Pertanika J. Soc. Sci. & Hum. 31 (3): 1015 - 1039 (2023)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Structures and Functions of Lexical Bundles in Student and Expert Timed Argumentative Writing: A Corpus-Driven Study

Nghia Xuan Nguyen

Department of Applied Linguistics and British-American Studies, Faculty of Foreign Languages, Hanoi University of Science and Technology, 100000 Hanoi, Vietnam

ABSTRACT

Despite rich research into lexical bundle use in documented writing on both expert and novice levels, little is known about this linguistic behaviour in undergraduate timed argumentative writing. This study aimed to narrow this gap by examining two self-compiled corpora: a student corpus composed of 200 timed argumentative essays written by first-year students at a Vietnamese university and an expert corpus comprising 200 essays of the same type produced by IELTS professional writers. Employing existing bundle classification frameworks, it compared structures and functions of four-word bundles across the corpora. Results showed that bundle use was more clausal than phrasal in both corpora, inconsistent with a few previous studies. In addition, the student corpus outnumbered the reference corpus in all broad structural categories, which can be ascribed to the over-manipulation of the "Pronoun/NP + be/VP fragment", "NP with *of*-phrase fragment", and "PP with *of*-phrase fragment" constructions. Function-wise, the ubiquity of stance and discourse organising bundles in both the corpora reflected the nature of argumentation, on the one hand, and the rigidly structured IELTS writing format, on the other. The study concluded with methodological implications for bundle research and pedagogical implications for L2 writing.

Keywords: Argumentative writing, bundle functions, bundle structures, corpus linguistics, lexical bundles, timed writing

ARTICLE INFO

Article history:
Received: 24 June 2022
Accepted: 08 March 2023
Published: 04 August 2023

DOI: https://doi.org/10.47836/pjssh.31.3.06

E-mail address:
nghia.nguyenxuan@hust.edu.vn (Nghia Xuan Nguyen)

ISSN: 0128-7702
e-ISSN: 2231-8534

INTRODUCTION

Over the past decades, investigations into large corpora such as the Bank of English (Sinclair, 1991) and the London-Lund Corpus of Spoken English (Svartvik, 1990) brought to the fore the fact that a large proportion of language is prefabricated or formulaic. This proportion was reported

to range between 32% (Foster, 2001) and 58% (Erman & Warren, 2000). Formulaic sequences, in Wray and Perkins' (2000) words, are "stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p. 1), so alleviating the cognitive load when learners process the language (L. Wei & Ying, 2011) and facilitating language production (Granger, 2018; Stengers et al., 2011; Wood, 2002).

Formulaic sequences come in several forms, such as idioms, collocations, and lexical bundles (Paquot & Granger, 2012). Among these, lexical bundles, or chunks (De Cock, 1998), n-grams (Stubbs, 2007a, 2007b), and clusters (Hyland, 2008a, 2008b), as termed by other scholars, have received growing attention for their unique semantic and structural attributes and a bottom-up empirical identification approach. Unlike idioms, they are stripped of idiomatic nature, and unlike collocations, they can be syntactically incomplete, as in "*the end of the"* (Biber & Barbieri, 2007; Y. Wei & Lei, 2011). Notably, lexical bundles are the result of computational corpus-driven retrieval. A word chunk must occur at a normalised frequency threshold of 20–40 times per million tokens and across at least 3–5 texts in the corpus to eliminate individual writers' idiosyncrasies to qualify as a lexical bundle. Biber et al. (1999) defined the construct as the most frequently recurring word sequences in each register or genre. Although this linguistic feature has inspired a decent body of research aimed at postgraduate and research writing, little is done relative to the specific genre of undergraduate argumentative writing under a timed condition, so further investigations are warranted. The next discussion provides a detailed account of this.

Previous Research on Structures and Functions of Lexical Bundles

Lexical bundle research can be summarised into at least three sublines of inquiry. Early researchers were interested in bundles within and across spoken and written registers with Biber et al.'s (1999) seminal work. The authors extracted bundles from the Longman Spoken and Written English corpus, based on which they developed a 12-category taxonomy for structural classification and found that bundles in conversation are more clausal (e.g., *I would like to*). At the same time, those in academic prose are more phrasal (e.g., *as a result of*). A series of other Biber's studies followed this work and echoed its findings (Biber et al., 2003, 2004; Biber & Barbieri; 2007), with Biber et al. (2004) adding that teaching discourse, as an intermediate register, structurally resembles both conversations in clausal bundles and academic prose and textbooks discourse in phrasal bundles.

The second set of studies, which has received the most rigorous consideration by far, focuses on academic research and disciplinary writing. They looked at discrepancies in bundle use between professional writers and student writers, among professional writers *per se*, and among student writers *per se*. It is agreed that professionals and students do not differ significantly in usage patterns of

bundle structures as both groups of writers employ clusters that are based more on noun phrases (NP; e.g., *this point of view*) and prepositional phrase (PP; e.g., *in the number of*) than verb phrase (VP; e.g., *it is possible to*) (Chen & Baker, 2010; Cortes, 2004, 2008; Hyland, 2008a, 2008b; Y. Wei & Lei, 2011). In a similar spirit, there is a consensus that native professionals produce more NP- and PP-based chunks while non-native professionals depend on VP-based forms (Akbulut, 2020; Esfandiari & Barbary, 2017; Fajri et al., 2020; Pan et al., 2016). As far as student writing is concerned, studies by Bychkovska and Lee (2017) and Lu and Deng (2019) converge that bundle use is verb-heavy in Chinese student writing compared to English student writing.

Function-wise, it is important to note that researchers utilised two distinctive taxonomies, i.e. that of Biber et al. (2004), with three broad categories of stance (e.g., *it is difficult to*), discourse organising (e.g., *on the one hand*), and referential expressions (e.g., *when it comes to*) (Adel & Erman, 2012; Bychkovska & Lee, 2017; Chen & Baker, 2010; Cortes, 2008), and Hyland's (2008a, 2008b) research-specific version consisting of research-oriented (e.g., *a wide range of*), text-oriented (e.g., *as shown in figure*), and participant-oriented functions (e.g., *it is suggested that*) (Akbulut, 2020; Lu & Deng, 2019; Pan et al., 2016; Y. Wei & Lei, 2011). Mixed findings have been reported. For example, while native English students use more stance bundles (Adel & Erman, 2012), non-native Chinese counterparts prefer bundles of this type (Bychkovska & Lee, 2017), or while professional writing represents remarkably more text-oriented than research-oriented bundles (Hyland, 2008a, 2008b), there is a similar amount of these functions between professional writing and student writing (Y. Wei & Lei, 2011).

The third subdomain of bundle inquiry pertains to undergraduate general academic writing, whose literature is scarce and methodologically inconsistent, so it interests the present study. Ping (2009) compared two subcorpora of argumentative scripts written by Chinese-L1 and English-L1 students. The scholar suggested that the Chinese corpus mainly contained clausal multiword and stance-oriented expressions, especially those with the embedded third-person plural pronoun *we*. It is congruent with Nam and Park (2020), who examined argumentative essays composed by Korean and American students and showed that the former writers relied more heavily on VP-based and stance sequences than the other group. Staples et al. (2013) and Chen and Baker (2016), dealing with bundle use across proficiency levels, were unanimous in that stance bundles persisted across all three levels. Chen and Baker (2016), however, added that bundles used in lower levels mirrored a more personal tone and those in higher levels more impersonal, and in terms of structure, lower-level essays were characterised typically by VP-based bundles while more proficient writings by more nominal constructions. With all that has been discussed, these conclusions should be treated with great caution because of the inconsistencies in the studies' methodological arguments.

The Gap in Bundle Literature and Research Questions

The impetus for this study was two-fold. Firstly, while studies on bundle use across registers and academic disciplines are many, we know very little about the domain of timed argumentative essays, one of the most widely accepted text genres of undergraduate academic writing (Johns, 1995). Secondly, those rare studies display methodological limitations that potentially render their findings questionable, the most problematic of which seems to be inadequate management of confounding influences. Specifically, the subcorpora in Ping (2009) contained both timed and untimed writings of different mean lengths, while intuitively, unmatched time conditions and length requirements may activate different language use mechanisms. Chen and Baker (2016), while trying to ensure data homogeneity as to time condition, essay length, and topic, did not discriminate argumentative writing from general expository writing. Staples et al. (2013) compared two unparallel datasets: a self-compiled corpus of TOFEL-iBT timed exam responses and Chen and Baker's (2010) corpus of discipline-specific academic writing. The present study compares bundles to use in timed argumentative scripts written by Vietnamese university students and professional IELTS writers, trying to fine-tune previous research from a methodological perspective. In so doing, it sought to address two questions below:

1. What are the structural patterns of lexical bundles used by Vietnamese student writers and expert writers in timed argumentative essays?
2. What are the functional patterns of lexical bundles used by Vietnamese student writers and expert writers in timed argumentative essays?

METHODOLOGY

Corpora Construction

The present study employed a whole corpus approach (Brezina, 2018). Materials used were two self-compiled corpora: a student corpus and an expert corpus (henceforth SC and EC). The SC was constructed based on Granger's (2013) learner corpus design model, which highlights two variables associated with the learner and the task. For the sake of comparability, variables persisting across the corpora were: task type (IELTS Task 2 simulated essay), text genre/rhetorical function (argumentative), conditions (timed and short writing), and discipline (general topics).

The SC consisted of 200 scripts produced mainly by first-year students undertaking the English major program and partly by a small group of other freshmen taking English modules as prerequisites at a university in Vietnam between 2017 and 2020. The essays were from mid-term and final exam papers with the task assimilated to IELTS Task 2, in which the students composed an extended response of at least 250 words to a general topic. The complete list of chosen topics can be found in Appendix A1. The request for access to these scripts was made to the faculty's management board, and

they were offered anonymity as the student information had been removed. Only essays scoring 6.0 and 8.0 (out of 10) were selected to limit the observed proficiency level to B2 upper-intermediate. These cut scores were determined concerning the Vietnamese Standardised Test of English Proficiency (VTSEP) score conversion guide, which is, in turn, based on the Common European Framework of Reference for Languages (CEFR) scale descriptors. All the chosen exam papers had previously been marked by EFL writing instructors, following an analytical rating procedure. As a step of the second rating, essays whose scores the researcher disagreed with were purposively removed.

The subsequent stage in the SC construction was typing. There was either no or insufficient documentation of how texts were typed in preceding studies until Bychkovska and Lee (2017), Huang (2015) and Shin et al. (2018) scrutinised bundle errors. Whether grammatical, lexical, and orthographical errors are corrected should largely affect the results. Not departing from an error-analysis approach but rather appreciating learners' attempts to use patterned language, the researcher corrected errors that would potentially jeopardise the status of a bundle. In so doing, the researcher referred to the authors mentioned above for common bundle errors and decided upon four main types: preposition (*in the one hand – on the one hand*), article (*as the result – as a result*), word form (*it is clearly that – it is clear that*), and spelling (*beleive – believe*). Chen and Baker's (2016) list of bundles was consulted since the study is closest to the present study in task and learner characteristics to filter out potential bundles for correction. Except for the erroneous bundles, the essays were typed verbatim in MS Word and converted into *.txt* files.

The reference corpus EC consisted of 200 essays written by IELTS experts. The tricky 'expert' quality was assured by accepting only compositions from (1) published IELTS books, (2) a reliable IELTS website, and (3) two renowned IELTS trainers: Simon and Mat Clark. The books were checked for credibility by referring to their publishing houses (the book list can be found in Appendix A2). The chosen website was *ielts-exam.net,* whose scripts considerably overlapped with those from published books; the two IELTS experts are well-known for their public dissemination of Band 9.0 compositions. As the rhetorical function was argumentation, only these questions were opted for: (1) *To what extent do you agree or disagree?* (2) *Do you agree or disagree?* (3) *Discuss both views and give your own opinion*, and (4) *Do the advantages outweigh the disadvantages?* The essays were typed verbatim in MS Word and converted to *.txt* files. Table 1 sketches out the corpora.

Retrieval and Filtration of Bundles

AntConc, a software program developed by Anthony (2014), enabled the automatic retrieval of lexical bundles. Input values were needed: the bundle length, the normalised cut-off frequency, and the dispersion range.

Table 1
Description of two corpora

Characteristics	SC	EC
Task type	IELTS Writing Task 2	
Text genre	Argumentative essay	
Writing condition	Timed	
Topics	General	
No. of texts	200	200
Mean text length	335 (SD = 71)	290 (SD = 30)
Range	210–582	251–405
Mean score	6.97	N/A
Corpus size (tokens)	67,006	57,951

This study considered four-word bundles under their salience in bundle scholarship. The widely documented frequency and dispersion criteria figures are 20–40 times per million words and at least five texts. While the observed corpora were small, a normalised frequency threshold of 40 times per million words was taken, equivalent to a raw frequency of 2.6 times for the SC and 2.3 times for the EC. As these values were still below the dispersion criterion, even accepting the minimum requirement of three texts, the dispersion was set first to four texts, raising the raw frequencies to four times for each corpus. This establishment retrieved 547 types and 4806 tokens in the SC and 98 types and 577 bundles in the EC.

Removing bundles that may inflate the overall results appears to be what past research ignored or poorly handled until Chen (2009), in her doctoral dissertation, worked out a 'pseudo' bundle treatment procedure. In this study, these 'disguised' bundles were manually filtered according to the exclusion criteria in Table 2.

Table 2
'Pseudo' bundles removal criteria

Removal criteria	Description
(a) Context-dependent bundles	
i. Prompt-based bundles	A great number of 'true' bundles stem from the prompts in the SC, e.g., *some people believe that* (prompt #9). The question was whether the writers truly possessed this bundle in their repertoire. To illuminate this, I checked the concordance lines and established two selecting criteria: the paraphrasing degree and the assigned score. If a chunk of text containing the prompt-based bundle rephrased the prompt at an acceptable level, plus the essay was awarded at least 7.0 (the SC's mean score), the bundle would be retained or otherwise deducted from frequency statistics. Other obvious prompt-based bundles, like *the main source of information* or *media, such as newspaper*s, were easily eliminated.

Table 2 *(Continue)*

Removal criteria	Description
ii. Topic-related bundles	In conformity with Biber et al.'s (1999) accentuation of register/genre-specific rather than topic-specific nature of bundles, those related to the topics were disqualified. For instance, the bundle *at an early age* in the EC was used in texts responding to different prompts on the same topic of young children's education, so it was taken off the list. Ambiguous bundles such as *have the right to* and *society as a whole* remained counted as they were found across texts with different topic areas.
(b) Overlapping bundles	
i. Complete overlap	Refers to two four-word bundles with matching frequencies derived from the same five-word bundle. For example, the two bundles, *there is no doubt* and *is no doubt that,* occur 16 times, so they were merged as *there is no doubt + that* with 16 counts.
ii. Complete subsumption	Refers to two four-word bundles being part of a five-word bundle but differing in occurrences. For example, *in conclusion, I think* in the EC occurs 8 times, and *conclusion I think that* occurs 6 times. The latter is completely subsumed in the former in frequency, so it was noted as *in conclusion, I think (+ that)* with a frequency of 8 times and the latter was removed.
iii. Partial subsumption	Refers to a situation in which two four-word bundles derived from the same five-word bundle, differing in occurrences but not completely subsumed in each other. For instance, *more and more popular* with 11 counts and *is more and more* with 6 counts; sharing 4 counts of the longer bundle *is more and more popular*. In this case, it was rewritten as *(is +) more and more popular*, with a finalised frequency being the sum of occurrences of the component bundles minus that of the shared five-word bundle, or 13.

Bundle types and tokens in both corpora dwindled after the manual extraction (see Table 3).

Analytical Frameworks

The present study analysed lexical bundles considering Chen and Baker's (2010)

Table 3
Bundles in two corpora before and after removal of 'pseudo' bundles

	SC		EC	
	Types	Tokens	Types	Tokens
Before filtration	547	4806	98	577
After filtration	139	1057	77	490

Note. Type = the number of different bundles; Token = the number of occurrences of the bundles

structural taxonomy and Chen and Baker's (2016) functional taxonomy. The structural framework originated from Biber et al. (1999). It was taxonomised by Chen and Baker (2010) into three broad categories, i.e., NP-based, PP-based, and VP-based, with each category entailing a few subcategories. It is the only framework to date that has been employed by researchers for categorising lexical bundles structurally.

The functional taxonomy underwent a more complex evolutionary process. It was rooted in Biber et al. (2004) and constantly modified by Biber and Barbieri (2007), Chen (2009), Chen and Baker (2010, 2016). Chen and Baker's (2016) version was adopted due partly to its recency and partly to its nature of being non-research-directed as opposed to Hyland's (2008a, 2008b), which caters better to research-based writing. Due to space constraints, the chosen taxonomies are not presented separately but are embedded in Tables 5 and 7.

Data and Statistical Analysis

The lexical bundles, upon retrieval, were each assigned a code corresponding to its subcategory, e.g., N1 for the "NP with *of*-phrase fragment" structure or S1 for an epistemic function, after having been meticulously checked with literature and concordance lines. They were then grouped by the letter code they shared. Intra-coding was carried out twice since the first coding, with the second coding two weeks after the first coding and the third another two weeks after the second. When there were inconsistencies in the coding, I consulted Chen's (2009) bundle list, in which each item was transparently structurally and functionally categorised. Type and token frequencies and percentages were then computed for descriptive analysis. As the two corpora were unequal, the occurrences were normalised to 100,000 words. The log-likelihood (LL) tests were applied to raw token occurrences of each main category and subcategory for significance testing. Lancaster University's Log-likelihood online calculator was employed to this end. LL critical values were as follows: an LL of 3.84 or higher was significant at $p < 0.05$, an LL of 6.63 or higher was significant at $p < 0.01$, an LL of 10.83 or higher was significant at $p < 0.001$, and an LL of 15.13 or higher was significant at $p < 0.0001$.

RESULTS AND DISCUSSION

Structural Differences in Bundle Use Between Two Corpora

Overall, VP-based bundle types dominated both the corpora, accounting for nearly two-thirds of each corpus (64% in the SC and 63.6% in the EC), while NP-based and PP-based bundle types combined made up a third of each corpus (34.5% in the SC and 33.8% in the EC; see Table 4). The figures suggest nearly identical distributional proportions of the three main categories across the corpora, with a difference of less than 1%. Token-wise, this pattern recurred despite a widened yet inconsiderable gap, signifying a highly consistent type and token distributional reality that bundles used in student writing and expert writing were more clausal than phrasal. This result

Table 4
Distribution of broad structural categories in two corpora

Category	Types (%)		Tokens (%)	
	SC	EC	SC	EC
NP-based	12.9	13	12.3	10.6
PP-based	21.6	20.8	25.4	25.1
VP-based	64	63.6	59.5	62.5
Others	1.5	2.6	2.8	1.8

contradicts previous findings in a few ways. Relating to register, scholars (Biber et al., 1999, 2004; Biber & Barbieri, 2007) agreed that clausal bundle use is an attribute of spoken rather than written discourse, while the examined discourse was primarily clausal. Also, while clausal bundles are used by student writers and phrasal bundles more frequently by expert writers (Chen & Baker, 2016; Cortes, 2008; Hyland, 2008a, 2008b), this does not completely hold for this study as clausal bundles were representative of both sets of writings.

Table 5 further illuminates how lexical bundles were structurally used across the corpora. If we look at each main category, it is visible that the between-corpus proportional pattern deduced above no longer applies. All major categories of bundles were used at least twice more frequently in the SC than in the EC; the differences were statistically significant at $p < 0.0001$. This finding agrees with previous studies (Chen & Baker, 2016; Nam & Park, 2020; Ping, 2009) in that non-native writing contains more verbal constructions, but at the same time, disagrees with others (Bychkovska & Lee, 2017; Pan et al., 2016) in that they found more nominal and prepositional constructions in expert writing. Scrutiny of the subcategories cast further light.

The corpora were heavily clausal due primarily to the "Pronoun/NP + be/VP fragment" construction with respective frequencies of 482 and 200.2 times per 100,000 words (p100kw). This construction is one that Chen and Baker (2010, 2016) modified by adding the VP fragment in the predicate position, allowing a substantial number of instances to be legitimised into this subcategory. Examples from the SC are *some people believe that, I completely agree with*, and *we cannot deny that*, and examples from the EC include *I would argue that, many people believe that*, and *I think it is*. As Chen (2009) contended, lower-level students are driven by language of vagueness (e.g., *some people believe that*) and personal involvement (e.g., *I completely agree with*). Notably, this construction was not uncommon among expert scripts, probably attributable to the nature of IELTS writing as being semi-academic rather than purely. It implies that students should eliminate the belief that minimising clausal formulas alone would help their scripts read native-like.

Table 5
Structural differences in bundle use between two corpora

Category/ Subcategory	Types		Tokens		LL
	SC	EC	SC	EC	
NP-based					
NP with *of*-phrase fragment	11	6	62 (92.5)	34 (58.7)	4.73 (+)*
NP with other post-modifier fragments	7	4	68 (101.5)	18 (31.1)	24.17 (+)****
Subtotal	18	10	130 (194)	52 (89.7)	24.17 (+)****
PP-based					
PP with *of*-phrase fragment	9	5	59 (88.1)	25 (43.1)	9.67 (+)**
Other PP fragments	21	11	109 (162.7)	98 (169.1)	0.08 (-)
Subtotal	30	16	268 (400.0)	123 (212.2)	36.09 (+)****
VP-based					
Anticipatory *it* + VP/AdjP	14	13	81 (120.9)	88 (151.9)	2.20 (-)
Passive verb + PP fragment	-	-	-	-	-
Copular be + NP/ AdjP	7	4	66 (8.5)	20 (34.5)	19.71 (+)****
(VP +) *that*-clause fragment	2	4	9 (13.4)	21 (36.2)	6.84 (-)**
(Verb/adjective) + *to*-clause fragment	4	7	24 (35.8)	45 (77.7)	9.91 (-)***
Adverbial clause fragment	4	3	20 (29.8)	16 (27.6)	0.05 (+)
Pronoun/NP + be/ VP fragment	43	18	323 (482)	116 (200.2)	73.85 (+)****
VP with an active verb	15	-	106 (158.2)	-	-
Subtotal	89	49	629 (938.7)	306 (528.0)	71.94 (+)****
Other expressions	2	2	30 (44.8)	9 (15.5)	9.09 (+)**
Total	139	77	1057 (1577.5)	490 (845.5)	138.56 (+)****

Note. AdjP = Adjective phrase; the number in brackets indicates normalised frequency per 100,000 words; LL = log-likelihood; [+] indicates overuse in the SC, [-] indicates underuse in the SC; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

Table 5 also statistically accounts for greater VP-based bundle use in the SC than in the EC. The above-mentioned "Pronoun/NP + be/VP fragment" construction accounted for the largest part of the difference at $p < 0.0001$. Perhaps this is because the student writers habitually imitated this expert practice before overdoing it in their essays. Meticulous concordance line checks uncovered that this construction was used no more than once per expert essay but a few times per student essay in a sizable portion of the student corpus. Another contributing factor is the "VP with active verb" construction with 158.2 tokens p100kw noted in the SC and none in the EC. Students' over-reliance on this subcategory can be exemplified by a vast token coverage of such bundle types as *plays an important part*, *play an important role*, and *become more and more*. According to Chen (2009), these bundles are often deemed clichéd and overgeneralising, so they are rarely used in expert writing.

Another finding is that whereas the SC outnumbered the EC in most subcategories, it was found underrepresented in the "(VP +) *that*-clause fragment" (e.g., *that they do not*) and "(Verb/adjective) + *to*-clause fragment" (e.g., *to be able to*) constructions at $p < 0.01$ and $p < 0.001$, respectively. Surprisingly, the "Passive verb + PP fragment" bundles (e.g., *can be seen as*) appeared in neither of the corpora. At the same time, according to Y. Wei and Lei (2011), this construction is over-manipulated in student writing.

Phrasal bundles of most types were overused in the SC. The "NP with *of*-phrase fragment" structure (e.g. *one of the most*), albeit its overrepresentation, was more than half realised by overgeneralising bundles such as *a large amount of* or *an important part in*, or bundles with the embedded colloquial quantifier "a lot of" such as *a lot of people*, *a lot of time* or *us a lot of*, which are speech-like (Chen, 2009; Chen & Baker, 2016). The "PP with *of*-phrase fragment" construction also contained speech-like, or L1 translated bundles such as *with the help of*, *with the development of* or *thanks to the development*. Bundles like these inflated both bundle types and tokens in the non-native corpus to a great extent. It calls for students' effort to convey messages using linguistically condensed and content-specific expressions. For example, the sentence *With the help of technology, a lot of people are able to pay online when they shop* can be rephrased as *Technology has enabled numerous shoppers to pay online*.

Functional Differences in Bundle Use Between Two Corpora

Chen and Baker's (2016) functional taxonomy for bundle classification consists of three broad categories: stance, discourse organising, and referential. Stance bundles express the writer's attitude or epistemic evaluation of the certainty of a proposition, so they are subdivided into epistemic bundles (e.g., *it could be argued*) and attitudinal/modality bundles (e.g., *it is difficult to*). Discourse organisers structure the link between the preceding and coming text. They are used to introduce a topic (e.g., *in this essay I*), elaborate or clarify a topic

(e.g., *on the other hand*), and emphasise a piece of information (e.g., *one of the most*). Referential expressions are used to make a direct reference to frame a given attribute or condition (e.g., *as a result of*), qualify a proposition in terms of size, number or extent (e.g., *a large proportion of*), and specify place, time, and text-deictic contexts (e.g., *in the first place*).

While discourse organisers accounted for the largest proportions in the SC (38.9% types and 46.5% tokens), the stance was the most identifiable function in the EC (42.9% types and 44.3% tokens; see Table 6). It is possibly because of the students' being tied to discourse markers that might have been learnt by rote beforehand and of the experts' sophisticated usage of impersonal writing in bringing out the nature of argumentation, respectively. To elaborate on the former, Vietnamese students are often instructed in EFL writing courses to construct their essays using a template that has fixed discourse organising expressions such as *on the one hand*, *on the other hand*, or *last but not least*. The fact that stance bundles (34.5% types and 33.7% tokens) ranked second only to discourse organisers in the SC and discourse organisers (35.1% types and 37.6% tokens) second only to stance bundles in the EC suggests that the students did try to maintain an argumentative tone. At the same time, the experts were not free from using pre-determined discourse organisers. It is likely how the IELTS writing is made and seen—a rigid writing format with a list of pre-set linking expressions.

Table 6
Distribution of broad functional categories in two corpora

Category	Types (%)		Tokens (%)	
	SC	EC	SC	EC
Stance	34.5	42.9	33.7	44.3
Discourse organisers	38.9	35.1	46.5	37.6
Referential	26.6	22	19.8	18.1

Regarding bundle counts instead of percentages (see Table 7), bundles of all main functions were consistently used more often in the SC than in the EC (types and tokens). The discrepancy in the stance function resulted mainly from epistemic expressions, used 426.8 times p100kw by the students and 277.8 times p100kw by the experts at a significance level $p < 0.0001$. As mentioned above, epistemic bundles depict the writer's assessment of the certainty of a claim, which is layered into three levels: least certain (e.g., *is more likely to*), neutral (e.g., *it has been suggested*), and strongest commitment (e.g., *there is no doubt*) and gauged by cautious or tentative language (Hyland, 1998; Poos & Simpson, 2002). It was found from the corpora that a large proportion of the bundles were of neutral quality, and those in the SC were inflated

Table 7
Functional differences in bundle use between two corpora

Category/Subcategory	Types		Tokens		LL
	SC	EC	SC	EC	
Stance					
Epistemic	40	25	286 (426.8)	161 (277.8)	19.63 (+)****
Attitudinal/modality	8	8	70 (104.5)	56 (96.6)	0.19 (+)
Subtotal	48	33	356 (531.3)	217 (374.5)	16.89 (+)****
Discourse organisers					
Topic introduction	1	-	10 (14.9)	-	-
Topic elaboration/ clarification	39	20	334 (498.5)	137 (236.4)	58.86 (+)****
Identification/focus	14	7	147 (219.4)	47 (81.1)	40.62 (+)****
Subtotal	54	27	491 (732.8)	184 (317.5)	103.87 (+)****
Referential					
Framing	19	9	100 (149.2)	44 (75.9)	14.99 (+)***
Quantifying	17	2	95 (141.8)	9 (15.5)	70.99 (+)****
Time/place/text-deixis	1	6	15 (22.4)	36 (62.1)	12.23 (-)***
Subtotal	37	17	210 (313.4)	89 (153.6)	34.42 (+)****
Total	139	77	1057 (1577.5)	490 (845.5)	138.56 (+)****

Note. The number in brackets indicates normalised frequency per 100,000 words; LL = log-likelihood; [+] indicates overuse in the SC, [-] indicates underuse in the SC; *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

by quite a few synonymous bundles of the "Pronoun/NP + be/VP fragment" structure, such as *some people believe that, some people think that*, and *some people argue that*. A smaller percentage was of strong commitment, but while the expert writers tended to prefer the "Anticipatory *it* + VP/AdjP" construction, such as *it is true that* and modal hedges as in *it could be argued*, the student writers resorted to boosting adverbs as in *I completely agree with* (Crompton, 1997; Skelton, 1988). It exhibits that students' exposure to cautious language at the expert level might be scant, so their attempt to use bundles of this type is often intuitive rather than procedurally guided. Kennedy and Thorp (2007) confirmed this, claiming that student writing is sometimes categorical and hyperbolic. Of note is that the student and expert writers seemed to develop an authorial voice by employing the personal pronouns *I* and *we* at roughly equal proportions. It disagrees with Granger and Rayson's (1998) and Petch-Tyson's (1998) statement that non-native learners use more personal pronouns in argumentative writing than professional writers.

Discourse organisers were more common in the SC than in the EC because of the overuse of the "topic elaboration/ clarification" bundles by 498.5 tokens p100kw compared to 236.4 tokens p100kw at $p < 0.0001$. While both groups of writers used complex lexico-grammatical markers such as *on the other hand* instead of *however*, the student writers used some others that the professional writers did not, for example, *last but not least* instead of *finally*, or *as well as the* instead of *and the*, an index of verbosity in student writing (Kennedy & Thorp, 2007). It should be acknowledged, however, that bundles like these were not ubiquitous. However, the actual reason for a surge in this subcategory in the SC was the overuse of informal bundles like *if we do not*, *people do not have*, and *for example, if you*, which were completely absent in the EC. It can be ascribed to students' lack of register consciousness, masking their ability to discriminate speech from writing. Finally, referential expressions were exploited in an unsystematic way in the corpora. The SC recorded an overuse of framing and quantifying bundles but an underuse of time/place/text deictic bundles. Of these, an obvious pattern emerged from the quantifying bundles in the SC that 9 out of 17 bundle types entailed the informal speech-oriented quantifying words/phrases *a lot of* and *many* while the only two bundle types in the EC *the rest of the* and *a large proportion of* were free of such quantifiers. As suggested earlier, framing and quantifying bundles would decline greatly if more content-specific words like *numerous* substituted such quantifiers.

Besides what has been discussed, the above-listed findings can be further elucidated from a few angles. Firstly, IELTS writing is semi-academic rather than fully academic, so experts are not completely independent of VP-based bundle use. They develop a range of discourse organisers and stance bundles at their disposal, such as *there is no doubt* or, *on the other hand*, to deal with a pre-determined writing format. In their learning process, students may copy or at least, in part, be instructed to transfer those expressions to their writing, resulting in an inevitable overuse. Secondly, expert writers may deliver their writing styles. A review of a sample of the expert essays reveals that the argumentation style differs from this essay to another, especially those written long ago, like in the first few books of the IELTS Cambridge series. Finally, according to Chen and Baker (2016), at the B2 proficiency level, students are transitioning from an informal writing style to a formal one, so their writing inexplicably reflects speech-driven language use behaviours such as cliché, overgeneralisation, and verbosity.

CONCLUSION

This study compared student and expert use of four-word bundles in timed argumentative writing and arrived at the following major findings. Bundles chosen by both writers were found to be structurally more clausal than phrasal due to excessive use of the "Pronoun/NP + be/VP fragment" pattern. Overuse of the broad categories in the SC was also observed, contributed by

the "Pronoun/NP + be/VP fragment", "NP with *of*-phrase fragment", and "PP with *of*-phrase fragment" constructions. Functionally, discourse organisers were most prevalent in the SC, while stance bundles prevailed in the EC, but stance bundles and discourse organisers were not uncommon in the respective corpora. Epistemic, topic elaboration/clarification, and framing bundles were the primary contributors to the greater use of each main function in the SC compared to the EC.

Implications

This study has important implications. Firstly, the findings summarised above were outputted based on the methodological reasonings presented throughout this paper, where an incongruence in data type and/or methodological arguments of earlier studies, even some most recent ones, were attended to, and where methodologically informed claims in previous literature were inherited. In this light, the present study suggests that lexical bundle research closely considers the following facets:

- Corpus construction should rest upon data homogeneity concerning task and learner characteristics if two or more corpora are to be compared. In this regard, Granger's (2013) model can be a starting point.
- At the onset, whether the researchers are interested in misused bundles it should be stated and justified, as this may affect their decision to accept or dismiss a certain bundle during the text treatment process.
- Upon software-based retrieval of bundles, a stringent filtration procedure is highly recommended. A transparent "blueprint" for handling context-dependent and overlapping bundles should be in place and strictly followed so as not to distort statistical results.

The study also has implications for pedagogical practices. Expanding on previous studies which support the coverage of lexical bundles in the L2 writing curriculum, it adds that they should only be taught concerning a corresponding text genre since they may not recur in a different genre. For example, the bundle list produced in this study should be used for teaching argumentative essays at an undergraduate level, but the ones previously developed appear more fitted for postgraduate and research writing levels.

Evidence as to bundle frequencies and preferences over certain structural and functional categories among student and expert writers from this study also calls for a need on the part of L2 writers to develop a register awareness. Not only should they be instructed to avoid heavily academic language and colloquial non-academic language use, but they should also be pointed to the integration of cautious language and impersonal writing practices and the minimisation of speech-like language behaviours. It can be achieved through recognition and production activities that

teachers design themselves or material developers procedurally introduce into L2 writing coursebooks. An example activity sequence is that students are asked to notice the forms and functions of lexical bundles in model texts, then do control practice exercises such as gap filling or error identification and correction, which arguably would help them acquire a bundle repertoire for subsequent sentence completion and free extended writing tasks. For this to happen, however, it might be essential that lexical bundle training be given to those parties not directly involved in bundle research, including teachers and material developers.

Limitations and Suggestions for Future Studies

This study has inevitable shortcomings. One is the lack of multiple formal ratings applied to the student essays. Although I was a second-rater and selected only the essays with assigned scores I agreed upon, a more robust marking procedure was desirable.

Another constraint is that this study directly compared the B2 proficiency level to the expert level without considering the middle levels of B1 and C1. It should be acknowledged, however, that lower-level writings not only contain fewer bundles but are also usually under length, making corpus data too small and possibly unrepresentative of the linguistic behaviour under concern. Meanwhile, writings of advanced levels are hard to collect, for participants rarely achieve this level at the time of research, especially in a corpus-driven study that requires a contribution by a large number of participants. If data is available, future studies are suggested to include B1 and C1 proficiency levels to enrich studies that investigate a continuous proficiency range.

The limitation of data size is also worth discussing. The two corpora in this study were relatively small in comparison with corpora of tens or hundreds of millions of words in bundle research. It is due mainly to the unavailability of expert essays. Though one may claim that a Band 9.0 score can guarantee the expert level, this is far from the truth since such a score does not seem to be awarded based on a strict marking procedure regarding scripts supplied online or in unaccredited books. Furthermore, large corpora often deal with multiple disciplines, text genres, and proficiency levels, while researchers have long agreed that corpora of thousands of words are not less meaningful than gigantic text bodies.

Finally, the present study was a comparative study whose purpose was limited to describing bundle usage. Further studies may utilise the bundle list in Appendix B and put it to the test to see whether explicit instruction of bundles can lead to L2 writers' enhanced recognition and/or production performance.

ACKNOWLEDGEMENT

This research is funded by the Hanoi University of Science and Technology (HUST), Vietnam under project number T2021-XH-004. The researcher is grateful to his friend, Nam Tran, for his continuous support and encouragement throughout this project. The researcher also thank two

anonymous reviewers for their valuable comments on the manuscript.

REFERENCES

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*, 81-92. https://doi.org/10.1016/j.esp.2011.08.004

Akbulut, F. D. (2020). A bibliometric analysis of lexical bundles usage in native and non-native academic writing. *Journal of Language and Linguistics Studies*, *16*(3), 1146-1166. https://doi.org/10.17263/jlls.803583

Anthony, L. (2014). AntConc (version 3.4.3) [Computer software]. Laurence Anthony's Website. http://www.laurenceanthony.net

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

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the Lune: A festschrift for Geoffrey Leech* (pp. 71-92). Peter Lang.

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

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

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52. https://doi.org/10.1016/j.jeap.2017.10.008

Chen, Y. H. (2009). *Investigating 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*, 30-49. https://doi.org/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. https://doi.org/10.1093/applin/amu065

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*, 397-423. https://doi.org/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*, 43-57. https://doi.org/10.3366/E1749503208000063

Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. *English for Specific Purposes*, *16*(4), 271-287. https://doi.org/10.1016/S0889-4906(97)00007-0

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, *3*(1), 59-80. https://doi.org/10.1075/ijcl.3.1.04dec

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29-62. https://doi.org/10.1515/text.1.2000.20.1.29

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. https://doi.org/10.1016/j.jeap.2017.09.002

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), 76-89. https://doi.org/10.17323/jle.2020.11271

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogical tasks: Second language learning, teaching, and testing* (pp. 75-93). Longman. https://doi.org/10.4324/9781315838267-11

Granger, S. (2013). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Benjamins. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2018). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228-247). Routledge. https://doi.org/10.4324/9781315206615-13

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). Addison Wesley Longman. https://doi.org/10.4324/9781315841342-9

Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System*, *53*, 13-23. https://doi.org/10.1016/j.system.2015.06.011

Hyland, K. (1998). *Hedging in scientific research articles*. John Benjamins. https://doi.org/10.1075/pbns.54

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41-62. https://doi.org/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. https://doi.org/10.1016/j.esp.2007.06.001

Johns, A. (1995). Teaching classroom and authentic genres: Initiating students into academic cultures and discourses. In D. Belcher & A. Hirvela (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 277-291). Ablex.

Kennedy, C., & Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS academic writing task. In L. Taylor & P. Falvey (Eds.), *IELTS collected paper: Research in speaking and writing assessment* (pp. 316-378). Cambridge University Press.

Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, *39*, 21-36. https://doi.org/10.1016/j.jeap.2019.03.008

Nam, D., & Park, K. (2020). I will write about: Investigating multiword expressions in prospective students' argumentative writing. *PLoS ONE*, *15*(12), Article e0242843. https://doi.org/10.1371/journal.pone.0242853

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. https://doi.org/10.1016/j.jeap.2015.11.003

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130-149. https://doi.org/10.1017/S0267190512000098

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107-118). Addison Wesley Longman. https://doi.org/10.4324/9781315841342-8

Ping, P. (2009). A study on the use of four-word lexical bundles in argumentative essays by Chinese English-majors: A comparative study based on WECCL and LOCNESS. *Teaching English in China*, *32*, 25-45.

Poos, D., & Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings form the Michigan Corpus of Academic Spoken English. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 3-23). Benjamins. https://doi.org/10.1075/scl.9.03poo

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. https://doi.org/10.1016/j.jslw.2017.09.004

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

Skelton, J. (1988). The care and maintenance of hedges. *ELT Journal*, *42*(1), 37-43. https://doi.org/10.1093/elt/42.1.37

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. https://doi.org/10.1016/j.jeap.2013.05.002

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics in Language Teaching*, *49*(4), 321-343. https://doi.org/10.1515/iral.2011.017

Stubbs, M. (2007a). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89-105). Radopi. https://doi.org/10.1163/9789401204347_007

Stubbs, M. (2007b). Quantitative data on multi-word sequences in English: The case of word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 163-189). Continuum.

Svartvik, J. (1990). *The London-Lund corpus of spoken English: Description and research*. Lund University Press.

Wei, L., & Ying, H. (2011). On the role of formulaic sequences in second language acquisition. *US-China Foreign Language*, *9*(11), 708-713.

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

Wood, D. (2002). Formulaic language in acquisition and production: Implications for teaching. *TESL Canada Journal*, *20*(1), 1-15. https://doi.org/10.18806/tesl.v20i1.935

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28. https://doi.org/10.1016/S0271-5309(99)00015-4

Appendix A1

Topics (Prompts) in the SC

1. Computers are being used more and more in education, and some people believe there will soon be no role for the teacher in education. how far do you agree with this opinion?
2. The role of education is to prepare children for a job in the future. Schools should, therefore, cut music and art out of the curriculum so that children can focus on useful and practical subjects such as IT or physics. To what extent do you agree with this point of view?
3. Will modern technology such as the Internet ever replace books or printed media such as newspapers as the main source of information? Discuss this matter and give your own opinion.
4. Having a good university degree can guarantee people a good job. To what extent do you agree with this view?
5. Some people say that the Internet is making the world smaller by bringing people together. How far do you agree that the Internet is making it easier for people to communicate with one another?
6. Some people believe that the media, such as the press, TV and Internet, should be more strictly controlled. Others feel that controls should be lessened to give people freer access to information. What's your opinion on this?
7. Since its coming into existence, fast food has been accused of contributing to the concurrent increase in the number of overweight people. Hence, some argue that heavier taxes should be imposed on these junk foods. To what extent do you agree with this opinion?
8. Some people think that university students should study whatever they like. Others believe that they should only be allowed to study subjects that will be useful in the future, such as those related to science and technology.
9. Some people believe that advances in technology are increasing the gap between the rich and the poor, while others think the opposite is happening. Discuss both views and give your own opinion.
10. Everyone should stay in school until the age of 18. To what extent do you agree or disagree?

Appendix A2

IELTS books used for the EC

Cambridge IELTS series 1-9
IELTS Foundation
IELTS Plus series 1-3
IELTS to Success
IELTS Trainer
Insight into IELTS
New Insight into IELTS
Prepare for IELTS
Academic Writing for IELTS
Collins Writing for IELTS
Barron's Writing for the IELTS
IELTS Advantage
15 Days' Practice for IELTS Writing
Cambridge Guide for IELTS
Exam Essential IELTS Practice Tests 2
IELTS Writing Analyse – Structure and Academic Essays Collection
Best Seller: Writing Task 2 for IELTS
High-score IELTS Writing

Appendix B

Lexical bundles and frequencies in two corpora

Freq	SC	Rank	EC	Freq
51	on the other hand (+ the/ there are/ I/ I believe/ some/ some people)	1	on the other hand (+ there/ some)	35
39	(play/ plays +) an important role in (+ our life/ the)	2	(so +) it is important to	16
26	last but not least	3	it is true that	14
25	(people +) all over the world	4	I think it is	12
24	some people believe that (+ the)	5	I strongly believe that (+ it is)	11
21	(you +) do not have to	6	there is no doubt + that	10
19	(from +) my point of view (+ I)	7	(is +) the best way to	10
19	on the one hand (+ there are)	8	are more likely to	9
18	(because of some following +) reasons first of all	9	on the one hand	9
18	in my opinion I (+ agree with)	10	some people believe that	9
17	is one of the (+ most)	11	as a result of	8
16	some people think that	12	at the same time	8
16	there is no doubt + that	13	I would argue that	8
15	(I +) agree with this opinion	14	in conclusion I think (+ that)	8

Appendix B *(Continue)*

Freq	SC	Rank	EC	Freq
15	(nowadays +) more and more people	15	it is difficult to	8
13	some people say that (+ that)	16	one of the most	8
13	(there +) are more and more	17	(+ is) one of the main	8
13	they do not have	18	part of the world (+ and)	8
13	(is +) more and more people	19	(I +) agree with the statement (+ that)	8
11	do not need to	20	in my opinion the	7
11	in conclusion it is	21	it is possible to	7
11	it is undeniable that	22	so that they can	7
11	some people argue that (+ the)	23	to sum up I	7
10	I strongly believe that (+ the)	24	all over the world	6
10	is the reason why	25	and it would be	6
10	in this essay I (+ will look at both)	26	(I +) do not believe that	6
9	a lot of people	27	I believe it is	6
9	is not the only (+ way to)	28	in a way that	6
8	a large amount of	29	in conclusion I believe + that	6
8	but it is not	30	it seems to me + that	6
8	it cannot be denied + that	31	many people believe that	6
8	for a long time	32	that this is a	6
8	for example if you	33	(there are +) a number of reasons	6
8	many people believe that	34	that it is better (+ for)	6
8	the best way to	35	as a means of	5
8	(+ is) a good way to	36	as long as they	5
7	a lot of time	37	have the right to	5
7	it is easy to	38	I believe that the	5
7	many people argue that	39	I believe that we	5
7	they will not have	40	in the first place	5
7	with the help of	41	it is clear that	5
7	I completely agree with (+ this)	42	it is important for	5
7	people who do not (+ have)	43	it is important that	5
6	a great number of	44	people around the around	5

Appendix B *(Continue)*

Freq	SC	Rank	EC	Freq
6	agree with this idea	45	society as a whole	5
6	agree with this statement	46	that it is not	5
6	another reason is that	47	the rest of the	5
6	as a result the	48	this point of view	5
6	as far as I	49	true to say that	5
6	because of some reasons	50	when it comes to	5
6	I do not agree (+ with this)	51	will be able to	5
6	have a lot of	52	a large proportion of	4
6	I partly agree with + this	53	an important part of	4
6	I think that the	54	at the expense of	4
6	if we do not	55	both positive and negative	4
6	in conclusion I think (+ that)	56	it could be argued + that	4
6	it is believed that	57	however I do not	4
6	it is clear that	58	I also believe that	4
6	it is true that	59	in conclusion I would	4
6	there are many people	60	in the form of	4
6	we cannot deny that	61	in the number of	4
6	we do not have	62	is also true that	4
6	when it comes to	63	it is certainly true + that	4
6	with the development of	64	is more important than	4
6	with this opinion because (+ of)	65	it is argued that	4
6	we cannot deny the (+ importance of)	66	it is my belief	4
5	other believe that they	67	many people feel that	4
5	a huge amount of	68	not to say that	4
5	as a matter of + fact	69	of the opinion that	4
5	(play +) a significant role in	70	should be able to	4
5	there are a lot + of	71	some people argue that	4
5	as a result of	72	that it would be	4
5	as a result they	73	the fact that the	4
5	becoming more and more	74	the only way to	4
5	do not have	75	there has been a	4
5	do not have enough	76	this can only be	4

Appendix B *(Continue)*

Freq	SC	Rank	EC	Freq
5	do not want to	77	while I accept that	4
5	first of all it (+ is)	78		
5	have a chance to	79		
5	I agree that the	80		
5	I agree with the	81		
5	in fact there are	82		
5	in my opinion the	83		
5	is the most important	84		
5	it is difficult to	85		
5	many people believe that	86		
5	no one can deny	87		
5	people do not have	88		
5	people in the world	89		
5	reason is that the	90		
5	that they do not	91		
5	(with +) the rapid development of	92		
5	there are many reasons	93		
5	there are so many	94		
5	this point of view	95		
5	to the fact that	96		
5	us a lot of	97		
4	a good way for	98		
4	a lot of benefits	99		
4	agree with this view	100		
4	an important part of	101		
4	and as a result	102		
4	and give my opinion	103		
4	and they do not	104		
4	and you do not	105		
4	there are a number + of	106		
4	are one of the	107		
4	as well as the	108		
4	become more and more	109		

Appendix B *(Continue)*

Freq	SC	Rank	EC	Freq
4	but it does not	110		
4	but it is the	111		
4	do not have much	112		
4	in every corner of (+ the)	113		
4	for some following reasons	114		
4	however I believe that	115		
4	however I think that	116		
4	however we cannot deny	117		
4	I partially agree with + this	118		
4	I strongly agree that	119		
4	I totally agree with	120		
4	I totally disagree with	121		
4	if they do not	122		
4	in my opinion I	123		
4	it is the fact (+ that)	124		
4	it is necessary to	125		
4	it is not only	126		
4	it seems to me + that	127		
4	plays an important part	128		
4	thanks to the development (+ of)	129		
4	that you do not	130		
4	the development of the	131		
4	the first reason why	132		
4	there are many things	133		
4	there are several reasons	134		
4	to be able to	135		
4	to sum up it	136		
4	to the development of	137		
4	when you want to	138		
4	with the idea that	139		