub-categories. In the rare cases of between-rater non-conformity, a third expert was consulted to resolve the emerging conflicts.

Data Analysis Procedure

The first analytical step entailed estimating the proportion of using various macro- and micro-structures and functions in the three sub-corpora through a cross-tabulation procedure. The cross-tabulation analysis was performed on both LB types and tokens to provide an in-depth picture of the comparative results. Given that the data included raw frequencies and percentages of various LB types, Chi-square tests were carried out to ascertain whether or not the structural or functional use of LBs was significantly associated with the language learning context where the texts were developed. The statistically significant Chi-square values were examined further through post-hoc residual analysis. Some screenshots of the GUI (Graphic Use Interface) of the freeware used in the current study (AntConc 4.1.1) are provided in the appendix to exemplify the process of importing the texts into the application and the computational analysis output. The sample screenshots are concerned with the native corpus.

Results

Results Related to the First Research Question

Table 3, a 3 x 3 contingency table, displays the occurrence frequency (N) and percentage (%) of the three macro-structures in each sup-corpus of the study. As shown in Table 3, NR was the macrostructure that fitted most of the LB types (EFL: 73.9%, ESL: 82.7%, and Native: 79.3%) and LB tokens (EFL: 80.6%, ESL: 88.2%, and Native: 81.1%) in the three sub-corpora. CR and VR followed NR, the widely-used macro-structure in all three sub-corpora. The pattern of using the three macro-structures seems partially similar between the sub-corpora.

Table 3
Cross-tabulation Results for the Macro Structures Used in the Three Sub-corpora

Corpus	Variable (Var.)	Descriptive Statistics (DS)	Macro Structure			Total
			NR	CR	VR	
EFL	LBs	N	116	26	15	157

	Type	%	73.9	16.6	9.6	100
	LBs	N	4857	756	407	6029
	Token	%	80.6	12.7	6.8	100
ESL	LBs	N	124	16	10	150
	Type	%	82.7	10.7	6.7	100
	LBs	N	4253	337	231	4821
	Token	%	88.2	7.0	4.8	100
Native	LBs	N	149	22	17	187
	Type	%	79.3	11.7	9	100
	LBs	N	4580	615	450	5645
	Token	%	81.1	10.9	8.0	100

The cross-tabulation analysis was also performed on LBs types and tokens representing the micro-structure used in the three sub-corpora. The results are displayed in Table 4. Before conducting the cross-tabulation analysis, four of the micro-structures underlying CR (i.e., as-f, wh-f, that-f, and there-f), whose frequency of occurrence was either zero or close to zero, were merged into a single category, namely prefix-fragments (PF). The microstructure if-f was earlier excluded from the analysis since none of the LBs in the whole corpus met such structure. This modification policy helped to reach a contingency table in which lower than 20% of cells had expected frequencies lower than five.

Table 4
Cross-tabulation Results for the Microstructures Used in the Three Sub-corpora

Micro Strategy	DS	Corpus					
		EFL		ESL		Native	
		Type	Token	Type	Token	Type	Token
PPE	C	48	2255	62	2203	82	2739
	%	30.6	37.4	41.3	45.7	43.6	48.5
of-p	N	43	1690	41	1304	43	1059
	%	27.4	28.0	27.3	27.0	22.9	18.8
other-p	N	20	759	16	501	15	502
	%	12.7	12.6	10.7	10.4	8.0	8.9
CE	N	5	153	5	245	9	280
	%	3.2	2.6	3.3	5.1	4.8	5.6
PV	N	8	181	7	162	13	348
	%	5.1	3.0	4.7	3.4	6.9	6.2
C *be*	N	7	226	3	69	4	102
	%	4.5	3.7	2.0	1.4	2.1	1.8
Ant. *it*	N	10	322	7	182	10	302
	%	6.4	5.3	4.7	3.8	5.3	5.3
AS	N	11	325	0	0	2	48
	%	7.0	5.4	0	0	1.1	0.8
HS	N	2	49	5	72	5	146
	%	1.3	.8	3.3	1.5	2.7	2.6
PF	N	3	69	4	83	5	119
	%	1.9	1.2	2.7	1.7	2.7	2.1

Based on the results in Table 4, three sub-categories of NR, including PPE, of-p, and other-p, were the micro categories with the highest proportion of occurrence within all three sub-corpora, respectively. Of all the NR structures, CE was used very occasionally in the three sub-corpora. Between the two VR sub-categories (i.e., PV & C be), PV was the most-favored structure type in the ESL and Native sub-corpora. Concerning the EFL corpus, the LBs type with PV and C *be* micro-structures (PV: 5.1%, C be: 4.5%) followed a pattern similar to the other two sub-corpora. Conversely, the LBs token with the C *be* structure in the EFL corpus (3.7%) exceeded those with the PV structure (3%), indicating that EFL writers used iterative instances of the same-type copular *be* structure. Compared to the micro-structures discussed above, those underlying CR (i.e., Ant. *it*, AS, HS, and PF) were found to be more different across the three sub-corpora. The AS (Type: 7%, Token: 5.4%) and HS (Type: 1.3%, Token: .8%) were the most- and least-favored CR strategies in the EFL corpus. In contrast, Ant. *it* was the most-used CR structure among the ESL (Type: 4.7%, Token: 3.8%) and Native (Type: 5.3%, Token: 5.3%) writers. AS was the absent and least-favored (Type: 1.1%, Token: .8%) CR structure in the ESL and Native sub-corpora, respectively. In sum, among the three sub-corpora, the LBs used in the ESL and Native ones shared a closer similarity in using the LBs micro-strategies.

Separate cases of the Chi-square test were performed on the raw frequencies representing macro-and micro-strategies to examine whether or not the association between language learning context and structural use of LBs was statistically significant (see the results in Table 5). The tests were performed on data representing LB types (not tokens) since in cases with large sample sizes (N > 500), Chi-square tests are almost always significant (Gravetter & Wallnau, 2013).

Table 5
Chi-Square Tests' Results for the Association between Learning Context and LBs Structure

Variable	N of Valid Cases	Pearson Chi-square Value	*df*	Sig. (2-sided)
Macro-strategies	495	4.05	4	.399
Micro-strategies	495	32.69	18	.018

The results in Table 5 testified to a significant association between language learning context and the LBs micro-strategies (Pearson $\chi^2 = 32.69$, $df = 18$, $p = .018$), but a non-significant association between language learning context and the LBs macro-strategies (Pearson $\chi^2 = 4.05$, $df = 4$, $p = .399$). Given the significant Chi-square results for the micro-structures, a residual analysis was carried out to identify the specific cells in the contingency table (Table 4) that made the highest contribution to the significant results. The results are shown in Table 6.

Table 6
Adjusted Residuals for the Association between Learning Context and LBs Micro-structures

Corpus	Micro-strategy									
	PPE	of-p	other-p	CE	VP	C be	Ant. it	AS	HS	PF
EFL	-1.6	.6	1.2	-.5	-.4	1.5	.9	**4.2**	-1.4	-.5
ESL	.8	.6	.2	-.4	-.4	-.7	-1.2	**-2.4**	1.6	.2
Native	1.7	-1.1	-1.3	.9	.9	-.7	.2	-1.7	-.2	.3

The absolute values of the adjusted residuals of two cells in Table 6 (marked in boldface) exceeded 2, indicating the significant contribution of these cells to the statistically significant chi-square results (Agresti, 2007). The two values were associated with the AS microstructure. The cell associated with the AS use in the EFL (4.2) had a positive value, indicating that the proportion of LBs of AS structure in the EFL corpus was more than that expected by chance. On the other hand, the adjusted residual value associated with AS in the ESL corpus (-2.4) was negative, implying that the proportion of this micro-structure in the ESL corpus was lower than the expected value.

Results Related to the Second Research Question

Table 7 displays the occurrence frequency and percentage of each macro-function in the three sub-corpora of the study.

Table 7
Cross-tabulation Results for the Macro-functions Used in the Three Sub-corpora

Corpus	Var.	DS	Macro Structure			Total
			RO	TO	PO	
EFL	LBs Type	N	50	50	17	157
		%	57.3	31.8	10.8	100
	LBs Token	N	3378	2177	474	6029
		%	56.0	36.1	7.9	100.0
ESL	LBs Type	N	87	47	16	150
		%	58.0	31.3	10.7	100
	LBs Token	N	2361	2033	427	4821
		%	49.0	42.2	8.9	100.0
Native	LBs Type	N	136	35	17	188
		%	72.3	18.6	9.0	100
	LBs Token	N	3602	1556	487	5645
		%	63.8	27.6	8.6	100

The results in Table 7 showed that the majority of the LB types in each sub-corpora (EFL: 57.3%, ESL: 58%, and Native: 72.3%) fulfilled the RO function. The pattern was the same regarding the LBs tokens, with the only difference that lower than half of the LBs tokens in the ESL corpus (49%) fulfilled this macro-function. TO (EFL: 31.8%, ESL:

31.3%, and Native: 18.6%) followed this dominant function in all three sub-corpora. Finally, PO was the function with the lowest occurrence proportion in all sub-corpora (EFL: 10.8%, ESL: 10.7%, and Native: 9%). Despite the abovementioned commonality in the macro-functional pattern, pair-wise comparison of the sub-corpora in LBs type and tokens of the three macro-functions implied greater levels of similarity between the EFL and ESL contexts, compared to other possible pairs (i.e., EFL vs. Native and EFL vs. Native), since the differences between the RO and TO types and tokens was much more conspicuous in the native corpus than the other two ones.

The cross-tabulation analysis was performed on the frequency data representing various micro-functions to further analyze the functional patterns of using LBs in the whole corpus. Based on the frequency and percentage values in Table 8, the highest proportion of the LB types (EFL: 31.8%, ESL: 40%, and Native: 45.7%) and tokens (EFL: 30.4%, ESL: 30.7%, and Native: 38.9%) in the three sub-corpora used to fulfill the Pro. function. Nonetheless, the proportion of the LB types that fulfilled this function in the Native corpus (45.7%) seemed remarkably higher than that of the EFL Corpus (31.8%). On the other hand, the lowest proportion of the LB types (EFL: 1.9%, ESL: 2%, and Native: 0%) and tokens (EFL: 1.8%, ESL: 1.3%, and Native: 0%) were used to fulfill the EF function. As revealed by comparing the two non-native corpora with each other and the native baseline, the between-corpus differences in FS, RS, and Des. seemed noteworthy.

Table 8
Cross-tabulation Results for the Micro-functions Used in the Three Sub-corpora

Micro Function	DS	Corpus					
		EFL		ESL		Native	
		Type	Token	Type	Token	Type	Token
Pro.	C	50	1831	60	1481	86	2195
	%	31.8	30.4	40.0	30.7	45.7	38.9
Loc.	N	5	389	10	350	16	538
	%	3.2	6.5	6.7	7.3	8.5	9.5
Quan.	N	11	484	13	449	21	602
	%	7.0	8	8.7	9.3	11.2	10.7
Des.	N	24	674	4	81	13	267
	%	15.3	11.2	2.7	1.7	6.9	4.7
TS	N	13	651	10	554	11	495
	%	8.3	10.8	6.7	11.5	5.9	8.8
RS	N	15	675	6	237	6	285
	%	9.6	11.2	4.0	4.9	3.2	5.0
SS	N	12	452	8	295	7	333
	%	7.6	7.5	5.3	6.1	3.7	5.9
FS	N	10	399	23	947	11	443
	%	6.4	6.6	15.3	19.6	5.9	7.8
SF	N	14	366	13	366	17	487
	%	8.9	6.1	8.7	7.6	9.0	8.6

EF	N	3	108	3	61	0	0
	%	1.9	1.8	2.0	1.3	0	0

The Chi-square test results on the frequency data representing the macro-and micro-functions (see Table 9) showed significant associations between language learning context and pattern of using LBs both on macro-functional (Pearson χ^2 = 11.69, df = 4, p = .020) and micro-functional (Pearson χ^2 = 49.34, df = 18, p = .000) levels.

Table 9
Chi-Square Tests' Results for the Association between Learning Context and LB's Function

Variable	N of Valid Cases	Pearson Chi-square Value	*df*	Sig. (2-sided)
Macro-functions	495	11.69	4	.020
Micro-functions	495	49.34	18	.000

Table 10 depicts the adjusted residuals calculated to ascertain which functional sub-categories yielded the significant Chi-square values. According to the results, six cells had adjusted residual values beyond the +/- 2 criteria. The LBs used in the EFL corpus to fulfill the Des. (3.9) and RS (2.7) functions were significantly more than the expected values. Similarly, the LBs fulfilled FS in the ESL and Pro. in the Native corpus were more than the expected-by-chance value. Conversely, the LBs that fulfilled the Pro. And Des. functions in the EFL and ESL corpora, respectively, were lower than the expected frequencies. All the cells enumerated above may contribute to the significant association between language learning context and functional use of LBs.

Table 10
Adjusted Residuals for the Association between Learning Context and LB's Micro-functions

Corpus	Micro-strategy									
	Pro.	Loc.	Quan.	Des.	TS	RS	SS	FS	SF	EF
EFL	**-2.4**	-1.9	-1.1	**3.9**	.8	**2.7**	1.5	-1.3	.0	1.0
ESL	.1	.2	-.2	**-3.0**	-.1	-.9	-.1	**3.3**	.1	1.1
Native	**2.2**	1.6	1.3	-.9	-.7	-1.7	-1.3	-1.9	.1	-1.9

Discussion

Findings Related to the Structural Patterns of LBs

The first inquiry of the study probed into the structural patterns of LBs used in the three sub-corpora under investigation to ascertain whether or not the structural use of LBs was associated with the type of language learning context. The proportional comparison of the 4-word LBs in the three sub-corpora revealed a substantial between-corpus similarity in using the three macro-structures of the study. More specifically, the results indicated that noun/preposition-related LBs were used dominantly throughout the three

corpora. This dominant structure was followed by clause-related and verb phrase-related ones in all sub-corpora under investigation. The Chi-square test results revealed that the proportion of LBs in the three structural categories was significantly independent of the contextual variations in the corpus. The novel scope of the study, which focused on both modes of a non-native L2 learning context while comparing the structural and functional use of LBs in non-native vs. native EAP texts, hindered establishing the meaningfulness of this finding in light of earlier empirical data. Nonetheless, the evidential data showing a similar structural pattern between native and non-native EAP corpora (e.g., Amirian et al., 2013; Hyland & Jiang, 2018; Shahmoradi et al., 2021) might endorse the between-corpora commonalities on a macro-structural level.

As the follow-up results revealed, the significant uniformity among the corpora in the structural use of LBs stemmed from equivalent proportions of using various micro-structures underlying noun/preposition-and verb phrase-related LBs. Bearing a marked similarity to both ESL and Native sub-corpora, the EFL corpus included a high proportion of prepositional phrase expressions (e.g., *in the current study, on the other hand)* as the most-favored structural category throughout their texts. Noun phrases with of-phrase fragments (e.g., *findings of the study, the effect of the*) constituted the second micro-structure with the highest proportion in the three corpora. The corpora were also homogeneous in using noun phrases with other post-modifier fragments (e.g., *the extent to which, the fact that they*) and comparative expressions (e.g., *as well as their, the same time as).*

Notwithstanding the minor differences in the structural model employed in the current study and the ones underpinned most of the earlier studies, the ascendency of phrasal LBs over clausal ones, as the shared findings of these studies (e.g., Biber et al., 2004; Biber, 2006; Biber & Barbieri; 2007; Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a, 2008b) corroborated the dominance of the noun/prepositional phrases in the three sub-corpora. Even studies that testified to the remarkable structural differences between EFL and native academic writings in using LBs (e.g., Amirian et al., 2013; Bao & Liu, 2021; Ghorbani et al., 2022; Yakut et al., 2021) introduce (preposition/noun) phrasal clusters superior to their clausal counterparts. The chief rationale for the fixed, dominant position of phrasal LBs in EAP texts of different genres, disciplines, and cultural backgrounds is that this specific structure suits two ultimate objectives of academic discourse: knowledge transmission and meaning conveyance (Hyland, 2008a; Swales, 2008).

The pattern of using verb phrase-related micro-structures was also very similar between the sub-corpora since the native and non-native writers used a partially higher proportion of LBs types of passive structure (e.g., *were found to be, is referred to as*) compared to that of the copular be structure (e.g., *are in line with, is one of the*). Following the same pattern, none of the three corpora used imperative LBs. The partial ascendency of passive verbs in all sub-corpora of the study lent additional support to the earlier empirical data (e.g., Biber et al., 1999; Chen & Baker, 2010; Hyland, 2008a) that point to this structural frame as the dominant verb-phrase-related structure.

Despite the between-corpus similarity in the overall structure and the micro-structures underlying verb phrase-and noun/preposition-related LBs, the three corpora differed significantly in the clause-related microstructures. The difference in Iranian EFL and ESL writers' preferences for abstract vs. human subject LBs was essential in the significant microstructural differences. Abstract subject LBs (e.g., *the study aimed to, the current study explored),* as the highest proportion of clause-related LBs in the EFL corpus, were absent or remarkably infrequent in the ESL and Native corpora, respectively. Instead of this structure widely preferred by EFL writers, those studying in an ESL context used human subject LBs (e.g., *I would argue that* or *I sought to examine)* in much the same way as their counterparts did. The finding is consistent with those of a couple of earlier studies (e.g., Bao & Liu, 2021; Li et al., 2018), showing that contrary to natives, EFL learners avoid using subject clause structures with a first-person singular subject (*I*) to diminish authorial stance in their discourse. Bao and Liu (2021) also showed that EFL learners avoid using the first-person plural subject (*we)* to refer to the communities where they belong, owing to the restriction of authorial stance and discourse subjectivity. Drawing on the latest versions of research documentation standards, ESL and native writers are more likely to be aware of the latest rhetorical conventions. On the contrary, EFL writers who mainly rely upon prototypical theses and electronically-available exemplars are more prone to adhere to outdated documentation conventions, such as substituting the human subjects *I* and *We* with abstract ones, such as the study and the research.

Among the clause-related micro-structures, anticipatory *it* LBs (e.g., *it is important to*, *it is possible to*) were of considerable interest to both non-native (EFL and ESL) and native writers. Nevertheless, the similarity between the ESL and native corpora in using this structure was more striking. This sort of bundle was also a pervasively used clause-related category in Jalali and Zarei's (2016) and Hyland's (2008b) studies on using LBs in published and postgraduate EAP texts. Thanks to their hedging role (Hyland, 2008a), anticipatory *it* clauses have been regarded as indicative of competent writing (Biber at, 1999; Biber & Barbieri, 2007). As a hypothetical justification, the extensive exposure of less proficient writers to the bundles representing this metadiscoursal structure in published works by competent writers may have provided room for the pervasive use of the structure in EAP texts written by EFL and ESL learners. The marked similarity between the ESL and native corpus in using the anticipatory *it* structure may be attributed to the higher frequency and quality of exposure to native-like and native academic discourse.

One plausible explanation for the great deal of structural commonality between the non-native corpora, as well as between the non-native corpora and the native baseline, could be the processing virtue of LBs, called formulaic advantage (Conklin & Schmitt, 2008). Stored, retrieved, and (re)processed as ready-to-use structural wholes (Wray, 2002), LBs are likely to be employed by even those writers who are not fully aware of their functional and rhetorical features. From a theoretical perspective, this holistic processing advantage seems to yield an optimal structural approximation concerning

those infrequent LBs which are structurally complete (i.e., *on the other hand*, *at the same time*) (Jeong & Jiang, 2019). Nonetheless, adequate exposure to language output replete with these formulaic language building blocks seems to compensate for the structural dependency of the vast body of in-complete LBs (Northbrook et al., 2021). Non-native writers on the threshold of research report development inevitably encounter extensive literature on their area of inquiry. This extensive exposure experience may trigger the formulaic advantage and improve the chance to use phrases structurally similar to those of natives.

Findings Related to the Functional Patterns of LBs

The second research question explored the association between language learning context and the functional use of LBs in EAP texts. As revealed by the results, the most and least-fulfilled functions in the three sub-corpora were the research-oriented and participant-oriented functions, respectively. The meaningfulness of the functional pattern could hardly be reinforced or challenged in light of the previous studies in the field since research on the functional use of LBs in academic texts appears deeply split on the mostly fulfilled macro function in EAP texts. Some evidential data corroborate the current study that the highest proportion of LBs in EAP texts are employed to fulfill research-oriented functions (e.g., Amirian et al., 2013; Shahmoradi et al., 2021). On the other hand, a great deal of evidence introduces the text-oriented function as the most dominant category (e.g., Hyland, 2008b; Ghorbani et al., 2022; Lu & Deng, 2019; Yakut et al., 2021). There is also evidence of the cultural-based nature of functional patterns in academic texts. For instance, Güngör and Uysal (2016) showed the ascendency of text-oriented LBs over research-oriented ones in an L2 corpus and vice versa in an L1 baseline. The inferiority of participant-oriented bundles in both non-native and native corpora of the study was consistent with all studies mentioned in this paragraph.

As the residual analysis results showed, the functional differences mainly stemmed from the differences in using procedural and descriptor LBs, two research-oriented micro-functions, and resultative and framing signals, two text-oriented micro-functions. Based on the results, the native and EFL writers' use of procedural chains (e.g., *the design of the*, *the process of the*) was higher and lower than the expected value, respectively. There was also another functional disequilibrium between the EFL and ESL corpora in using descriptors (e.g., *the age of the*, *the characteristics of the*), showing the EFL writers' tendency against the ESL writers' reluctance to use LBs influential in describing research and its underlying elements. Similar to the current study, procedural LBs were the most common micro-function in Ghorbani et al.'s (2022) EFL corpus of academic texts.

The other differentiating functional features included resultative and engagement features. The EFL corpus contained a heavier than expected use of resultative signals (e.g., *according to the results*, *given the fact that*). In contrast, the ESL writers mainly used framing signals (e.g., *in the case of*, *with regard to the*) to fulfill the text organization function. The native writers, on the other hand, used partially similar proportions of the four text organizer functions, including transition (e.g., *on the other hand*), resultative

(e.g., *due to the fact*), structuring (e.g., *in the next chapter*), and framing signals (e.g., *in the form of*). The pervasive use of causative signals, a micro-function comparable to resultative signals, has been validated as a text-organizing strategy used by EFL learners in lengthy research reports (Allen, 2009; Bao & Liu, 2021; Qin, 2014). The finding could be supported by Ghorbani et al.'s (2022) research, showing the tendency of Iranian writers to use resultative LBs in their written discourse.

Compared to research-and text-oriented functions, which differentiated between the sub-corpora, the sub-categories of participant-oriented function (i.e., engagement and stance features) followed the same pattern. Nonetheless, contrary to some of the previous evidential data (e.g., Ghorbani et al., 2022; Hyland, 2008a) that testified to the superiority of engagement features over stance ones in research articles, the current study's corpora, irrespective of their contextual variations, all included a higher proportion of stance features. The finding was consistent with that of Yakut et al. (2021), showing the ascendency of stance features over their engagement counterparts in doctoral dissertations by native and non-native writers. Reflecting writers' concern for conveying their content evaluation, the heavier use of stance features in research reports does not necessarily negate report developers' reluctance to address their potential readers throughout the text. To endorse the reverse order of stance and engagement markers in research articles and reports, it suffices to consider that research articles are reader-focused texts written to be read by the target discipline-specific community members. On the contrary, research reports have a more limited range of direct audiences and are mainly developed to be evaluated by supervisors and referee boards.

In a nutshell, the differences in several micro-functions between the non-native corpora and the native baseline, as discussed above, contributed to significant between-corpus differences not only on a micro-functional level but also on a macro-functional one. This finding seems in line with the empirical data showing functional differences in LBs used by native and non-native (EFL) writers in academic texts (e.g., Amirian et al., 2013; Bao & Liu, 2021; Yakut et al., 2021). Taking a long, hard look at each of the differentiating micro-functions and the peculiarities of the learning contexts under investigation may help to justify the significant functional differences between the three sub-corpora. Feeling quite confident about developing a coherent discourse that facilitates effective knowledge transition, native writers tend to invest their time and cognitive effort in describing, situating, and contextualizing the research. Along with research contextualization concerns, those struggling with academic writing in an ESL context may have regard for a native-like organizational scheme. As a result, they exploit their phraseological lexicon to effectively and coherently convey their intended meanings. The ESL writers' tendency to overuse framing signals may signify this genuine concern. On the other hand, those struggling with writing in institutional language learning (EFL) are desperate to strike a balance between meaning construction and research contextualization, using higher proportions of resultative and descriptor LBs.

Conclusion

This study sought to elucidate how the peculiarities of the two distinct non-native language learning contexts contribute to lexical variations on a phraseological level. To this end, the 4-word LBs used in Applied Linguistics research reports written by Iranian EFL and ESL learners at MA and Ph.D. levels were analyzed in structure and function compared to a structurally similar native baseline. The results helped to conclude that LBs used in research reports developed by Iranian EFL learners could structurally approximate the same-genre written discourse composed by their ESL and native counterparts, probably thanks to their immense exposure to discipline-specific EAP texts. Nonetheless, infrequent clausal LBs, such as abstract/human subject clauses, are prone to violate the expected structural conformity. As a general rule of thumb, the formulaic advantage of adequate exposure to LBs might be responsible for efficient storage and retrieval of salient, commonly-used LB micro-structures. On the other hand, the functional variations in lexical bundle use caused by contextual factors, though much more severe between EFL and native writings, may even differentiate between texts written by EFL and ESL writers of the same cultural backgrounds. The functional variations in texts developed in both non-native learning modes (EFL and ESL), compared to those written by English-speaking natives, are likely to be manifested through a heavier load of descriptor, resultative, or framing LBs overused to maintain discourse legitimacy and comprehensibility.

Aside from narrowing the substantial gap in the literature for an inclusive operationalization of the non-native learning context, the detailed comparative profile of LBs plotted by the current study might bear new insights for EFL academic writers engrossed in the secret of developing commendable EAP texts. Previous research on native vs. non-native lexical variations highlighted the differentiating structures and functions of LBs in EAP texts by native and EFL writers. Nevertheless, the insights provided by earlier evidential data could hardly propose implications for academic writing pedagogy owing to the cultural and L1 differences between native and non-native writers. Turning the spotlight on the structural and functional characteristics of LBs used in two sorts of culturally-parallel non-native writings, as compared with the standard (native) conventions of formulaic language use, the current study could enlighten those in charge of EFL academic writing programs about the significance of simulating the language input and exposure that dominate English-mediated contexts worldwide. The simulated model might be more likely to meet with success when strategies for encouraging extended exposure to multimodal use of formulaic language accompany input-enhancement policies aimed at promoting the salience of the scarce micro-functions in which ESL and native writings varied. The need for theorizing, contextualizing, and implementing these formulaic enhancement strategies may act as fertile ground for further domain-relevant investigation. Undoubtedly, this novel attempt to explore the role of contextual influences in formulaic language variations needs replication and expansion to produce conclusive remarks. Those interested in the topic may have due foresight to

overcome the chief limitation of the current study, that is, the lack of control over interpersonal differences between writers, such as gender and writing proficiency.

Acknowledgments

We would like to thank the editorial team of TESL Quarterly for granting us the opportunity to submit and publish the current synthesis. We would also like to express our appreciation to the anonymous reviewers for their careful, detailed reading of our manuscript and their many insightful comments and suggestions.

Declaration of conflicting interests

The authors declare no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for this article's research, authorship, and/or publication.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley.

Allen, D. (2010). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education*, *1*, 105–127.

Azadnia, M. (2021). A corpus-based analysis of lexical richness in EAP texts written by Iranian TEFL students. *Teaching English as a Second Language Quarterly*, *40*(4), 61-90.

Amirian, Z., Ketabi, S., & Eshaghi, H. (2013). The use of lexical bundles in native and non-native postgraduate writing: The case of applied linguistics MA theses. *Journal of English Language Teaching and Learning*, *5*(11), 1–29.

Bao, K., & Liu, M. (2022). A corpus study of lexical bundles used differently in dissertations abstracts produced by Chinese and American PhD students of linguistics. *Frontier in Psychology*, *13*, 1–13. https://doi.org/10.3389/fpsyg.2022.893773

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*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell, (Eds.), *Out of corpora: Studies in honor of Stig Johansson*, (pp.181–189). Rodopi.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Bhatia, V. K. (2014). *Worlds of written discourse: A genre-based view*. Bloomsbury Publication.

Brown, H. D. (2001). *Teaching by principles*. Longman

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. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays. *Applied Linguistics*, *37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Cortes, V. (2002). *Lexical bundles in academic writing in history and biology* (Unpublished doctoral dissertation). Northern Arizona University.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56–71. https://doi.org/10.1515/9783484604674.56

Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speaker *Brno Studies in English*, *38*(2), 37–58. https://doi.org/10.5817/BSE2012-2-3

Ellis, N. C. & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61–78. https://doi.org/10.1515/CLLT.2009.003

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *Tesol Quarterly*, *42* (3), 375–96. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Gao X. (2017). A comparable-corpus-based study on native English and Chinese academic writers' use of English lexical bundles. *For. Lang. Their Teach.* 3, 42–52. https://doi.org/10.1515/CJAL-2019-0029

Gholaminejad, R. (2021). A Comparison of two genres: lexical bundles in the discourse of applied linguistics. *Journal of the Spanish Association of Anglo-American Studies*, *43*(2), 90–109. http://doi.org/10.28914/Atlantis-2021-43.2.05

Ghorbani, A., Okati, F., & Lotfi Gaskaree, B. (2022). Lexical bundles in the conclusion section of applied linguistics articles: a comparative study of international and Iranian journals. *Journal of Critical Applied Linguistics Studies*, *1*(1), 67–84. http://dx.doi.org/10.22034/jcals.1.1.67

Gil, N. N., & Caro, E. M. (2019). Lexical bundles in learner and expert academic writing. *Bellaterra Journal of Teaching & Learning Language & Literature*, *12*(1), 65–90. https://doi.org/10.5565/rev/jtl3.794

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychon. Bull. Rev.*, *20*(6), 1161–1169. https://doi.org/10.3758/s13423-013-0458-4

Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the behavioral science* (9th ed.). Wadsworth.

Güngör, F. & Uysal, H. H. (2016). A comparative analysis of lexical bundles used by native and non-native scholars. *English Language Teaching*, *9*(6), 176–188. https://doi.org/10.5539/ELT.V9N6P176

Håkansson, G., & Norrby, C. (2010). Environmental influence on language acquisition: comparing second and foreign language acquisition of Swedish. *Language Learning*, *60*, 628–650. http://doi.org/10.1111/j.1467-9922.2010.00569.x

Huang K. (2018). Register features of lexical bundles used by Chinese EFL majors: a contrastive analysis of spoken and written English. *For. Lang. World*, 5, 71–79.

Hyland, K. (2008a). As can be seen: lexical bundles and disciplinary variation. *Eng. Specific Purposes*, *27*, 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*, 41–62. https://doi.org/10.1111/j.1473 4192.2008.00178.x

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150-169. https://doi.org/10.1017/S0267190512000037

Hyland, K. & Jiang, F. (2018). Academic lexical bundles. *International Journal of Corpus Linguistics*, *23*(4), 383–407. https://doi.org/10.1075/ijcl.17080.hy

Jalali, H. (2009). *Lexical bundles in applied linguistics: variations within a single discipline* (Unpublished master's thesis). University of Isfahan.

Jalali, H. (2013). Lexical bundles in applied linguistics: Variations across postgraduate genres. *Journal of Foreign Language Teaching and Translation Studies*, *2*(2), 1–29. https://doi.org/10.22034/EFL.2013.79199

Jeong H., & Jiang N. (2019). Representation and processing of lexical bundles: Evidence from word monitoring. *System*, *80*, 188–198. https://doi.org/10.1016/j.system.2018.11.009

Karabacak, E., & Qin, J. (2013). Comparison of lexical bundles used by Turkish, Chinese, and American university students. *Procedia-Social and Behavioral Sciences*, *70*(0), 622–628. http://doi.org/10.1016/j.sbspro.2013.01.101

Kashiha, H., & Chan, S. H. (2014). Cross-linguistic and cross-disciplinary investigation of lexical bundles in academic writing. *Pertanika Journal of Social Sciences & Humanities*, *22*(4), 937–951.

Kazemi, M., Katiraei, A., & Rasekh, E. (2014) the impact of teaching lexical bundles on improving Iranian EFL students' writing skills. *Procedia – Social and Behavioral Sciences*, *98*, 864–869. https://doi.org/10.1016/j.sbspro.2014.03.493

Krashen, S. D. (1985). *The Input Hypothesis: Issues and Implications*. Longman.

Li, L., Franken, M., & Wu, S. (2018). Chinese postgraduates' explanation of the sources of sentence initial bundles in their thesis writing. *RELC Journal*, *50*, 37–52. https://doi.org/10.1177/0033688217750641

Liu, C., & Chen, H. H. (2020). Analyzing the functions of lexical bundles in undergraduate academic lectures for pedagogical use. *English for Specific Purposes*, *58*, 122–137. doi: http://doi.org/10.1016/j.esp.2019.12.003

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 Academic Purposes*, 39, 21–36. http://dx.doi.org/10.1016/j.jeap.2019.03.008

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: a test of the critical mass hypothesis. *Journal of Child Language*, *21*, 339–366. https://doi.org/10.1017/S0305000900009302

Meunier, F., & Granger, S. (2008). *Phraseology in foreign language learning and teaching*. John Benjamins. https://doi.org/10.1075/z.138

Mitchel, A. D., Christiansen, M. H., & Weiss, D. J. (2014). Multimodal integration in statistical learning: Evidence from the McGurk illusion. *Frontiers in Psychology*, *5*, 1–6. https://doi.org/10.3389/fpsyg.2014.00407

Northbrook, J., & Conklin. K. (2019). Is what you put in what you get out? Textbook-derived lexical bundle processing. *Applied Linguistics*, *40*(5), 816–833. https://doi.org/10.1093/applin/amy027

Northbrook, J., Allen, D., & Conklin, K. (2021). Did you see that? The role of repetition and enhancement on lexical bundle processing in English learning materials. *Applied Linguistics*, 1–20. http://doi.org/10.1093/applin/amab063

Ortega, L. (2014). Experience and success in late bilingualism. Paper presented in Keynote address at the 17th AILA world congress, Brisbane, Australia.

Pang, W. (2010). Lexical bundles and the construction of an academic voice: A pedagogical perspective. *Asian EFL Journal*, *47*(1), 10–11.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System*, *42*(1), 220–231. https://doi.org/10.1016/j.system.2013.12.003

Rashtchi, M., & Ali Mohammadi, M. (2017). Teaching lexical bundles to improve academic writing via tasks: Does the type of input matter? *Electronic Journal of Foreign Language Teaching*, *14*(2), 201–219.

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 55–86). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.05sch

Shahmoradi, N., Jalali, H., & Ghadiri, M. (2021). Lexical bundles in the abstract and conclusion sections: the case of applied linguistics and information technology. *Applied Research on English Language*, *10*(3), 47–76. https://doi.org/10.22108/ARE.2021.128024.1703

Shirazizadeh, M., & Amirfazlian, R. (2021). Lexical bundles in theses, articles and textbooks of applied linguistics: investigating intradisciplinary uniformity and variation. *Journal of English for Academic Purposes*, *49*, 100–126. https://doi.org/10.1016/j.jeap.2020.100946

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list (AFL). *Applied Linguistics*, *31*, 487–512.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214-225. https://doi.org/10.1016/j.jeap.2013.05.002

Swales, J. M. (2008). Foreword. In D. Belcher & A. Hirvela (Eds.), *The oral-literate connection: Perspectives on L2 speaking, writing, and other media interactions* (p. v-viii). Universityof Michigan Press.

Vercellotti, M. L. (2015). The development of complexity, accuracy, and fluency in second language performance: a longitudinal study. *Applied Linguistics*, *38*, 90–111. http://doi.org/10.1093/applin/amv002

Vold, E. T. (2022). Learner spoken output and teacher response in second versus foreign language classrooms. *Language Teaching Research*, 1–35. https://doi.org/10.1177/13621688211068610

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. RELC Journal, *42*(2), 155–166. http://doi.org/10.1177/0033688211407295

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. Bloomsbury.

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

Yakut, I., Yuvayapan, F., & Bada, E. (2021). Lexical bundles in L1 and L2 English doctoral dissertations. *Journal of Teaching English for Specific and Academic Purposes*, *9*(3), 475–493. https://doi.org/10.22190/JTESAP2103475Y

Zhang, S., & Kang, C. (2022). A comparative study on lexical and syntactic features of ESL versus EFL learners' writing, *Frontier in Psychology*, *13*, 1–11. http://doi.org/ 10.3389/fpsyg.2022.1002090

Zhou, J., & Lü, C. (2022). Enhancing syntactic complexity in L2 Chinese writing: effects of form-focused instruction on the Chinese topic chain. *Frontier in Psychology*, *13*, 789–843. http://doi.org/10.3389/fpsyg.2022.843789

Appendix

Sample Screenshots of the GUI of the Freeware Used in the Study