DOI: 10.1111/weng.12605

ORIGINAL ARTICLE

WORLD ENGLISHES WILEY

Bundles in advanced EAL authors' articles: How do they compare with world Englishes practices?

Ken Hyland[1] | Feng (Kevin) Jiang[2]

[1]School of Education and Lifelong Learning, University of East Anglia, Norwich, UK

[2]Jilin University, Changchun, China

Correspondence
Ken Hyland, University of East Anglia, Norwich, UK.
Email: K.Hyland@uea.ac.uk

Abstract

With increasing numbers of scholars from around the world now engaged in international publishing to further their careers, many authors for whom English is not their first language worry about the acceptability of their language to journal editors. In this article we explore this issue by focusing on a key component of fluent academic writing: the high frequency fixed-word collocations known as lexical bundles, strings which are 'glued together' and help characterize smooth production. Here we compare their use in the pre-submission drafts of authors with different first languages with published papers in leading international journals. Our results suggest that language background certainly contributes to the differences between our EAL texts and published papers, but that seniority and discipline also significantly impact these language choices.

1 | INTRODUCTION

A central concern of English for specific purposes (ESP) research and teaching in recent years has been with the needs of academics writing for publication. Participation in the global exchange of information has become a prerequisite for a successful career for academics around the world, and this increasingly has to be done in 'international' English-medium journals. As a result, many academics for whom English is not their first language worry about the 'correctness' of their language, turning to 'text mediators' of various kinds to polish their prose (Na & Hyland, 2016) or even translate it wholesale into English (Na & Hyland, 2019). With English now firmly entrenched, for the time being, as the academic lingua franca, the question arises of how far native varieties differ from the conventions accepted by editors of published papers.

In this study we explore this issue by examining high frequency fixed-word strings known as lexical bundles, which are often seen as a defining characteristic of fluent academic production (Cortes, 2006). We do this by comparing their use in both published papers and the pre-submission drafts of L2 users of English. In particular, we set out to answer the following questions:

1. To what extent do lexical bundles differ between the manuscripts of writers with different L1s and those published in leading international journals?
2. How far does L1 background, discipline, and academic experience influence the use of bundles?

We hope that any similarities and differences we find in these uses will tell us something about non-native varieties of English in a global communicative context. In the following sections we briefly discuss the significant presence of English as an Additional Language (EAL) speakers in global publishing and the importance of fixed phrases in academic linguistic production. We then go on to present and discuss our study.

2 | ENGLISH, GLOBAL PUBLISHING, AND THE NON-NATIVE ENGLISH AUTHOR

There is little dispute about the current status of English as the lingua franca of the academic world, although there is less agreement about what this means. For some, the globalization of knowledge production has opened borders to relationships between individuals and to the exchange of ideas, creating an imagined research community which is the very embodiment of Enlightenment science. For others, it disadvantages those who do not speak English as a first language while undermining other academic languages. What is clear, however, is that this is a variety of English which differs dramatically from all others and requires a specialist expertise to be pulled off successfully.

The growth of English in academic communication has been accompanied by a huge increase in scholarly publishing. There are perhaps 8 million academics now working in 17,000 universities around the world seeking to publish in English-language journals each year (UNESCO, 2017). One of the largest journal publishers, Elsevier, for example, reported over 2 million articles submitted and 1 billion read in 2019 (Page, 2020). This massive expansion of academic publishing is driven not only by the ease of collaboration and access afforded by new technologies, but also by the fact that career opportunities of academics across the globe are increasingly tied to an ability to gain acceptance for work in high-profile journals. The appraisal culture is spreading like a pandemic, requiring greater participation in publishing, and this growth has been led by EAL authors, especially those from emerging economies (Tollefson, 2018). The Scimago country ranking for 2019,[1] for example, puts China at the head of the list with India, Japan, and Russia also in the top 10, along with EU countries. Most publishing authors, then, are now writing in an additional language.

Participation in this global activity clearly puts a considerable strain on the communicative resources of all writers, whatever their language background (Hyland, 2016) and even Native English Speakers (NESs) struggle to create effectively persuasive texts (Belcher, 2007). Most attention, however, has been devoted to the *linguistic* difficulties faced by non-Anglophone authors (Clavero, 2010; Guardiano et al., 2007). EAL authors themselves often express a sense of frustration having to write in English and a feeling that NES scholars have it easier (Ferguson et al., 2011; Hanauer & Englander, 2011). But while the importance of a certain proficiency in a foreign language should not be underestimated, often it is the *rhetorical* requirements of academic disciplines which confound authors. The expectations which surround the appropriate presentation of arguments and the use of community-familiar expressions are frequently the biggest obstacles they face.

At the same time, however, these conventions themselves are under pressure to change from native varieties of English. With international research dependent on the use of one shared language, cross-cultural and cross-linguistic collaborations are likely to impact the structures of this register. Ever more interactions between native and non-native speakers of English, where expectations are culturally and situationally embedded, create a more complex picture of 'acceptable' academic writing. In this paper we consider the influence of first language, discipline, and academic experience on writing for publication by exploring a key feature of assured and appropriate writing: lexical bundles. Using a corpus of 150 manuscripts written by EAL scholars with 10 different language backgrounds and various levels of publishing experience, we compare their unmediated texts with those successfully published in established international journals of the same fields.

3 | LEXICAL BUNDLES IN ACADEMIC WRITING

An important component of fluent linguistic production is control of lexical bundles, or multi-word expressions. These are strings of three or more words which are 'glued together' in everyday discourse such as *it was found that* and *in the case of* in academic registers. Simply, bundles are statistically the most frequent recurring sequences of words in any collection of texts: extended collocations which appear more repeatedly than by chance. They are made visible and retrieved by corpus analysis software with specified frequency and distribution criteria (Biber, 2006). As a result, bundles are generally neither idiomatic nor complete grammatical units, but they are familiar to experienced users of a language and have customary pragmatic or discoursal functions (Biber, 2009; Hyland, 2008a). Academic writing, for instance, draws on a much larger stock of prefabricated phrases than either news or fiction (Hyland, 2012) while Biber et al. (1999, p. 994) suggest that four-word bundles occur over 5,000 times per million words in academic prose.

Lexical bundles seem to reflect a very real part of users' communicative experiences. As suggested by Sinclair's (1991) 'idiom principle,' there is a phraseological tendency in language use whereby speakers and writers co-select words in routine ways. We are mentally 'primed' to expect co-occurring words through our experience of them in frequent associations (Hoey, 2005). So by making language more predictable to the reader they act as processing short-cuts and work to facilitate pragmatically efficient communication. In academic discourse they help to reduce processing time by using familiar patterns to guide readers through a text (*in the next section*, *we can see that*) or by linking ideas (*is due to the*, *in contrast to*). In addition, by signaling appropriate use of a disciplinary code, they allow writers to display solidarity with colleagues (Cortes, 2006) and to construct a disciplinary competent voice (Hyland, 2008a; Pang, 2010).

In other words, bundles reveal the lexico-grammatical, community-authorized ways of making meanings and at the same time help define expertise and disciplinary membership. They are familiar to writers and readers who regularly participate in a particular discourse, their very 'naturalness' signaling competent input in a given community. Conversely, this means that their absence might indicate the writing of a novice or newcomer. Haswell (1991), for example, suggests that:

> there can be little doubt that as writers mature they rely more and more on collocations and that the lesser use of them accounts for some characteristic behaviour of apprentice writers. (Haswell, 1991, p. 236)

Research, in fact, indicates that the bundles used by novices and students differ markedly from those in published academic writing (Chen & Baker, 2010; Cortes, 2004; Hyland, 2008b; Scott & Tribble, 2006). Studies have found, for example, that Chinese writers often overuse strings such as *first of all*, *on the other hand*, and *in a nutshell*, compared with NES writers (Lee & Chen, 2009; Ma, 2009).

The study of high-frequency bundles and their possible variations can therefore tell us something about the influence of contexts on written academic text production, indicating how scholarly authors both use and perceive acceptability. In the following section we describe our corpus and method and then go on to discuss our findings.

TABLE 1 The ten L1 categories in the SciELF corpus

	First author's L1	No. of articles	No. of words	% of words	Avg. words per article
1	Finnish	25	123,153	16.22	4926.12
2	Czech	22	109,173	14.38	4962.41
3	French	16	91,186	12.01	5699.13
4	Chinese	21	84,807	11.17	4038.43
5	Spanish	13	79,038	10.41	6079.85
6	Russian	13	71,376	9.40	5490.46
7	Swedish	13	60,060	7.91	4620.00
8	Italian	11	58,685	7.73	5335.00
9	Portuguese	12	56,625	7.46	4718.75
10	Romanian	4	25,197	3.32	6299.25

TABLE 2 Professional status of authors in the SciELF corpus

	Experience of first author	No. of articles	No. of words	% of words	Avg. words per article
1	Research students	30	165775	21.83	5525.83
2	Junior academics	86	424161	55.86	4932.10
3	Senior academics	34	169364	22.31	4981.29

TABLE 3 Disciplinary characteristics in the SciELF corpus

	Discipline of first author	No. of articles	No. of words	% of words	Avg. words per article
1	Humanities	23	144,857	19.08	6298.13
2	Social sciences	49	280,876	36.99	5732.16
3	Life sciences	19	72,881	9.60	3835.84
10	Natural sciences	59	260,686	34.33	4418.41

4 | CORPUS AND METHOD

4.1 | SciELF and academic BNC corpora

Two data sets form the basis of our analysis: the SciELF and the academic section of the British National Corpus (BNC). We focus principally on the SciELF (2015) corpus which comprises research papers written by L2 users of English. These are final drafts of unpublished manuscripts which have not been edited by professional proofreading services or a native speaker of English. It is thus a corpus of second-language use in written academic communication. There are 150 papers (759,300 words) in the corpus written by authors with 10 different L1 backgrounds in a range of disciplines. The language backgrounds of the authors are shown in Table 1:

To explore the contexts of these research papers in greater detail, we also grouped the texts into the professional status of the authors (Table 2) and broad disciplinary groups (Table 3).

To determine if these authors used lexical bundles in a distinctive way, we used the academic section of the BNC as a reference corpus. The BNC is a huge corpus designed to characterize contemporary British English in its

TABLE 4 Disciplinary breakdown in the academic sub-corpus of BNC

	Disciplines	No. of articles	No. of words	% of words	Avg. words per article
1	Humanities	23	153,452	19.70	6671.83
2	Social sciences	49	288,639	37.05	5890.59
3	Life sciences	19	75,646	9.71	3981.37
10	Natural sciences	59	261,398	33.55	4430.47

various uses, and we extracted written academic texts from this to construct a parallel corpus with similar disciplinary characteristics and the same number of texts in each division (Table 4).

4.2 | Identification of lexical bundles

While lexical bundles are automatically identified on the basis of frequency of occurrence and breadth of use, researchers have used different criteria to determine what counts as a bundle. The threshold frequency has ranged from 10 (Biber, 2006) through 20 (Cortes, 2004; Hyland, 2008a, 2008b) to 40 times per million words (Biber et al., 2004), and even raw frequencies (Chen & Baker, 2010). A second identification criterion is that sequences have to occur in a specified number of files in the corpus, such as three to five texts (Biber & Barbieri, 2007) or 10 per cent of texts (Hyland, 2008a) to avoid idiosyncratic uses. Analysts must also decide on the length of strings they select. Two-word bundles are extremely common and are therefore less useful for research purposes (Staples et al., 2013) while 5- and 6-grams are comparatively rare and often subsume shorter ones. Four-word bundles seem to be most often studied, perhaps because they are over 10 times more frequent than five-word sequences and offer a wider variety of structures and functions to analyze (Biber et al., 1999).

For the present study, we took a conservative approach by following Hyland (2008a, 2008b) and Cortes (2004) in setting a high-frequency cut-off of 20 occurrences per million words and including only those bundles which appeared in at least 10 per cent of texts. We also decided to focus on four-word bundles due to their frequency and their variety. We manually excluded bundles with text-dependent noun phrases (for example, *the second world war*) and removed overlapping word sequences where two four-word bundles are actually part of a five-word string (such as *play an important role* and *an important role in*) (Chen & Baker, 2010).

4.3 | Presentation of lexical bundles

We conducted the search for four-word bundles using *AntGram* (Anthony, 2020), a freeware n-gram generation tool, following our specific criteria. The results were transferred into an Excel file where we coded each example for its function and grammatical structure. The two authors worked independently to code a 10 per cent sample, refining agreement through successive passes to achieve an inter-rater reliability of 98 per cent (structure) and 97 per cent (functions).

When comparing lexical bundles in corpora of different sizes, we are aware that smaller corpora may show more bundles than larger corpora after normalization as phrases which repeat just a few times could meet the lower cut-off point (Cortes, 2015, p. 205). Having divided the SciELF sub-corpus into smaller groups, normalization could skew our results. Because of this, we studied the *proportion* of bundle types in the sub corpora rather than the frequencies, but followed Hyland (2008a, 2008b) and Hyland and Jiang (2018) in using *log Likelihood* tests to determine statistically significant differences. We also considered effect size (*%DIFF*) for the veracity of *log Likelihood* results (Gabrielatos, 2018).

TABLE 5 Frequency of four-word lexical bundles in the two corpora

	SciELF	Academic BNC
Types (unique bundles)	134	458
Tokens (frequency of types)	3316	9221
type/token	0.04	0.05
Tokens per paper	22.11	61.47
% of corpus	0.44	1.18

TABLE 6 Distribution of four-word bundles by L1 backgrounds in the SciELF

	Chinese	Czech	Finnish	French	Italian	Portuguese	Romanian	Russian	Spanish	Swedish
Type	83	113	116	98	32	38	9	49	65	46
Token	378	573	628	512	174	212	48	250	337	218
Type/token	0.22	0.20	0.18	0.19	0.18	0.18	0.19	0.20	0.19	0.21
% of corpus	0.43	0.51	0.49	0.54	0.28	0.36	0.19	0.34	0.41	0.35

5 | FREQUENCIES OF BUNDLE USE BY EAL WRITERS

In the SciELF corpus of NES articles we identified a total of 134 different four-word bundles (types) occurring a total of 3316 times (tokens of these types). Each paper contained an average of 22.11 bundles. In contrast, the BNC corpus of published papers contained both more types (458) and more cases of these types (9221), averaging 61.47 cases per paper. Therefore, the L2 authors used significantly fewer four-word bundles than the published authors (*log Likelihood = 2744.32, %DIFF = –63.10, p<0.0001)*. This considerable difference is illustrated in Table 5.

It is clear that many clusters used by the writers of published papers are not found in the EAL papers while others appear far less frequently. This reliance on a narrower range of formulaic expressions at much lower frequencies may be due to a lack of familiarity with the common ways published writers create cohesive texts. While there is nothing to suggest the EAL manuscripts are less effective or somehow inadequate, it may seem to readers that the arguments they find in them are less fluent and assured than they might expect and the writers are not shaping meanings in anticipated ways. We also see from Table 5 that the bundles comprise a higher proportion of the published papers and that these papers also have a higher type to token ratio, indicating greater variation or 'richness' in terms of the number of unique bundles in the corpus. In other words, the EAL writers, in this corpus anyway, use fewer different bundles and use them less often, demonstrating a more restricted and perhaps more repetitive way of communicating.

We also find that writers with different first languages tend to use bundles differently from each other. Table 6 shows a wide variation in the use of number and frequencies across the languages. While the larger corpora are likely to produce more bundles, and the figures for the Romanian L1 writers obviously reflect the low number of texts in that sub-corpus, the type/token ratios fall into a very narrow range across all groups, indicating more accurately the true variety of bundles across the corpora.

What seems apparent from these results is that the very different contexts in which these writers may be working, or their experiences of learning English in the past, have relatively little impact on the extent of their use of academic four-word bundles. Writers from geographically more remote countries such as Brazil (Portuguese speakers) and those raised using a character-based script (China), seem no more disadvantaged in the frequency of their use of bundles than those working in leading publishing countries (France and Spain) or with high levels of spoken English in the community (Sweden).

6 | PREFERENCE FOR BUNDLE TYPES BY EAL WRITERS

When turning to the most frequent bundles, however, we find a different story. While writers favor *on the other hand* and *in the case of* as the top two in both the SciELF and BNC lists, there is a wide variation by language background. Table 7 shows the top 10 most commonly used bundles in the articles by the different language groups, with those overlapping with the published texts shaded.

Only eight of the top 20 bundles in the BNC corpus appear among those most used 10 by EAL writers, ranging from six by the Italian writers to just two in the Romanian, Finnish, and Swedish lists, and none in the French list. High frequency items in the published texts, such as *in terms of the*, *the way in which*, and *the extent to which* do not feature among those most favored by writers of different languages at all. We can see that writers with different first languages have predilections for particular forms, so we find *than when compared to* at the top of the Chinese writers' list but not in any of the other, similarly with *a better understanding of* at the top of the French list. In fact, it appears that the individual language lists comprise almost completely different sets of items, with only a few bundles overlapping in three or more lists and over 50 appearing as unique items. The list of items most frequently used by the French writers, for example, contains only one bundle which occurs in another list.

Once again, we should stress that these differences do not imply poor academic writing or a lack of proficiency in English. Far from it, the bundles here seem perfectly acceptable for their purposes and are likely to be read as such by those familiar with published academic texts. They may, however, characterize the academic writing of particular language groups and suggest familiarity with ways of patterning arguments which are less widely used in published articles.

Another important contextual aspect underlying writers' language choices is the experience of writing and publishing they bring to the act. We can see from Table 8 that junior academics employed far more bundles and used them more frequently than the other groups, although the figure is skewed by the fact that the corpus contains over 50 more texts and 60 per cent more words.

We can see, however, that although the type/token ratio was similar to the others, four-word bundles comprised a much larger proportion of their texts. We are unsure how to account for this, although it may be that the junior scholars devoted more time to their writing and expression of arguments, deliberately seeking to produce bundle-rich texts which displayed their awareness of a range of conventions. We find, however, that each group had different preferences for the forms of bundles they used. It is true that the top three bundles in the published corpus, *on the other hand*, *in the case of*, and *at the same time* occur in the 10 most frequent bundles in each list, with the first two of those most used by junior and senior scholars. However, all the other items differ from the other groups and from the BNC monitor corpus, indicating once again a distinctive take on fluent academic production.

A final important contextual influence on the decisions writers make is the discipline they are working in and here, once again, there are considerable differences in the variety, frequency, and types of bundles in the two corpora. Table 9 illustrates these comparisons. We can see that the published authors use significantly more types, and use them far more often, in the social and life sciences, that the differences are less dramatic in the humanities, but that the EAL authors working in the natural sciences employ well over twice the number of bundles as the published authors.

Presumably, the natural science writers are familiar with a larger range of formulaic ways of presenting what are often numerical and quantitative results, but they are using bundles not found in the published corpus to do so. Thus, we find that *one of the most*, *at the end of*, and *greater than or equal* are heavily used in the natural science texts but are not among the top 20 in the published corpus or in any of the other fields. While the top 10 bundles in the humanities and social sciences show considerable overlap with those in the BNC corpus, none of those used by EAL writers in the top 10 of the life sciences correspond with those in the published texts. Clearly, there are considerable variations not only in the bundles favored by writers in different disciplines (Hyland, 2008a; Hyland & Jiang, 2018), but among those writing English with other language backgrounds.

TABLE 7 Ten most frequent four-word bundles by different L1 writers (overlap with BNC shaded)

Chinese	Czech	Finnish	French	Italian
than when compared to	on the other hand	on the other hand	a better understanding of	the end of the
at the same time	is one of the	than or equal to	a function of the	at the end of
on the other hand	on the basis of	in the present study	a large number of	than or equal to
at the end of	as well as the	of the present study	a part of the	in the case of
is one of the	one of the most	as well as the	and the other one	on the basis of
the end of the	in the case of	the beginning of the	are reported in table	on the other hand
an important role in	at the end of	the state and the	as a function of	with respect to the
by means of a	in the area of	greater than or equal	as the ratio between	it is possible to
in accordance with the	in the field of	in the case of	as well as in	one of the most
the size of the	as a result of	a part of the	as well as the	the beginning of the
Portuguese	**Romanian**	**Russian**	**Spanish**	**Swedish**
on the other hand	in accordance with the	in the case of	on the other hand	as well as the
it is possible to	on the other hand	at the same time	in the case of	on the other hand
it is important to	taking into account the	in the context of	one of the most	at the end of
the analysis of the	the fact that the	on the basis of	that is to say	are shown in fig
the use of the	in the case of	and at the same	at the same time	the end of the
at the same time	of the most important	as a result of	in relation to the	can be seen in
for the purposes of	is one of the	on the other hand	is one of the	to be able to
in relation to the	one of the most	a large number of	as a result of	be referred to as
the fact that the	the structure of the	it is important to	the analysis of the	in the present study
the other hand the		one of the most	the point of view	in this case study

TABLE 8 Distribution of four-word bundles by experience in the SciELF

	Research students	Junior academics	Senior academics
Type	215	1053	265
Token	1107	5681	1365
Type/token	0.19	0.19	0.19
% of corpus	0.67	1.34	0.81

TABLE 9 Disciplinary distribution of four-word bundles in SciELF and BNC academic

	Humanities		Social science		Life science		Natural science	
	ELF	BNC	ELF	BNC	ELF	BNC	ELF	BNC
Type	181	224	279	471	100	380	549	213
Token	1018	1115	2188	5817	473	2187	2959	1498
Type/token	0.01	0.20	0.01	0.08	0.01	0.17	0.01	0.14
% of corpus	0.66	0.73	0.75	2.02	0.65	2.89	1.10	0.57

TABLE 10 Classification of four-word lexical bundles in academic writing

Verb phrase-related	• passive verb (*is shown in fig*, *can be noted that*) • copular be (*is one of the*, *is the number of*) • imperative (*should note that the*, *let us observe that*)
Clause-related	• anticipatory it (*it is important to*, *it follows that the*) • abstract subject (*the goal is to*, *fig b shows the*) • human subject (*we shall have to*, *one should note that*) • *as*-fragments (*as can be seen*, *as shown in fig*) • *if*-fragments (*if and only if*, *if we look at*) • *there*-fragments (*there seems to be*, *there has been a*) • *wh*-fragments (*which is to be*, *which is equivalent to*) • *that*-fragments (*that the effect of*, *that need to be*)
Noun/preposition-related	• noun phrase with *of*- fragment (*the nature of the*, *the case of the*) • noun phrase with other post-modifier fragment (*the fact that the*, *the extent to which*) • prepositional phrases (*in terms of the*, *with respect to the*) • comparative expressions (*as well as the*, *as far as the*)

7 | STRUCTURAL DIFFERENCES IN BUNDLES BY EAL WRITERS

Among the distinguishing features of academic discourse is the formal properties of its lexical bundles (Biber et al., 1999). In academic writing, bundles are frequently prepositional phrases with -*of* fragments (*as a result of*), noun phrase + *of* fragments (*the nature of the*) (Scott & Tribble, 2006, p. 138; Hyland, 2008b) and anticipatory *it* fragments (*it is argued that*) (Salazar, 2014; Hyland & Jiang, 2018). Together, these three forms comprise over 70 per cent of all four-word patterns in academic discourse but rarely figure in conversation, where the majority of bundles contain a verb phrase, particularly 'personal pronoun + verb phrase' (for example, *I don't know what*). In this study we followed Hyland and Jiang's (2018) categorization of four-word bundles (Table 10) and coded the sequences we found in the corpora.

We can see from Figure 1 that while the *relative* use of structural bundles are similar in the two corpora, the published papers are far more heavily dominated by bundles containing a noun or preposition (77% of the total compared with 65% by the EAL writers). The noun phrase with *of*-phrase fragment is the most common structure in both cor-

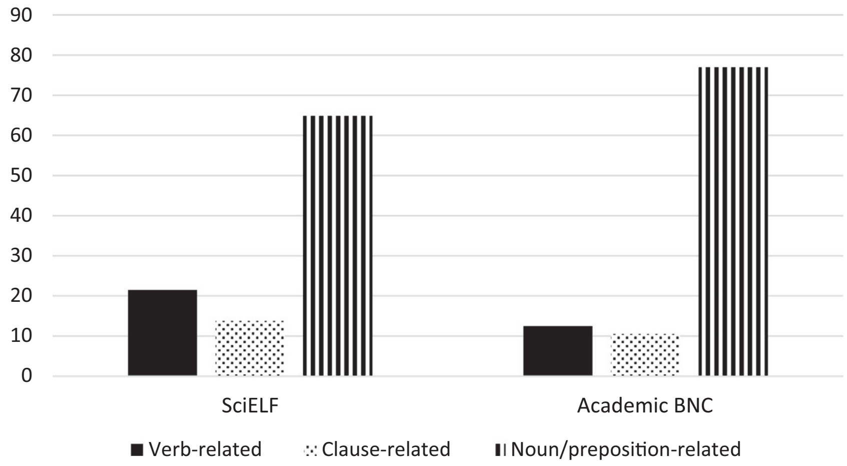

FIGURE 1 Structural distribution of four-word bundles in the two corpora (%)

TABLE 11 Structural distribution of four-word bundles by L1 (per 10,000 words)

	Chinese	Czech	Finnish	French	Italian	Portuguese	Romanian	Russian	Spanish	Swedish	BNC corpus
Verb phrasal	7.76	7.54	7.83	11.65	2.76	5.14	3.51	5.85	5.01	11.81	14.82
Clausal	2.51	2.84	5.63	6.72	1.79	5.48	0.00	5.17	6.60	2.07	12.43
Noun/prepositional	32.85	40.44	35.69	35.38	23.24	25.70	15.23	23.00	29.60	20.26	91.14

pora, comprising over 25 per cent of all forms. This covers a range of meanings in academic discourse and in particular is widely used to identify quantity, place, or size (*the temperature of the*, *the base of the*), to mark existence (*a wide range of*, *the presence of the*), or highlight qualities (*the nature of the*, *a function of the*). Passive verb bundles are the second most popular structural pattern, normally followed by a prepositional phrase fragment typically marking a location or a logical relation as writers seek to either guide readers through the text (*is shown in Figure*, *are summarized in Table*) or identify the basis for an assertion (*be related to the*, *is based on the*).

Table 11 shows how the different language groups used these bundle structures, with the Chinese, Czech, and Finnish writers turning to these noun and preposition-related forms most frequently. The Swedish and French speakers made particularly heavy use of verb phrase bundles as in (1) and (2), and these comprised over a third of all types in the texts of the former. The Italian texts had exceptionally low frequencies of both clausal and verb phrase-related bundles.

1. The lower and upper frequency limits **are shown in Fig**. 6 with dashed and solid lines respectively. (Swedish)
2. this study could be completed **taking into account the** values emitted by existing plants… (French)

Table 12 shows how the different professional and writing experience of these EAL writers influenced their use of structural bundles. What is most striking about this table is not only the extent to which they differ from the published texts, but also how similar they are to each other. Irrespective of status and seniority, the proportions of structural bundles were remarkably consistent across the texts of all EAL authors. Verb phrase-related structures are very high across all groups compared to the published papers, with passive verb structures predominating. This may suggest that preferences for broad structural patterns of bundles may have more to do with writing in an additional language rather than seniority of writing experience.

TABLE 12 Structural distribution of four-word bundles by status (per 10,000 words)

	Research students	Junior academics	Senior academics	BNC Academic
Verb phrase-related	15.09 23.85%	25.24 19.27%	16.30 21.32%	14.81 12.52%
Clausal	8.40 13.28%	19.55 14.92%	9.91 12.97%	12.40 10.50%
Noun/prepositional	39.78 62.87%	86.23 65.82%	50.24 65.71%	91.12 76.99%
Totals	63.27 100%	131.02 100%	76.45 100%	118.33 100%

The influence of discipline as a context of writing plays a more significant role in these EAL writers' language choices. Table 13 shows that while the overall frequencies of bundles in each structural category are considerably lower in the EAL texts, the proportion of the three main types broadly corresponds with those in the published papers.

As noted earlier, noun phrases with *of* and prepositional phrases are the most frequent patterns, but while these are particularly heavily used in the life and natural sciences papers in the published texts, the EAL authors tend towards passive bundles. There is also a greater preference for abstract entities among the ELF scientists (3) and (4), with uses in the natural sciences significantly exceeding those in the BNC corpus.

1. The **data were analysed using** SPSS for Windows, version 20.0 (41). (Life sciences)
2. our **results show that the** definition of social status can make a great difference for the model predictions. (Natural sciences)

In the humanities and social sciences there is a much stronger preference for anticipatory *it* bundles than in the published papers (5) and (6), and we also find more clausal fragments in the humanities, particularly *if*- and *that*- fragments (7) and (8).

1. **It is important to** stress how the conjunction but causes certain counter-expectancy onto the reader, which helps the character to construe a specific picture: an ugly woman. (Humanities)
2. For measurement **it is necessary to** understand and describe the whole innovation process and to identify factors that may affect the ultimate realisation of innovation. (Social sciences)
3. **If we take the** discourse events of each passage, we will see that their descriptions help clarify the subtle meanings of evaluation… (Humanities)
4. the disagreement can't be resolved by pointing to the fact **that one of the** two subjects is less biased. (Humanities)

These overall differences suggest that the EAL writers, in general, appear to take a more cautious approach to authorial visibility, preferring to disguise their role in the research than to explicitly take credit for it. Speculatively, this may be a result of their training in academic writing and textbook invocations to adhere to norms of objectivity to gain greater credibility for one's arguments.

8 | FUNCTIONAL DIFFERENCES IN LEXICAL BUNDLES BY EAL WRITERS

Finally, we considered the rhetorical functions performed by the bundles writers used. Here we follow Biber, Conrad, and Cortes (2004) and Hyland (2008a, 2008b, 2012) in grouping bundles into three main functional groups: research-oriented, dealing with referential functions in the real world; text-oriented, concerned with the organization of the discourse; and participant-oriented, concerned with stance and evaluation. Each category is further subdivided into the main focus of the bundle as shown in Table 14.

TABLE 13 Structural distribution of four-word bundles by field (per 10,000 words & %)

	SciELF				Academic BNC			
	Humanities no %	**Social science no %**	**Life science no %**	**Natural science no %**	**Humanities no %**	**Social science no %**	**Life science no %**	**Natural science no %**
Verb phrase	8.12 12.32	11.50 15.24	20.33 31.33	27.91 25.47	8.22 11.31	28.82 14.30	80.89 27.98	13.20 22.98
Clausal	7.48 11.35	10.15 13.45	9.18 14.14	13.02 11.88	7.43 10.22	23.24 11.53	38.22 13.22	6.03 10.50
Noun/prep	50.33 76.34	53.80 71.31	35.39 54.53	68.67 62.66	57.05 78.47	149.46 74.17	170.01 58.80	38.20 66.52
Totals	65.93 100	75.45 100	64.90 100	109.60 100	72.70 100	201.52 100	289.13 100	57.43 100

TABLE 14 Functional categories of four-word bundles (Hyland, 2008a)

Research-oriented

- location — indicating time and place (*at the same time*, *in the present study*);
- procedure (*the use of the*, *the role of the*, *the purpose of the*, *the operation of the*);
- quantification (*the magnitude of the*, *a wide range of*, *one of the most*);
- description (*the structure of the*, *the size of the*).

Text-oriented

- transition signals – establish additive or contrastive links (*on the other hand*, *in addition to the*, *in contrast to the*);
- resultative signals – mark inferential or causative relations (*as a result of*, *it was found that*, *these results suggest that*);
- structuring signals – organize stretches of discourse or direct reader elsewhere in text (*in the present study*, *in the next section*, *as shown in fig.*);
- framing signals – situate arguments by specifying limiting conditions (*in the case of*, *with respect to the*, *on the basis of*, *in the presence of*, *with the exception of*).

Participant-oriented

- stance features – convey the writer's attitudes and evaluations (*are likely to be*, *may be due to*);
- engagement features – address readers directly (*it should be noted*, *as can be seen*).

TABLE 15 Functional distribution of four-word bundles (per 10,000 words)

	SciELF	BNC
Research-oriented	**18.64 (42.68%)**	**47.72 (40.32%)**
Location	5.51	12.61
Procedure	5.11	14.18
Quantification	3.92	9.31
Description	4.10	11.62
Text-oriented	**18.43 (42.20%)**	**51.71 (43.69%)**
Transition	5.93	16.42
Resultative	2.71	7.00
Structuring	2.51	7.61
Framing	7.28	20.68
Participant-oriented	**6.42 (14.70%)**	**18.92 (15.99%)**
Stance	5.91	16.20
Engagement	0.51	2.72

The functional classifications show a strong connection to the structural patterns discussed earlier, with *noun phrases + of* structures prominent in research-oriented functions, *prepositional phrase* patterns in text-oriented functions, and *anticipatory it* largely occurring in participant functions. We can also see, in Table 15, a roughly even split between research and text-oriented bundles overall, with participant strings being far less frequent. While this pattern is mirrored across the two corpora, we can see a slight preference for research bundles among the EAL writers, particularly in those concerned with location and quantification, and an almost negligible use of engagement bundles. Proportions of text-oriented bundles in both corpora are similar, largely used to frame arguments and establishing links between ideas.

When we turn to the uses by the L1 writer groups, in Table 16, we see considerable differences in functional preferences. The Chinese and French authors, for example, make heavy use of research-oriented bundles, imparting a greater real-world, laboratory-focused sense to their writing (9 and 10) while devoting minimal attention to explicit expressions of stance. The Swedish writers, on the other hand, use more bundles to ensure their texts are set out

TABLE 16 Distribution of four-word bundle functions by L1 (per 10,000 words & %)

	Chinese		Czech		Finnish		French		Italian	
Research-oriented	31.03	71.96	23.15	45.55	24.18	49.20	34.86	64.84	13.82	48.85
Text-oriented	10.72	24.87	21.11	41.54	21.05	42.83	17.32	32.23	11.54	40.81
Participant-oriented	1.37	3.17	6.56	12.91	3.91	7.96	1.57	2.93	2.93	10.35
	Portuguese		**Romanian**		**Russian**		**Spanish**		**Swedish**	
Research-oriented	13.02	35.85	5.86	31.24	12.65	37.20	17.12	41.54	11.49	33.02
Text-oriented	14.39	39.62	8.98	47.91	16.19	47.60	17.49	42.43	18.82	54.12
Participant-oriented	8.91	24.53	3.91	20.83	5.17	15.21	6.60	16.02	4.47	12.85

TABLE 17 Functional distribution of four-word bundles by status (per 10,000 words & %)

Bundle orientation	Research students		Junior academics		Senior academics		BNC academic	
Research	29.62	46.80	59.61	45.50	33.62	44.00	47.72	40.32%
Text	25.76	40.70	54.23	41.40	32.32	42.30	51.71	43.69%
Participant	7.91	12.50	17.16	13.10	10.47	13.70	18.92	15.99%

clearly, especially in framing arguments and showing unambiguous connections between ideas and sentences (11). Portuguese language authors, however, are at greater pains to bring their readers into their texts and make certain that their stance is conveyed unambiguously (12).

1. The mechanical properties and proton conductivity of the acid doped blend membranes were improved **at the same time**. (Chinese)
2. **The beginning of the** growing of these two flowstone could be contemporaneous according to the U/Th dating. (French)
3. **In the light of** what has been stated above this is a misunderstanding. (Swedish)
4. The representation (Figure 2) shows this return by **the fact that the** network of peers is on the two extreme positions of the scheme. (Portuguese)

Regarding the influence of research experience, we noted earlier that the normed frequency of bundles differs considerably across the language groups and in comparison with the published texts. Looking at the percentages in Table 17, however, a similar *proportion* of functions is apparent across the different seniority groups and every EAL author group exceeds the published group in its use of research-oriented bundles. Increasing experience seems, gradually, to bring the proportionate use of bundle categories closer to those found in the BNC texts, with research bundles gradually falling and text and participant types increasing. While we recognize our corpus is too small to allow firm conclusions, we might tentatively suggest this could indicate an increasing awareness of more widely used forms by the EAL writers and a growing correspondence towards familiar uses in published articles.

Finally, we once again examined the disciplinary context in which the EAL writers worked. Here, in Table 18, we see relatively high use of research-oriented bundles by writers of life and natural science texts.

In the science fields, research bundles ae mainly used to depict research procedures, showing the ways that studies were conducted (13 and 14) or help to specify aspects of models, equipment, materials, or the research environment, and were typically realized by noun phrase + of structures (15 and 16).

TABLE 18 Functional distribution of four-word bundles by disciplines (per 10,000 words & %)

Bundle orientation	Humanities		Social science		Life science		Natural science		BNC academic	
Research	27.94	42.40	32.27	42.80	28.69	44.20	49.54	45.20	47.72	40.32%
Text	26.95	40.90	30.16	40.00	26.22	40.40	43.51	39.70	51.71	43.69%
Participant	11.01	16.70	12.97	17.20	9.99	15.40	16.55	15.10	18.92	15.99%

1. This weight concern sum score **was used as a** continuous variable to describe the level of overall weight concerns in the regression model. (Life sciences)
2. Powder X-ray diffraction **experiments were carried out** on a Stoe Stadi MP in vertical set-up using the transmission mode. (Natural sciences)
3. **The formation of a** solid carbamate increases the absorption capacity of the solution … (Natural sciences)
4. Oxidative stress can damage **the structure of the** synapse and inhibit the synaptic transmission. (Life sciences)

This emphasis on the ways the research was conducted plays an important role in conveying the grounded, experimental basis of research in the hard sciences. In addition, it conveys a scientific ideology which emphasizes the empirical over the interpretive, minimizing the presence of researchers and contributing to the objective claims of the sciences. This may be a conscious effort by the EAL authors to stress their research rather than its presentation by placing greater emphasis on research practices and the methods, procedures, and equipment used. In this way it is possible to present credible generalizations rather than show research to be the result of interpreting individuals.

9 | CONCLUSION

This study has explored an important aspect of world Englishes research and a growing issue for practitioners in ESP: that of the participation of EAL authors in international publishing. In particular, we have sought to understand the contribution made by writers with different L1s to the formulaic patterns of academic writing, examining the types, structures, and functions of the most frequently repeated building blocks of texts and considering the impact of various contexts on these uses.

Overall, we have found that EAL writers use a more restricted range of types, deploy fewer bundles in their texts, and use more verbal and clause-based bundles than those found in published papers. This means that many patterns used by the writers of published papers are not found in the EAL papers while others appear far less frequently. It also means that the EAL writers' texts contain relatively more forms which package information as passives and fronted by anticipatory it structures, and that they use more research-oriented bundles, focusing on the topic of the study rather than on its presentation and engagement with readers. We found that the L1 background of the writers seems to have some impact on their choice of bundles, with relatively little overlap in preferred forms and the extent to which they make use of them. There were also slight preferences for different structural patterns and major divergences in preferred functions.

It is difficult to conclusively attribute these differences to the influence of first language, however, as the results are cross-cut by seniority and discipline. Students use fewer types and tokens than academics, and junior scholars employed far more bundles and used them more frequently than the other groups, although this result may be influenced by the greater number of texts in our corpus by this group. There does seem, however, to be a movement towards greater correspondence to published uses in the preferred rhetorical functions which the EAL writers use. Discipline also seems to be an important contextual variable in EAL writing, as we might expect. ELF authors working in the natural sciences, for example, employ considerably more bundles than those working in other fields and well over twice the number as the published authors. There seems, however, to be little variation in the structures or the functions although EAL scientists employ more research-oriented types.

In general, I think we would want to say that these results suggest that when considering lexical bundles, writers with different language backgrounds all bring something different to the act of academic writing. They construct their texts using collocational building blocks which often differ from those found in published texts and from the texts of other L1 writers. It is important to underline here that our findings are descriptive: they attempt to paint a picture of the way academics write and not a view of how they should write. These language choices do not make their texts 'wrong,' 'inappropriate,' or 'non-native' in any way, but because texts are, at least to some extent, characterized by repeated customary combinations, readers may notice unfamiliar uses or the absence of more familiar ones. Every additional paper by an EAL writer, however, contributes to an ever-changing code as academic English gets appropriated and adapted in its use. Academic discourse is a melting pot of Englishes; a place where different varieties are constantly in contact so that the ever-increasing participation of EAL authors in global publishing will, very likely, slowly enlarge the variety of bundles we see in professional texts.

ENDNOTES

[1] https://www.scimagojr.com/countryrank.php?year=2019

REFERENCES

Anthony, L. (2020). *AntGram (1.2) [Computer software]*. Tokyo: Waseda University.

Belcher, D. (2007). Seeking acceptance in an English-only research world. *Journal of Second Language Writing*, *16*, 1–22.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

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

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

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

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

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

Clavero, M. (2010). 'Awkward wording. Rephrase': Linguistic injustice in ecological journals. *Trends Ecology Evolution*, *25*, 552–553.

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

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, *17*, 391–406.

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

Ferguson, G., Pérez-Llantada, C., & Plo, R. (2011). English as an international language of scientific publication: A study of attitudes. *World Englishes*, *30*, 41–59.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor, & A. Marchi, Eds., *Corpus approaches to discourse: A critical review*. (pp. 225–258). London: Routledge.

Guardiano, C., Favilla, M., & Calaresu, E. (2007). Stereotypes about English as the language of science. *AILA Review*, *20*, 28–52.

Hanauer, D., & Englander, K. (2011). Quantifying the burden of writing research articles in a second language: Data from Mexican scientists. *Written Communication*, *28*, 403–416.

Haswell, R. (1991). *Gaining ground in college writing: tales of development and interpretation*. Dallas: Southern Methodist University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

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

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

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, *32*, 150–169.

Hyland, K. (2016). Academic publishing and the myth of linguistic disadvantage. *Journal of Second Language Writing*, *31*, 58–69.

Hyland, K., & Jiang, F. K. (2018). Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics*, *23*, 383–407.

Lee, D., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, *18*, 281–296.

Ma, G. H. (2009). Lexical bundles in L2 timed writing of english majors. *Foreign Language Teaching and Research*, *41*, 54–60.

Na, L., & Hyland, K. (2016). Chinese academics writing for publication: English teachers as text mediators. *Journal of Second Language Writing*, *33*, 43–55.

Na, L., & Hyland, K. (2019). “I won’t publish in Chinese now”: Publishing, translation and the non-English speaking academic. *Journal of English for Academic Purposes*, *39*, 37–47.

Page, B. (2020). Elsevier sees 2019 profits and revenues lift. The Bookseller. Retrieved from https://www.thebookseller.com/news/elsevier-sees-profit-and-revenue-lift-1192787

Pang, W. (2010). Lexical bundles and the construction of an academic voice: A pedagogical perspective. *Asian EFL Journal*, *47*, 1–13.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. Amsterdam: John Benjamins.

SciELF. (2015). *The SciELF corpus*. Available at http://www.helsinki.fi/elfa/

Scott, M., & Tribble, C. (2006). *Textual patterns*. Amsterdam: John Benjamins.

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

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*, 214–225.

Tollefson, J. (2018). China declared world’s largest producer of scientific articles. Nature. Retrieved from https://www.nature.com/articles/d41586-018-00927-4

UNESCO. (2017). *Science report: Towards 2030*. Retrieved from https://en.unesco.org/unesco_science_report/

How to cite this article: Hyland, K., & Jiang, F. K. (2022). Bundles in advanced EAL authors’ articles: how do they compare with world englishes practices?. *World Englishes*, 1–17. https://doi.org/10.1111/weng.12605