National Research University Higher School of Economics
Journal of Language & Education Volume 6, Issue 4, 2020

Fajri, M. S. A., Kirana, A. W., & Putri, C. I. K. (2020). Lexical Bundles of L1 and L2 English Professional Scholars: A Contrastive Corpus-Driven Study on Applied Linguistics Research Articles. *Journal of Language and Education, 6*(4), xxx-xxx. https://doi.org/10.17323/jle.2020.11271

Lexical Bundles of L1 and L2 English Professional Scholars: A Contrastive Corpus-Driven Study on Applied Linguistics Research Articles

Muchamad Sholakhuddin Al Fajri, Angkita Wasito Kirana, Celya Intan Kharisma Putri

Airlangga University

Correspondence concerning this article should be addressed to Muchamad Sholakhuddin Al Fajri, Faculty of Vocational Studies, Universitas Airlangga, Jl. Dharmawangsa Dalam No. 28-30, Surabaya, Jawa Timur, Indonesia, 60286. Email: m-sholakhuddin-al-fajri@vokasi.unair.ac.id

The current study examined the structural and functional types of four-word lexical bundles in two different corpora of applied linguistics scientific articles written by L1 English and L1 Indonesian professional writers. The findings show that L2 writers employed a higher number of bundles than L1 writers, but L2 writers underused some of the most typical lexical bundles in L1 English writing. Structurally, unlike previous studies, this study reports the frequent use of prepositional phrase (PP) - based bundles in the articles of L2 writers. However, besides the high frequency of PP-based bundles, L2 authors also used a high number of verbal phrase-based bundles, suggesting that these L2 writers were still acquiring more native-like bundles. In terms of functional types, L2 writers employed fewer *quantification* bundles than their counterparts. This study has potential implications for teaching English for academic writing. Teachers need to raise their students' awareness of the most frequently used lexical bundles in a specific academic discipline and pay attention to the discourse conventions of academic writing, helping L2 students transition from clausal to phrasal styles.

Keywords: lexical bundles, academic writing, corpus linguistics, applied linguistics

Introduction

Lexical bundles (henceforth LBs) are defined as "recurrent expressions that usually co-occur in natural language use, regardless of their idiomaticity and their lexical status" (Biber et al., 1999, p. 990), and they can be "identified empirically by running a computer program in a corpus of language texts" (Cortes, 2015, p. 205). They are identified automatically by using a computer program with frequency and distribution thresholds set by researchers (Hyland, 2012). LBs play a significant role in improving the quality of scientific writing for both native and non-native speakers. LBs are seen as a significant aspect of fluent linguistic production and a noticeable feature of academic written texts (Hyland & Jiang, 2018). Hyland (2012, p, 153) emphasises the importance of LBs for writers and speakers in three points: "(1) their repetition offers users (and particularly students) ready-made sets of words to work with; (2) they help define fluent use and therefore expertise and legitimate disciplinary membership; (3) they reveal the lexico-grammatical community-authorized ways of making-meanings". LBs are therefore seen to be very important in the formulation of texts.

Regarding the second point, LBs could help writers claim their membership in a particular discourse community (Ädel & Erman, 2012). Wray (2006) explains that, when speaking, people choose a specific turn of phrase that they consider to be related to certain values, styles, and groups. In other words, they help registered community members show solidarity with other members (Esfandiari & Barbary, 2017) and build a disciplinary experienced voice (Pang, 2010). Thus, LBs tend to reflect an authentic part of users' communicative experiences (Hyland & Jiang, 2018).

LBs have also drawn the interest of linguists to explore the role of LBs in teaching and learning academic writing. For instance, in English for Academic Purposes (EAP), exposure to multi-word constructions helps

students gain a better understanding of the language style of academic textbooks (Wood & Appel, 2014) since LBs make up 21-52.3% of written discourse (Biber et al., 1999). The absence of multiword expressions thus might indicate a writer's lack of expertise in academic contexts (Wray, 2002). In other words, LBs enable us to distinguish novice and expert users of a certain language in different contexts both in oral and written forms, which are seemingly useful in teaching and learning activities especially in enhancing speaking and writing skills. Noticing the significance of those studies on LBs, this article aims to explore the application of LBs by professional writers for whom English is their native language (L1) and those for whom English is their L2, or foreign language, in academic articles within the discipline of applied linguistics.

Structural and Functional Characteristics of Lexical Bundles

Most LBs are incomplete structural units that comprise two or more words, and they can be categorised into different types of structures as they have strong grammatical correlations (Cortes, 2004). Biber et al. (1999) categorised the grammatical structures of LBs into three common forms: verb-phrase bundles which refer to any word combinations with a verb component such as *it is also possible, can be noted that,* and *it is likely that*; noun-phrase fragments which refer to any noun phrases with post-modifier fragments such as *the use of the*, *the nature of the*, and *the way in which*; and prepositional phrase bundles which include any bundles starting with a preposition plus a noun-phrase fragment such as *in addition to the*, *in the context of,* and *at the end of*. Different registers require different grammatical structures. For example, Biber et al. (1999) argued that LBs in a conversation contain mostly clause fragments (60%) and only 15% of them were phrases. In contrast, in academic prose, LBs were mostly in the form of phrases and less than 5% were clausal constructions.

In terms of functions, Biber et al. (2004) divided LBs into three main categories: stance expression (e.g. *the fact that the*), discourse organizers (*as well as the*), and referential expression (e.g. *one of the most*). Like in the LBs structures, they also found a dramatic difference between oral and written registers in their dependence on LBs' functional types. Conversations or spoken registers mostly use stance expression bundles, while academic writing mostly uses referential expressions. On the basis of Biber et al.'s taxonomy, Hyland (2008a) developed a similar functional taxonomy including research-oriented, text-oriented, and participant-oriented bundles. Their taxonomies differ in the sense that Biber et al.'s (2004) classification was based on both written and oral registers that covered various genres, while Hyland's (2008a) taxonomy was far more specific, focusing on written registers only. Therefore, this study used the functional taxonomy proposed by Hyland (2008a).

Previous Studies on Lexical Bundles

To discover the application of LBs, a corpus has been widely employed for analysing various types of texts (written and spoken) in different languages (e.g. Kim, 2009; Ruan, 2017; Wang, 2017; Wright, 2019). Biber, Conrad, and Cortes (2004) investigated LBs of oral and written registers, including conversations, lectures, textbooks, and academic prose. In several other academic writing genres, linguists have investigated LBs in theses, dissertations, and students' academic writing in a range of academic disciplines. For example, Hyland (2008a) examined the forms, structures, and functions of four-word clusters in a corpus of research articles, dissertations, and theses in four academic disciplines: engineering, microbiology, business, and applied linguistics, while Cortes (2004) compared research articles and students' writing within the disciplines of history and biology. Broadening the scope of the field of knowledge, Kwary et al. (2017) analysed the use of LBs in journal articles of four wide academic disciplines: life sciences, physical sciences, health sciences, and social sciences. These previous studies generally suggest that LBs vary in their discourse functions and their use differs from one discipline or register to another.

Three studies have explored the use of LBs in L2 writing at different proficiency levels. Staples et al. (2013) investigated LBs used by non-native English speakers with different proficiency levels in prompted TOEFL writing. Their research shows that learners at lower proficiency level preferred to employ more clusters than those at higher proficiency levels, lending weight to the second language acquisition theory that as students acquire more proficiency in a second language, they have a tendency to use fewer formulaic structures (Ellis, 1996). Chen and Baker (2016) studied second language development by comparing the use of LBs in L2 English (L1 Chinese) rated learner essays across three levels of Common European Framework of Reference (B1, B2 & C1). Their findings indicate that learners' writing at lower levels is likely to share more features with conversation, relying more on colloquial quantifiers, while the discourse of more advance writing has a more

impersonal tone, closer to that of academic prose. The findings also show that the CEFR-B2 level seems to be a transition stage in which learners start to recognise the differences between formal and informal writing. In the field of applied linguistics, the LB studies compared English learner writing with professional writing. For instance, Wei and Lei (2011) examined LBs used by advanced Chinese EFL learners and professional writers in the field of applied linguistics, and Qin (2014) compared clusters used by non-native English graduate students at different levels of study and authors of applied linguistics journal articles. These comparisons may make results difficult to interpret because learners and professional writers have different English and writing proficiency levels and specific writing purposes for unique audiences.

Other studies have compared the use of LBs by L1 English and L2 English writers. For instance, Chen and Baker (2010) compared the use of LBs in academic writing by non-native English students with native peer students and native expert writers. They concluded that the structures and functions of LBs in L1 and L2 student writing is similar. However, L2 students tended to underuse some typical bundles in professional academic writing. In a similar vein, Ädel and Erman (2011) analysed the use and the functions of English LBs in advanced undergraduate writing by L1 English and L1 Swedish students. The findings indicated that L1 writers deployed a larger number of LBs with a wider variety, which are generally similar to the results of phraseological analysis tradition in Second Language Acquisition (SLA) (Ädel and Erman, 2011).

There are only a few studies that investigated the use of LBs by L1 and L2 English academic professionals in international journals. Perez-Llantada (2014), for example, analysed the convergent and divergent usage in academic articles from twelve disciplines. However, the wide variety of scientific disciplines is likely to skew the results of the study since registers, genres, and disciplines all affect the structure and function of LBs (Esfandiari & Barbary, 2017). Pan, Reppen, and Biber (2016) compared the use of LBs by L1 and L2 English professional academics in telecommunications articles. Esfandiari and Barbary (2017) analysed the use of LBs by L1 and L2 English professional academics in psychology research articles. However, there is little work that has been devoted to the use of LBs by L1 English and L2 English professional writers, especially L1 Indonesian writers, in the field of applied linguistics. This study tried to fill this gap by analysing and comparing the use, structure, and function of LBs in applied linguistics academic articles written by English professional writers (L1 English) and Indonesian professional writers (L2 English). Examining applied linguistics articles is significant since journal article authors in this field, whose expertise is related to language, seem to be more aware of the use of formulaic language or bundles, which may affect the use of LBs in their writing. This study therefore can contribute to the ongoing discussion regarding the influence of academic disciplines on the use of LBs in professional academic writing.

Methodology

Corpus Construction

The corpora of the present study are a collection of applied linguistic scientific articles published in a three-year period from 2016 to 2018. The decision to include articles from a three-year period intended to mitigate the over-usage of LBs in special issue publications that probably occur in a certain journal, thus avoiding idiosyncrasies of specific issues or topics. The two corpora are the English corpus (EC) comprising articles written by L1 English academics and the Indonesian corpus (IC) consisting of research articles written by L1 Indonesian professional authors.

The EC was collected from scientific research articles published by internationally reputable journals in the field of applied linguistics that have high Impact Factors (IF) and are indexed in the Scopus database and Social Science Citation Index (SSCI). Meanwhile, the IC was taken from research articles published in Indonesian applied linguistics international journals that are indexed by Scopus or Directory of Open Access Journal (DOAJ) and accredited by Indonesian Ministry of Research, Technology and High Education. The selected Indonesian journals only publish articles in English (see Appendix 1 for the journal list). Both corpora have a similar total number of words, with approximately 1,300,000 words in each corpus. The number of words in each corpus was kept equal since LBs are significantly more sensitive to the number of words in a corpus than the number of articles (Cortes, 2004). Therefore, in our corpus there are fewer articles and journals in the EC as the EC articles were typically longer than the IC articles (see Table 1).

Table 1
Distribution of the corpora

Corpus	Number of Journals	Number of Articles	Number of words
English Corpus (EC)	4	158	1,325,986
Indonesian Corpus (IC)	7	274	1,334,752

We considered the selected journals for the IC to be equivalent to the journals for the EC for several reasons. First, the journals for the IC are peer-reviewed journals, following the academic conventions of international journals. Second, the journals in the IC are indexed by international research article databases. Both EC and IC articles consisted of Introduction, Methods, Results/Findings, and Discussion sections (Martinez et al., 2009). Additionally, articles published in Indonesia tend to reflect the language produced by Indonesian authors in Indonesian contexts for international audiences.

To ascertain the first language of the author(s), we followed the method proposed by Wood (2001) which defines L1 English writers as those whose first and last name are considered as typical native English speaker names and those who are affiliated with institutions in countries that use English as their first language. Therefore, L1 Indonesian writers are also categorised as all writers whose first and last names are considered typical Indonesian names and those who are affiliated with Indonesian institutions. We, thus, excluded articles from Indonesian journals where any of the authors did not fulfil both criteria.

Identification of Lexical Bundles (LBs)

This research considered two criteria in identifying LBs, namely frequency and dispersion. For these relatively big corpora, the standardised frequency threshold was set as 40 occurrences per million words to identify bundles that are often considered as characteristics of target texts (Pan et.al., 2016). The cut-off frequency of 40 per million words was equivalent to minimum raw frequency of 53 occurrences for both corpora. This was calculated by multiplying the cut-off frequency by the corpus size and then dividing the result by one million (Wood & Appel, 2014).

The dispersion threshold holds a significant role to avoid individual author idiosyncrasies. Thus, it needs to be clearly determined to guarantee that the bundles are not only used by a handful of authors or texts. Following Chen and Baker (2010) and Hyland (2008a), the current study only included those bundles that occurred in at least 10% of the total texts in each corpus (Hyland, 2008a). The bundles therefore must occur in at least 27 and 16 articles in the IC and EC respectively. Besides, the length of the word sequences (LBs' length) included in the study must also be determined. This study focused on 4-word bundles "because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles" (Hyland, 2008b, p.8).

The corpus software AntConc[1] was used to retrieve the bundles. However, context/content dependent bundles such as *teaching and learning process, as a foreign language* or *in the united states* were excluded since "they are not the 'building blocks' which carry a distinct discourse function" (Chen, 2009, p. 58). In addition, overlapping clusters were also checked manually via concordance analyses to avoid inflated results of quantitative analyses (Chen & Baker, 2010). For example, in the IC *it can be seen* and *can be seen that* occurred as a subset of the 5-word sequence *it can be seen that.* Thus, the lower frequency sequence was combined into the higher frequency one: *it can be seen (that)*. After identifying LBs with the above-mentioned criteria, we compared the frequency, structure, and function of our LBs. In analysing LBs' structures, we used Biber et al.'s (1999) classification which includes noun phrase-based (e.g. *the use of the*), prepositional phrase-based (e.g. *in the case of*), and verb phrase-based bundles (e.g. *it can be seen*). For functional analysis, we employed Hyland's classification (2008b) since it is more relevant to academic writing domain. The classification includes research-oriented bundles, which are used to structure writers' experiences (e.g. *the use of the*), text-oriented bundles, which are concerned with the organisation of the text or discourse (e.g. *in addition to the*), and participant-oriented bundles, which are focused on stance and engagement (e.g. *are likely to be*).

To classify the bundles, each author worked independently and the inter-rater reliability was 97% (structure) and 94% (functions). The discrepancies were then discussed to reach 100% agreement based on their contexts.

[1] Anthony, L. (2018). AntConc (3.5.7) [Computer Software]. Waseda University. http://www.laurenceanthony.net/software

We acknowledged that there were several multi-functional bundles (e.g. *in the present study*), which also have been recognised by the previous studies (Güngör & Uysal, 2020; Salazar, 2014). In these cases, the ambiguous bundles were categorised according to their primary function after cross-checking their contexts.

Results and Discussion

Comparison of Frequency

After excluding the content/context dependant and overlapping bundles, we identified 2,700 and 4,874 lexical bundle tokens in the EC and IC respectively, which comprised 31 different bundle types in the EC, and 51 bundle types in the IC (see Appendix 2 for the list of LBs). This finding is congruent with several previous studies including Pan et al.'s (2016) research, which found that L2 professional academic writers in telecommunications used LBs more frequently than their L1 counterparts, and Güngör and Uysal's (2016) study, which showed that L2 English academic authors in educational sciences used a larger number of LBs than LI English writers. This is, however, contrary to previous studies that compared the use of LBs in the corpora of native and non-native student writing (e.g. Adel & Erman, 2012; DeCock, 2004), in that L2 learners employed a lower number of bundles than native English students. One of the reasons for the lower number of LBs in L2 student writing is the learners' incorrect use of articles (e.g. the omission of required definite articles within LBs) (Shin, Cortes, & Yoo, 2018), which may not apply to L2 professional writing since academics are comparatively competent writers. Thus, it seems that when it comes to professional academic writing, L2 writers including Indonesians are likely to employ a substantially higher frequency of bundle types and tokens than L1 authors. Hyland (2008a) points out that both groups of expert writers use the fewest clusters compared to master's and doctoral students. This indicates that L2 expert writers still rely on formulaic expressions to some extent. This greater reliance on LBs might also suggest the comparatively smaller vocabulary of L2 writers, while L1 professional writers might be able to present their arguments in a more flexible manner.

Additionally, a comparative analysis showed that 14 bundles were shared between both groups (e.g. *on the other hand*, *at the same time,* and *as well as the*), indicating that nearly half of the LBs used by native writers (14 of 31) were also employed by non-native writers. Most of the shared bundles (9 of 14) are text-oriented ones, which might not be surprising as text-oriented bundles are common in soft sciences research articles (Hyland, 2008b). However, only two LBs (*on the other hand* and *the results of the*) were shared among the top ten most frequently used bundles. The top two most frequent LBs in the EC (*the extent to which, in the present study*) were not used by IC writers. Qin (2014) also reported the absence and low frequency of *the extent to which* in the master's and doctoral student writing in applied linguistics. This indicates that these L2 advanced writers were still not aware of some typical academic LBs in the field of applied linguistics.

Comparison of structural types

Table 2 shows the distribution of structural subcategories of the LBs used by both English and Indonesian writers. Log-likelihood tests comparing the number of tokens in each category were conducted to measure significant differences across the corpora. The results demonstrated that IC writers used considerably more VP-based and PP-based bundle tokens than EC writers. For NP-based bundles, although this category does not indicate substantial differences, the subcategory shows a different pattern. L1 English authors employed significantly more NP with other post-modifier fragments, while L1 Indonesian authors used significantly more NP with *of*-phrase. Examples of NP-based bundles are presented below:

> This paper explores *the extent to which* corpus linguistics can contribute to the study of language ideology in both explicit and implicit forms in news media. (EC: NP-other)
>
> Moreover, *the use of the* nomination strategy also indicates that the Jakarta Post tries to avoid ambiguity. (IC: NP-*of*)

In terms of the structural distribution of LBs (see Table 3), the comparison of the percentages of the main structural categories in both corpora shows that EC writers mostly employed phrasal bundles (NP- and PP-based bundles), accounting for 84% of bundle types and tokens. This finding is congruent with previous studies that point out that the frequency and percentage of phrasal bundles are higher than clausal bundles in English academic prose (Biber et al., 2011, 2013; Biber & Gray, 2011), and L1 English professional writers used

considerably more NP- and PP-based bundles than VP-based bundles in academic research articles (Pan et al., 2016).

Table 2
Distribution of structural subcategories

Categories	Subcategories	Types		Tokens		LL
		EC	IC	EC	IC	
NP-based	Noun phrase + of (e.g. *the use of the*)	6	9	464	828	101.57**
	Noun phrase with other post-modifier fragment (e.g. *the extent to which*)	4	1	388	54	286.78**
	Total	10	10	852	882	0.34
PP-based	Prepositional phrase with embedded *of* phrase (e.g. *in the context of*)	10	10	787	1031	31.26**
	Other prepositional phrase fragments (e.g. *in relation to the*)	6	10	635	911	47.74**
	Total	16	20	1422	1942	77.31**
VP-based	copula *be* + noun phrase/prepositional phrase (e.g. *is one of the*)	-	2	-	281	387.70**
	Anticipatory *it* + verb phrase/adjective phrase (e.g. *it can be seen*)	1	7	95	644	453.86**
	Passive verb + prepositional phrase (e.g. *can be seen in*)	2	4	138	362	102.54**
	(Verb/adjective) to-clause fragment (e.g. *to be able to*)	1	3	61	289	159.87**
	(Verb phrase) + *that* clause fragment (e.g. *that there is a*)	-	3	-	268	369.76**
	Total	4	19	294	1844	1241.51**
Others	as well as the	1	2	132	206	15.85**
Total		31	51	2700	4874	-

Table 3
Distribution of structural categories

Categories	Types (%)		Tokens (%)	
	EC	IC	EC	IC
NP-based	32.26	19.61	31.56	19.10
PP-based	51.61	39.22	52.67	39.84
VP-based	12.90	37.25	10.89	37.83
Others	3.23	3.92	4.89	4.23
Total	**100.00**	**100.00**	**100.00**	**100.00**

On the other hand, IC writers predominantly used PP-based and VP-based clusters, accounting for 77% of the types and 78% of the tokens. This result contrasts somewhat with previous studies on writing from other disciplines that revealed that VP-based bundles were more frequent than the other categories (NP and PP) in L2 writing. For example, Pan et al. (2016) found more VP-based bundles (58%) than NP- and PP-based bundles (34%) in L1 Chinese professional writing in telecommunications journals, and Chen and Baker (2010) found more VP-based bundles (52.3%) than NP- and PP-based bundles (47.5%) in L2 learners' writing. With the high frequency of PP-based bundles in the IC corpus, it therefore suggests that L2 professional Indonesian writers in the field of applied linguistics demonstrate relatively higher academic writing proficiency since both L1 and L2 writers will shift from the clausal to phrasal style as their writing skills increase (Bychkovska & Lee, 2017; Pan et al., 2016; Staples et al., 2013). Wei and Lei (2011) also found that advanced Chinese EFL learners in the discipline of applied linguistics used more NP- and PP-based four-word bundles than VP-based formulaic sequences, which is in contrast to previous research on Chinese EFL learner writing in other disciplines (e.g. Bychkovska & Lee, 2017; Pang, 2009). The use of more phrasal bundles may be due to the fact that applied linguistics majors require a more advanced level of English even at the undergraduate level so students and researchers in this field might be more aware of the conventions of English academic register. However, it should be noted that the considerable use of clausal constructions or VP-based bundles in the IC (37% of the types and 38% of the tokens) indicates that these L1 Indonesian expert writers were still in the process of obtaining more appropriate academic English or acquiring more native-like LBs.

Comparison of functional types

As shown in Table 4, the two corpora contained a relatively similar proportion of the three main functional categories. Text-oriented bundles (types and tokens) comprised the largest proportion in both the EC and IC, with similar percentages at 55% and 51% respectively for types, and 57% and 55% respectively for tokens, while participant-oriented bundles constituted the smallest proportion, accounting for 6% of types and tokens in the EC, and 8% of types and 9% of tokens in the IC. Text-oriented bundles are significantly used in applied linguistics journal articles, more generally in the social sciences, to "provide familiar and shorthand ways of engaging with a literature, providing warrants, connecting ideas, directing readers around the text, and specifying limitations", representing the more discursive and evaluative patterns of arguments in the soft sciences (Hyland, 2008b, p. 16). This finding echoes previous related research on LBs used by Persian writers in psychology research articles (Esfandiari & Barbary, 2017), and LBs used by Chinese writers in telecommunications journal articles (Pan et al., 2016). This indicates that L1 and L2 English professional writers do not differ much in the proportion of the main functional distributions of LBs.

Table 4
Distribution of functional types

Categories	Types (%)		Tokens (%)	
	EC	IC	EC	IC
Research-oriented	38.71 (39)	41.18 (41)	36.59 (37)	35.80 (36)
Text-oriented	54.84 (55)	50.98 (51)	57.30 (57)	54.97 (55)
Participant-oriented	6.45 (6)	7.84 (8)	6.11 (6)	9.23 (9)
Total	**100.00**	**100.00**	**100.00**	**100.00**

However, Table 5 shows profound differences in the frequency of bundle tokens across the two corpora in nearly all subcategories of functional types. The results of log-likelihood tests comparing the number of tokens in each category demonstrate that IC (L2) writers used LBs significantly more frequently than EC (L1) writers in most functional subcategories (*procedure, description, transition, resultative, structuring, framing,* and *engagement*), but less recurrently in the subcategories of *location* and *quantification*. While L2 writers used fewer bundle tokens of *location* than L1 writers, the difference in the number of tokens was not significant. *Quantification* was the only subcategory that L2 writers employed bundle tokens significantly less recurrently than L1 writers did. EC writers used four types of *quantification* bundles (*the extent to which, the total number of, a wide range of*, and *the degree to which*), which were not used by IC writers. Cortes (2004, p. 415) found the absence of *quantification* LBs in the learner academic writing in biology. Chen and Baker (2010) also noticed the absence of extent/degree modifiers such as *the extent to which* and *the degree to which* (quantifying bundles) in learner writing. From this, we can assume that L2 writers including professional academic writers pay less attention to the use of quantifying bundles in their academic writing.

Table 5
Distribution of sub-functional types

Categories	**Subcategories**	**Types**		**Tokens**		**LL**
		EC	**IC**	**EC**	**IC**	
Research-oriented	Location (e.g. *at the end of*)	4	4	339	302	2.39
	Procedure (e.g. *the use of the*)	2	9	128	841	582.13**
	Quantification (e.g. *one of the most*)	4	3	356	269	12.73**
	Description (e.g. *the ways in which*)	2	5	165	333	56.70**
	Total	12	21	988	1745	207.49**
Text-oriented	Transition signals (e.g. *on the other hand*)	2	4	276	474	51.60**
	Resultative signals (e.g. *the results of the*)	2	7	194	695	296.30**
	Structuring signals (e.g. *as shown in figure*)	5	6	437	509	5.02*
	Framing signals (e.g. *in the case of*)	8	9	640	1001	77.71**
	Total	17	26	1547	2679	299.55**

Categories	Subcategories	Types		Tokens		LL
		EC	IC	EC	IC	
Participant-oriented	Stance features (e.g. *it is possible that*)	-	-	-	-	
	Engagement features (e.g. *it should be noted*)	2	4	165	450	135.39**
	Total	2	4	165	450	135.39**
Total		31	51	2700	4874	-

Framing is the subcategory that makes up the largest proportion of bundles in both corpora, which is congruent with the findings of Hyland's (2008a) and Hyland and Jiang's (2018) studies on applied linguistics research articles. This subcategory is used to elaborate arguments by specifying cases (*in the case of*) and pointing to limitations (*with the exception of*) (Hyland, 2008a). For example:

> Moreover, students needed to draw on similar skills and knowledge to those required for their disciplinary assignments, especially *in the case of* students from Applied Linguistics/ TESOL and Education. (EC)
>
> It will also benefit those teaching Computer-Assisted Language Learning (CALL) courses in EFL contexts, particularly *in the context of* higher education in Indonesia. (IC)

However, unlike previous studies that suggest that L2 expert writers employed four-word framing bundles significantly less frequently than their counterparts (e.g. Esfandiari & Barbary, 2017; Pan et al., 2016), this study shows the opposite. This may be due to the higher frequency of PP-based bundles in the IC corpus, since framing clusters consist mainly of PP-based bundles (Pan et al., 2016).

In the participant-oriented category, we did not find any *stance* bundles in either corpora. *Stance* refers to how writers explicitly convey their attitudes, epistemic and affective judgments, and evaluations (e.g. *it is obvious that* and *are likely to be*) (Hyland, 2008a, 2008b). This finding might lend weigh to Hyland and Jiang's (2018) analysis that showed a dramatic decline (-38.2%) in the *stance* bundle tokens of applied linguistics journal articles over a 50-year period from 1965 to 2015. Engagement bundles also experienced a decrease, but not significantly, around -9.2% (Hyland & Jiang, 2018). In the present study, similar to Esfandiari and Barbary's (2017) and Pan et al.'s (2016) studies, we found that L2 English writers used relatively more *engagement* bundle types and tokens than L1 English writers to engage readers and guide them to particular interpretations. For instance:

> *It is important to* note that this study uses the term reader knowledge rather than reader characteristics to specifically refer to linguistic knowledge in terms of grammatical knowledge. (IC).
>
> From the teachers' responses, *it can be seen* that teachers in this study share similar beliefs in grading the students although they come from different schools. . . (IC)

It is important to note that three of the four *engagement* clusters used by IC writers are in the form of anticipatory *it* and passive constructions (*it can be seen, it can be concluded*, *it can be said*), which indicate an impersonal tone. This might be influenced by the writers' preference for impersonality in their academic writing. Like in Hong Kong (Hyland, 2008a) and mainland China (Wei & Lei, 2011), impersonality devices such as passive patterns in academic writing are also suggested in Indonesian universities, accounting for the overuse of passive structure bundles. This might support Li, Franken, and Wu's (2018) study, which argued that one of the reasons for the different bundle selections of L2 learners is classroom learning.

Conclusion

In the present study, we compared the frequency, structure, and function of four-word LBs of L1 and L2 professional writers in research articles published in applied linguistics journals. In terms of frequency, L2 writers employed a higher number of bundles than L1 writers, but L2 writers underused some of the most typical LBs in native writing, such as *the extent to which*. Structurally, unlike previous studies (e.g. Chen & Baker,

2010; Pan et al., 2016), in this study, L2 writers, more specifically Indonesian writers, predominantly used PP-based clusters, which may indicate that L2 professional writers in the field of applied linguistics demonstrated relatively high academic writing proficiency. However, despite the high frequency of PP-based bundles, L2 authors still used a significant number of VP-based bundles, suggesting that these Indonesian professional writers (L2) were still acquiring more native-like LBs, such as NP- and PP-based formulaic sequences. Functionally, L1 English and L2 English professional writers did not differ much in the proportion of the main functional distributions of LBs, but showed marked differences in nearly all subcategories of functional types. L2 writers employed *quantification* bundles significantly less frequently than their counterparts, which also had been reported by the previous studies (Chen & Baker, 2010; Cortes, 2004). These findings show that L2 English journal article authors in the discipline of applied linguistics are more aware of the use of LBs, compared to other disciplines. This lends weight to the argument that academic disciplines influence the use of LBs in professional academic writing.

It is important to point out the two limitations of this research. First, the method of determining L1 and L2 writers may not be a satisfactory approach since it may not be accurate to identify the first language of the author(s) by means of their names and institutions. Second, log-likelihood tests have recently been suggested as problematic for measuring lexical variations by Bestgen (2017). However, the test is still commonly used by many lexical bundle researchers including Hyland and Jiang (2018) so we chose this test together with the calculation of percentages for easier comparisons with other bundle studies. Therefore, future studies can employ different methods of determining L1 and L2 writers, such as by looking at their profile and education background, and different lexical variation measures, such as proportions. Also, examining crosslinguistic influence of L1 Indonesian on the use of L2 English bundles can be conducted.

Despite these limitations, the findings of this study have potential implications for academic writing pedagogy. Teachers are suggested to raise learners' awareness of the most frequently used LBs in a specific academic discipline and to pay attention to the discourse style of academic writing, helping L2 students shift from clausal to phrasal bundle use. Focusing on the most typical bundles in native writing that are underused by L2 writers, such as *quantification* bundles, is also suggested.

Conflict of Interest

The authors declare that they have no conflicts of interest.

References

Ädel, 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*(2), 81-92. http://dx.doi.rg/10.1016/j.esp.2011.08.004

Bestgen, Y. (2017). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns de Filologia: Estudis Lingüístics*, *22*, 33-56. http://dx.doi.rg/10.7203/qf.22.11299

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. http://dx.doi.rg/10.1093/applin/25.3.371 %J Applied Linguistics

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics, 15*(2), 223-250. http://dx.doi.rg/10.1017/S1360674311000025

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35. http://dx.doi.rg/10.5054/tq.2011.244483

Biber, D., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond t-units—A response to WeiWei Yang. *TESOL Quarterly, 47*(1), 192-201. http://dx.doi.rg/10.1002/tesq.84

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

Bychkovska, T., & Lee, 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. http://dx.doi.rg/10.1016/j.jeap.2017.10.008

Chen, Y.-H. (2009). *Lexical bundles across learner writing development* [Unpublished doctoral dissertation]. Lancaster University.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49. http://dx.doi.rg/10125/44213

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880. http://dx.doi.rg/10.1093/applin/amu065 %J Applied Linguistics

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

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora, 3*(1), 43-57. http://dx.doi.rg/10.3366/E1749503208000063

Cortes, V. (2015). Situating lexical bundles in the formulaic language spectrum: origins and functional analysis development. In V. Cortes & E. Csomay (Eds.), *Corpus-based research in applied linguistics: Studies in honor of Doug Biber* (pp. 197-218). John Benjamins.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, 2*(2), 225-246.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*(1), 91-126. http://dx.doi.rg/10.1017/S0272263100014698

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. http://dx.doi.rg/10.1016/j.jeap.2017.09.002

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. http://dx.doi.rg/10.5539/elt.v9n6p176

Gungor, F. & Uysal, H. H. (2020). Lexical bundle use and crosslinguistic influence in academic texts. *LINGUA, 242*, 1-22. http://dx.doi.rg/10.1016/j.lingua.2020.102859

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62. http://dx.doi.rg/10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21. http://dx.doi.rg/10.1016/j.esp.2007.06.001

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

Hyland, K., & Jiang, F. (2018). Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics, 23*(4), 383-407. http://dx.doi.rg/10.1075/ijcl.17080.hyl

Kim, Y. (2009). Korean lexical bundles in conversation and academic texts. *Corpora, 4*(2), 135-165. http://dx.doi.rg/10.3366/E1749503209000288

Kwary, D. A., Ratri, D., & Artha, A. F. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics, 7*(1), 131-140. http://dx.doi.rg/10.17509/ijal.v7i1.6866

Li, L., Franken, M., & Wu, S. (2019). Chinese postgraduates' explanation of the Sources of sentence initial bundles in their thesis writing. *RELC Journal*, *50*(1), 37-52. http://dx.doi.rg/10.1177/0033688217750641

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198. http://dx.doi.rg/10.1016/j.esp.2009.04.003

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. http://dx.doi.rg/10.1016/j.jeap.2015.11.003

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

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. http://dx.doi.rg/0.1016/j.jeap.2014.01.002

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

Ruan, Z. (2017). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC Journal, 48*(3), 327-340. http://dx.doi.rg/10.1177/0033688216631218

Salazar, D. (2014). Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching (Vol. 65). Amsterdam: John Benjamins Publishing Company. http://dx.doi.rg/10.1075/scl.65

Shin, Y. K., Cortes, V., & Yoo, I. W. (2018). Using lexical bundles as a tool to analyze definite article use in

L2 academic writing: An exploratory study. *Journal of Second Language Writing*, *39*, 29-41. http://dx.doi.rg/10.1016/j.jslw.2017.09.004

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. http://dx.doi.rg/10.1016/j.jeap.2013.05.002

Wang, Y. (2017). Lexical bundles in spoken academic ELF. *International Journal of Corpus Linguistics, 22*(2), 187-211. http://dx.doi.rg/10.1075/ijcl.22.2.02wan

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

Wood, A. (2001). International scientific English: The language of research scientists around the world. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 71-83). Cambridge University Press.

Wood, D. C., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1-13. http://dx.doi.rg/10.1016/j.jeap.2014.03.002

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

Wray, A. (2006). Formulaic language. In E. K. Brown & A. Anderson (Eds.), *Encyclopedia of language and linguistics* (pp. 590-597). Elsevier.

Wright, H. R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes, 54*, 1-14. http://dx.doi.rg/10.1016/j.esp.2018.09.001

Appendix 1

Journal list

No	English Journals	Indonesian Journals
1	Annual Review of Applied Linguistics	Indonesian Journal of Applied Linguistics
2	Applied Linguistics	TEFLIN journal
3	Journal of Second Language Writing	PAROLE: Journal of Linguistics and Education
4	Studies in Second Language Acquisition	Celt: A Journal of Culture, English Language Teaching & Literature
5		IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)
6		Lingual: Journal of Language and Culture
7		TELL-US Journal

Appendix 2

Lexical bundle lists (shared bundles are bolded)

No	Bundles in the English corpus		Bundles in the Indonesian Corpus	
	Bundles	Frequency	Bundles	Frequency
1	the extent to which	178	**in the form of**	300
2	in the present study	153	**on the other hand**	245
3	**on the other hand**	144	the result of the	186
4	**as well as the**	132	it can be seen + (that)	162
5	in the case of	123	(this) + is in line with + (the)	147
6	**at the same time**	118	to be able to	145
7	**the results of the**	115	it was found that + (the)	143
8	**in the context of**	107	(it) + can be concluded that (the)	139
9	in the current study	100	is one of the	134
10	the ways in which	99	**the results of the**	133
11	**in terms of the**	97	can be seen in + (the)	121
12	**it is important to**	95	**the use of the**	117
13	**the end of the**	92	**in the context of**	112
14	**on the basis of**	82	that the use of	107
15	**as a result of**	79	can be seen from (the)	105
16	**at the end of**	74	in the process of	105
17	as can be seen	70	**as well as the**	99
18	can be seen in	68	in this study the	99
19	**the use of the**	67	on the use of	91
20	the nature of the	66	**in terms of the**	90
21	the total number of	66	in this case the	85
22	participants were asked to	61	**at the same time**	84
23	with regard to the	61	to find out the	83
24	of the present study	59	**the end of the**	82
25	a wide range of	58	as one of the	79
26	in the field of	58	it can be said + (that)	79
27	**in the use of**	58	the meaning of the	79
28	the fact that the	57	**at the end of**	77
29	at the beginning of	55	in other words the	76
30	**in the form of**	54	can be used to	75
31	the degree to which	54	in relation to the	73
32			in accordance with the	72
33			that there is a	72
34			it means that the	71
35			**it is important to**	70
36			the findings of the	65
37			**in the use of**	64
38			of this study is	62
39			can be found in	61
40			in order to make	61
41			in this research the	61

No	Bundles in the English corpus		Bundles in the Indonesian Corpus	
	Bundles	**Frequency**	**Bundles**	**Frequency**
42			it is found that	60
43			above it can be	59
44			**on the basis of**	58
45			that there is no	57
46			the implementation of the	57
47			one of the most	56
48			**as a result of**	55
49			an important role in	54
50			in addition to the	54
51			the findings of this	53