

“To Whom It May Concern”: A Study on the Use of Lexical Bundles in Email Writing Tasks in an English Proficiency Test

Zhi Li & Alex Volkov

Lexical bundles are worthy of attention in both teaching and testing writing as they function as basic building blocks of discourse. This corpus-based study focuses on the rated writing responses to the email tasks in the Canadian English Language Proficiency Index Program® General test (CELP-IP-General) and explores the extent to which lexical bundles could help characterize the written responses of the test-takers of different English proficiency levels. Three subcorpora of email writing responses were created based on test-takers' proficiency levels. AntConc 3.4.4 was used to identify 2- to 6-word lexical bundles, which were then manually coded for their discourse functions. The results showed that test-takers of higher proficiency levels used more lexical bundles in terms of both bundle types and tokens compared with test-takers of lower proficiency level. The writing samples of different proficiency levels shared some lexical bundles, and overall they had similar proportional distributions of lexical bundle functions. Nevertheless, noticeable differences among the proficiency levels were observed in the proportional distributions of the subfunctions of stance bundles, discourse organizing bundles, and bundles of other functions. The identification of differential use of lexical bundles can contribute to a better understanding of English learners' email writing performance.

Les expressions figées méritent notre attention tant en enseignement qu'en évaluation de l'écrit puisqu'elles jouent le rôle de composantes de base du discours. Cette étude basée sur des corpus porte sur des réponses écrites par courriel à des tâches et évalue la mesure dans laquelle les expressions figées pourraient aider à caractériser les réponses écrites des participants de différentes compétences en anglais. Trois sous corpus de réponses écrites par courriel ont été créés en fonction des niveaux de compétence des participants. AntConc 3.4.4 a servi dans l'identification d'expressions figées composées de 2 à 6 mots qui ont ensuite été codées manuellement selon leurs fonctions discursives. Les résultats indiquent que les participants de niveaux de compétence plus élevés employaient plus d'expressions figées et de types plus variés comparativement aux participants de niveaux de compétence plus bas. Les échantillons de textes de différents niveaux de compétence partageaient quelques expressions figées et, globalement, avaient des distributions proportionnelles d'expressions figées similaires. Toutefois, des différences notables entre les niveaux de compétence ont été observées dans la distribution proportionnelle des sous fonctions discursives. L'identification des différences dans l'emploi des expressions figées peut contribuer à une meilleure

KEYWORDS: lexical bundles, email writing task, English proficiency test, discourse function, email structures

Email is a common means of communication in today's world, and email literacy has been recognized as important for effective communication in both workplace and schools (Chen, 2016; McKeown & Zhang, 2015). Email writing, like other forms of correspondence, can be highly conventional when used in formal contexts, while in informal contexts it may vary greatly in terms of style and language choice (Crystal, 2001; Gains, 1999; Lan, 2000). When used in an institutionalized way, email writing may be characterized by the use of frequently recurring multiword lexical units, or lexical bundles, and readers may also expect to encounter certain specific structures and language forms. Therefore, a good understanding of lexical bundles may facilitate English language learners' email writing and promote effective communication. This study focuses on the lexical bundles extracted from the email writing tasks on a high-stakes English proficiency test and examines the discourse functions of the bundles across three proficiency levels.

Literature Review

Emails as a Communicative Medium

Emails are regarded as a variety of language with relatively fixed discourse elements fitting into the composing spaces in email programs or apps (Crystal, 2001). These elements can be obligatory, such as the body of the message and the sender and recipient(s) in the header, or they can be optional, such as subject line, greetings, and complimentary closing. Depending on the purpose of communication and situational factors, the language of emailing varies greatly (Gains, 1999) and the writing styles are extremely idiosyncratic (Baron, 1998), which makes it difficult for email writing to be considered and defined as a unified genre.

Despite the small number of studies on the linguistic features of email writing as a whole, the pragmatic features associated with the opening and closing lines have attracted much attention, especially in regards to cross-cultural communication and workspace communication (Bjørge, 2007). The work by McKeown and Zhang (2015) is one of the recent studies on the relationship between a number of situational factors and the choice of opening and closing in British workplace emails. Using a quantitative approach to modelling these relationships, McKeown and Zhang were able to pinpoint the influential factors from a multitude of variables. For example, it was found that formality

in the openings was enhanced due to factors such as external communication, social distance between parties, gender of the senders, and so on.

In a similar vein, efforts have been made to understand English language learners' email writing. A common theme from these studies is that English language learners (ELLs) generally lack the adequate pragmatic competence and appropriate linguistic devices to make proper request acts through email writing to their teachers (Biesenbach-Lucas, 2007; Chen, 2016). In these studies, requestive strategies are usually coded as conventional directness (with imperatives, performatives, want statements, and expectation statements), conventional indirectness (query preparations), or nonconventional indirectness (strong hints). Along with these strategies, a number of syntactic and lexical devices are identified in email writing to mitigate the imposition level of requests, such as if/whether clauses and downtoner phrases. Zhu (2012) compared the upward request acts in the emails written by Chinese-speaking students with an English major and those of non-English majors in light of three situational factors: social distance, power, and rank of imposition. Overall, non-English-major students used more direct strategies, and their requests tended to appear less appropriate than their English-major peers.

These studies on email writing reveal the influential factors on email styles and formality. However, they rarely touched the frequently recurring lexical units and discourse functions related to the body of the message. This study intended to treat email writing as a whole and to analyze the discourse functions of these lexical units. Furthermore, it is noteworthy that the email writing samples used in this study were elicited by a set of tasks in a testing context. Test-takers are generally aware of the nature of the task, which has explicit information about the topic, target audience, and evaluative criteria. Therefore, the email samples from this study may resemble professional discourse in formal correspondence in terms of linguistic features and conventionality, and may appear dissimilar from personal discourse found in casual emails (McKeown & Zhang, 2015).

Features of Lexical Bundles

The term "lexical bundle" was coined in *The Longman Grammar of Spoken and Written English* edited by Biber, Johansson, Leech, Conrad, and Finegan (1999), when the authors took an inductive approach to identifying and analyzing corpus-driven outcomes of recurring continuous word sequences. Lexical bundles are defined as frequently occurring multiword lexical units (Biber & Barbieri, 2007). Lexical bundles, as artifacts of frequency-based queries, are usually not idiomatic in meaning. Another feature of lexical bundles is their usual incompleteness in structure as they tend to appear as fragments from a larger grammatical structure or even as crossover bundles bridging two phrasal or clausal units (Biber, 2009).

Because lexical bundles exist as parts of larger units, they can fulfill certain discourse functions. As Biber (2009) put it, lexical bundles provide "a kind

of pragmatic ‘head’ for larger phrases and clauses” or “interpretive frames for the developing discourse” (p. 285). For this reason, lexical bundles have been regarded as “building blocks” of discourse in both spoken and written registers (Biber, Conrad, & Cortes, 2004).

According to Cortes (2004), three types of discourse functions are usually realized by lexical bundles: stance expression, referential expression, and discourse organizer. The same classification of discourse functions is used in a number of studies on lexical bundles (Biber et al., 2004; Chen & Baker, 2016; Hyland, 2008). Stance expression bundles deal with personal or impersonal epistemic stance such as *I don't know*, as well as attitudinal or modality stance such as *I would like to* and *I am unable to*. Referential expression bundles are used to refer to time or place such as *at the same time*, or to specify certain attributes of an object such as *the time to read*. Discourse organizer bundles are used to introduce topics or focus in an email response, for example, *I am writing to, to let you know*. Discourse organizing bundles can be further classified into two general categories depending on their relation to the neighboring clausal units: introducing or focusing on a topic and elaborating or clarifying a topic (Biber et al., 2004; Chen & Baker, 2016). In addition to these three types of functions, Conrad and Biber (2005) identified another group of lexical bundles that serve special conversational functions such as expressing politeness, making simple inquiry, and reporting. In our study, some of the lexical bundles appear to be common in email writing, for example, the bundles expressing politeness such as *thank you very much*, the opening phrases such as *dear sir or madam*, and the closing phrases such as *kind regards*. These bundles were collectively labelled as “other functions” in this study.

Lexical bundles of various lengths have been studied. However, 4-word lexical bundles are arguably the most widely studied. This is usually because the number of extracted 4-word lexical bundles is more manageable than 3-word bundles and, in some cases, 4-word bundles tend to include a certain number of 3-word bundles (Cortes, 2013). Also, 4-word bundles are found to be more common than 5-word bundles (Hyland, 2008). In practice, two distributional criteria are used to identify lexical bundles, namely frequency of occurrence and coverage or range of occurrence (Gray, 2016). They are applied to ensure that the identified lexical bundles are representative and not overly idiosyncratic in a given corpus. There are no agreed cut-off values for these two criteria as of yet and, consequently, a variety of values have been used in previous studies. The cut-off value for frequency may range from 10 to 40 times per million words, depending on the corpus size as well as the length of the lexical bundles (Cortes, 2013). In Biber and Barbieri (2007), for example, the threshold of 40 occurrences per million words is deemed conservative. Similar concerns about frequency in a relatively small corpus have been expressed by Gray (2016). Gray (2016) dealt with subcorpora of about 100,000 words each and decided to choose a more conservative approach (10 occurrences in at least five different texts in the corpus) in order to avoid

overidentifying lexical bundles. With regard to the criterion of coverage, some studies specified particular values for their corpora, such as three to five texts (Ädel & Erman, 2012; Biber & Barbieri, 2007; Cortes, 2013), while others use percentages, such as at least 5% or 10% of all texts (Hyland, 2008; Pan, Reppen, & Biber, 2016). For example, the word sequences in Cortes (2013) need to be used in five or more texts (out of 1,372 texts) to be identified as lexical bundles. Hyland (2008), on the other hand, used at least 10% of texts in each subcorpus as the threshold of coverage, which yielded an equivalent of at least five texts in that corpus. Another approach to setting cut-off values is to use dynamic values in the case where corpora of different sizes are compared, as shown in Chen and Baker (2016). Considering the differences in size of the three subcorpora, Chen and Baker used dynamic thresholds for both frequency and range to ensure the extracted lexical bundles from the subcorpora were representative and comparable. For example, the smaller corpus had lower thresholds (three or more occurrences in at least three different essays for the smaller corpus vs. four or more occurrences in at least three different essays for the bigger corpus).

Studies on Lexical Bundles

Previous studies on the use of lexical bundles by English learners of various proficiency levels suggest that lexical bundles can be indicative of proficiency level as they are used differentially by expert and novice second language (L2) writers. For example, Chen and Baker (2016) studied 4-word lexical bundles as potential criterial discourse features in a corpus of 585 expository or argumentative essays written by Chinese learners of English as collected in the Longman Learner Corpus (LLC). Three subcorpora were constructed to represent three proficiency levels in accordance with the Common European Framework of Reference for Languages (CEFR), namely B2, B1, and C1. Chen and Baker found that the lexical bundles did exhibit different features in the subcategories of both structures and functions while they shared similar distributional patterns as a whole. Their in-depth linguistic analyses indicated that lower-proficiency-level writers employed more oral language-like lexical bundles while the lexical bundles used by the higher-proficiency-level writers appeared more academic in style. In addition, they noticed that some lexical bundles in the subcorpus of lower-proficiency-level essays were not appropriately used. Chen and Baker (2016) maintained that their findings about lexical bundles showed some distinctive features in terms of formulaicity and stylistic features of the essays across proficiency levels and claimed that their findings could help validate and refine the CEFR descriptors of writing proficiency.

Studies on lexical bundles have been used in the area of language testing research. For example, in a corpus-driven study of lexical bundles in TOEFL iBT writing tasks, Staples, Egber, Biber, and McClair (2013) found that test-takers of different proficiency levels showed similar uses of lexical bundles

in terms of functions and degree of fixedness despite the fact that lower-proficiency writers used more lexical bundles including some taken from the writing prompts. Appel and Wood (2016) studied the use of recurring word combinations of 4- to 7-word sequences by lower-proficiency-level writers and higher-level writers in the Canadian Academic English Language (CAEL) Assessment. They compiled two subcorpora of the CAEL Assessment argumentative writing samples from non-native English-speaking test-takers and compared the functional types of the 4–7 word sequences in these corpora, namely, stance, discourse-organizing, and referential. Appel and Wood found that there were larger percentages of stance expressions as well as discourse-organizing expressions for lower-level writers than for the higher-level writers, while higher-level writers tended to use more referential expressions. In addition, Appel and Wood reported that lower-level writers used more expressions borrowed from the source materials in the test than higher-level writers did.

Research Question

The reviewed studies have much to offer in revealing the relationship between uses of lexical bundles and the situational or learner factors. However, the majority of the studies on lexical bundles focused on academic discourses while fewer efforts were devoted to English for general purposes—in our case, email writing. To address this gap, this study investigated the lexical bundles used by test-takers of different proficiency levels on the email writing tasks on a general English proficiency test. Specifically, we aimed to answer the following research question.

Do test-takers of different writing proficiency levels use lexical bundles differently in terms of discourse functions?

Method

The Email Writing Task

The test of interest is called the Canadian English Language Proficiency Index Program-General or the CELPIP-General test, which is developed and administered by Paragon Testing Enterprises in Canada. The CELPIP-General test is a standardized and computer-delivered English proficiency test that measures language performance in four modalities: reading, listening, speaking, and writing. The CELPIP scores are mainly used as proof of English-language proficiency by applicants for permanent residency or citizenship in Canada.

Two types of writing tasks are currently used in the CELPIP-General test: Writing an email and Responding to survey questions. This study focuses on the first writing task, as email writing is one of the most common writ-

ing tasks in daily life. This task requires a test-taker to write an email of 150–200 words to address day-to-day matters in 27 minutes (see Appendix for a sample task). The prompt consists of a short description of a scenario and three subtasks that the email would be expected to fulfill (Paragon Testing Enterprises, 2015). Trained raters use analytical rating scales to evaluate test-taker performance using four dimensions: coherence/meaning, lexical range, readability/comprehensibility, and task fulfillment. The rating scores are reported using CELPIP levels from Minimal to 12, which are calibrated against the Canadian Language Benchmarks (CLB).

The Corpus of CELPIP Email Writing Responses

After we retrieved essays from the CELPIP database, we grouped the essays by proficiency levels and then selected the most recent 2,500 essays in each group to build a balanced corpus of email writing responses at the CELPIP levels 4, 7, and 10, which correspond to three broad stages of the CLB proficiency levels (Stages I, II, and III). A summary of the three subcorpora is presented in Table 1. The total count of running words is 1,357,911, with a total of 27,117 unique words or type. The average length of the email responses varies from 157 words per text at Level 4 to 193 words per essay at Levels 7 and 10. It is worth noting that we did not control for writing prompts and test-takers’ demographic characteristics such as gender and first language in the compilation of this corpus.

Table 1
Summary of the CELPIP Email Writing Corpus

Proficiency level	Number of texts	Number of words (tokens)	Number of word type	Average length
CELPIP 4	2,500	392,625	13,505	157
CELPIP 7	2,500	482,605	13,773	193
CELPIP 10	2,500	482,681	14,989	193
Total	7,500	1,357,911	27,117	181

Analytical Tool

Lexical bundles were identified using AntConc 3.4.4 (Anthony, 2014) using its Clusters/N-Grams function. Due to the fact that our corpus contains a large number of short email responses, which is different from the corpora used in other studies, it is challenging to determine the optimal criteria of frequency and range based on the literature. In short email writing, a lexical bundle is less likely to appear multiple times in the same email response. As a result, a frequency criterion such as 20 occurrences per million words may also establish the threshold value for range. In our study, the frequency cut-off value is able to guarantee an appropriate dispersion of the writing samples and ad-

dresses the concern of idiosyncrasy. Consequently, a cut-off value for range in our case becomes redundant and we decided to drop the criterion of range in our study.

In this study, we employed a criterion of 40 occurrences per million words for frequency. This criterion is regarded as relative conservative (Biber & Barbieri, 2007) and such a cut-off value can help prevent overidentifying lexical bundles. In addition, the lists of lexical bundles extracted with this criterion were proven to be manageable, compared with a lower cut-off value (e.g., 20 occurrences per million words), while it helped capture some unique bundles that were of medium frequency of occurrence used by lower- or higher-proficiency-level test-takers only. Considering the differences in the size of the subcorpora, we followed the practice in Biber and Barbieri (2007) to normalize the required frequencies and convert them into 16 occurrences for the subcorpus of proficiency Level 4 ($40/1,000,000 \times 392,625 = 15.7$) and 20 occurrences for the subcorpora of proficiency Levels 7 and 10 ($40/1,000,000 \times 482,681 = 19.3$).

Procedures

Once the 4-word lexical bundles were extracted with AntConc 3.4.4, several steps were taken to clean the bundle list. The first step involved identifying and removing prompt-specific bundles, such as *dear Mr. Smith I* and *my daughter's birthday*. This is because these bundles are not likely to be used in general email writing. The prompt-specific bundles were identified through a manual check for the overlaps between bundle components and the content words that were unique in the CELPIP writing prompts. Second, some overlapping lexical bundles were identified based on their shared elements as in *to whom it may* and *whom it may concern*. Another example is the lexical bundles containing elements from two adjacent sentences such as *from you kind regards* as taken from two independent structures *I look forward to hearing from you* and *Kind regards*. We decided to adjust the length of the lexical bundle to reflect this formulaicity and reran the analysis for lexical bundles of varied length from 2- to 6-word bundles. The same frequency criterion (40 occurrences per million words) was used in the search for bundles of different lengths. However, both bundle frequency and structural completeness were considered in determining new bundles. As a result, fewer new lexical bundles of different lengths were added to the list, including 2-word bundles (e.g., *kind regards* and *best regards*), 3-word bundles (e.g., *my name is* and *be able to*), 5-word bundles (e.g., *to whom it may concern* and *I am looking forward to*), and 6-word bundles (e.g., *hope this email finds you well*). Lastly, we combined the lexical bundles with and without contraction as in *I'm going to* and *I am going to* in order to avoid inflating types of bundles.

The lexical bundles were then manually labelled for their discourse functions. In addition to the three primary discourse functions of the lexical bundles — stance expression, referential expression, and discourse organizer — we

labelled the bundles of other functions. Both researchers independently labelled lexical bundles for their discourse function after a brief familiarization and calibration session. The intercoder agreement for labelling discourse function varied from 85% to 93% across the three subcorpora of proficiency levels, making an average agreement of 89%. Disagreement was resolved through further discussions and in-depth analyses of the concordance lines.

Results and Discussions

Overview of the Lexical Bundles

A summary of the lexical bundles extracted from the email writing task is presented in Table 2. The corpora of higher proficiency levels (CELPiP Levels 7 and 10) yielded more lexical bundles in terms of both type and token, compared with the corpus of the lower proficiency level (CELPiP 4), while the differences between these two higher proficiency levels were less salient (see the second column of Table 2). For example, the total number of token of lexical bundles used in the subcorpus of CELPiP Level 4, the lower proficiency level, was 3,901, which is about 60% of the lexical bundle tokens found in the subcorpora of the higher proficiency levels. Comparisons of the normalized tokens against the corpus size still suggest that, as a whole, test-takers of higher proficiency levels used more lexical bundle tokens. Considering the differences in the size of these corpora in terms of running words, the differences in the number of lexical bundles are understandable because shorter responses tend to employ fewer lexical bundles.

Table 2
Summary of Lexical Bundles Across Three Proficiency Levels

<i>Proficiency level</i>	<i>Lexical bundle types</i>	<i>Lexical bundle tokens</i>	<i>Number of words in corpus</i>	<i>Normalized lexical bundle tokens (per 1,000 words)</i>
CELPiP 4	81	3,901	392,625	9.94
CELPiP 7	98	6,566	482,605	13.61
CELPiP 10	96	6,666	482,681	13.81

The patterns of lexical bundles revealed in Table 2 are similar to the findings in the other studies on lexical bundles in that lower-proficiency-level learners or non-native English speakers were more likely to use a narrower range of lexical bundles while higher-proficiency-level learners and native English speakers had more types of lexical bundles at their disposal (Ädel & Erman, 2012; Appel & Wood, 2016).

The top 40 lexical bundles from the three subcorpora and their frequency information are listed in Table 3. Eyeballing the table reveals that 14 bundles, or 35% of the top 40 bundles, are shared across the three proficiency levels although the frequency of occurrences differed. These bundles appear to form

a bare-bones structure of email writing, covering from the salutation to the closing, as shown by *to whom it may concern, dear sir or madam, I am writing this, I would like to, to let you know, so that I can, as soon as possible, I hope you will, be able to, thank you very much, and best regards*. It is noteworthy that some lexical bundles only differ in one slot, such as *I would like/love to, I/we would like to, to hear/hearing from you*, suggesting some phrase frames or concgrams (bundles with variable or fixed slots) may be appropriate units to capture these variations (Biber, 2009).

Table 3
Top 40 Lexical Bundles from the Subcorpora of Three Proficiency Levels

<i>CELP 4 (Frequency)</i>	<i>CELP 7 (Frequency)</i>	<i>CELP 10 (Frequency)</i>
I would like to (430)	I would like to (783)	I would like to (741)
my name is (309)	my name is (341)	I am writing to (441)
dear sir madam (275)	kind regards (324)	be able to (345)
good day (160)	I am writing this (321)	kind regards (324)
dear sir I am (124)	be able to (243)	to whom it may concern (318)
thank you very much (104)	best regards (236)	my name is (316)
as soon as possible (103)	I am writing to (233)	I look forward to (280)
thank you so much (94)	dear sir madam I (219)	best regards (236)
I am writing to (89)	to whom it may concern (197)	thank you for your (222)
I am going to (88)	as soon as possible (192)	please let me know (221)
best regards (88)	thank you for your (122)	to hearing from you (186)
I am writing this (86)	to hear from you (117)	I am writing this (136)
<i>how are you</i> (72)	thank you very much (111)	as soon as possible (127)
<i>I don't have</i> (71)	I am looking forward (103)	if you have any (90)
to whom it may concern (63)	to inform you that (99)	dear sir or madam (59)
be able to (61)	to hearing from you (91)	thank you in advance (59)
I hope you will (60)	to let you know (86)	<i>and I have been</i> (58)
<i>I do not know</i> (60)	I look forward to (85)	to bring to your (56)
<i>and I want to</i> (59)	dear sir I am (83)	I would like you (56)
<i>have a good day</i> (56)	please let me know (81)	to inform you that (54)
I just want to (55)	I hope you will (81)	<i>would it be possible</i> (53)
thank you for your (52)	I am writing you (76)	to hear from you (51)
to let you know (51)	dear sir or madam (72)	to let you know (50)
<i>I hope you understand</i> (51)	I would like you (70)	I would really appreciate (49)

continued on next page

<i>CELPIP 4 (Frequency)</i>	<i>CELPIP 7 (Frequency)</i>	<i>CELPIP 10 (Frequency)</i>
<i>first of all I (47)</i>	first of all I (60)	<i>please feel free to (48)</i>
<i>because I have a (46)</i>	<i>at the same time (58)</i>	<i>appreciate if you could (48)</i>
<i>because I want to (43)</i>	thank you so much (54)	<i>I have noticed that (47)</i>
dear sir or madam (40)	thank you in advance (53)	<i>do not hesitate to (46)</i>
I hope you can (40)	I hope you can (51)	it would be a (45)
<i>have a lot of (39)</i>	I would really appreciate (50)	<i>and would like to (45)</i>
I am writing you (33)	<i>for your kind consideration (47)</i>	<i>I am unable to (45)</i>
<i>is very important to (33)</i>	<i>I am planning to (47)</i>	<i>at your earliest convenience (44)</i>
so that I can (33)	to inform you about (43)	<i>the end of the (44)</i>
<i>have a nice day (31)</i>	if you have any (43)	<i>in regards to the (42)</i>
<i>a lot of people (28)</i>	<i>the reason why I (41)</i>	<i>I would love to (41)</i>
I hope that you (28)	<i>I am writing in (40)</i>	thank you very much (39)
I am looking forward (28)	so that I can (39)	I hope you will (39)
and I hope you (27)	<i>we would like to (39)</i>	<i>I would also like (38)</i>
<i>don't have any (26)</i>	to bring to your (37)	<i>to inform you of (37)</i>
<i>if you don't (26)</i>	I just want to (36)	so that I can (35)

Note. The lexical bundles shared across the three proficiency levels are in boldface and the unique bundles at each level are in italics.

Unique bundles were found at each proficiency level. Excluding the lexical bundles that appeared in two or three of the lists, we identified 14 unique bundles each at CELPIP Levels 4 and 10, and 6 at CELPIP Level 7. This pattern matched our expectation of CELPIP Level 7, as it is the midground between the lower proficiency level and the higher one, thus featuring more overlapping lexical bundles with adjacent levels.

A closer look at the unique bundles used at CELPIP Levels 4 and 10 groups suggests some differences in their writing, as the unique bundles at CELPIP Level 10 seem to be more polite and formal as shown in *do not hesitate to*, *at your earliest convenience*, *I would greatly appreciate*, whereas the ones in CELPIP Level 4 appear to be more casual as in *how are you*, *have a nice day*, *if you don't*, and *because I want to*. This observation is roughly in line with what Biesenbach-Lucas (2007) found in her comparative study of students' email communication with faculty members made by native and non-native English speakers. That is, native English-speaking students were more polite in making their requests than non-native English-speaking students. More discussions about politeness in email writing are presented in the subsection discussing the bundles of other functions.

The Function Features of Lexical Bundles

Table 4 describes the frequencies of occurrence and percentage of the lexical bundles in each of the discourse function categories as well as across the three proficiency levels. Overall, a similar distributional characteristic was found in the lexical bundles across the proficiency levels. The bundles of stance, discourse organizer, and other functions made up more than 90% of the total occurrences at each proficiency level, while the occurrences of referential bundles were much less frequent.

Table 4
Frequency of Occurrence and Percentage of Lexical Bundles of Different Discourse Functions

<i>Proficiency level</i>	<i>Stance</i>	<i>Referential</i>	<i>Discourse organizer</i>	<i>Other</i>
CELPIP 4	1,247 (32%)	229 (6%)	1,003 (26%)	1,422 (36%)
CELPIP 7	2,050 (31%)	513 (8%)	1,695 (26%)	2,308 (35%)
CELPIP 10	2,053 (31%)	507 (8%)	1,887 (28%)	2,219 (33%)

The percentages of the bundle function types are similar for the four types as well (see Table 4 and Figure 1). The stance bundles showed almost the same percentages across the three proficiency levels (CELPIP Level 4: 32%,

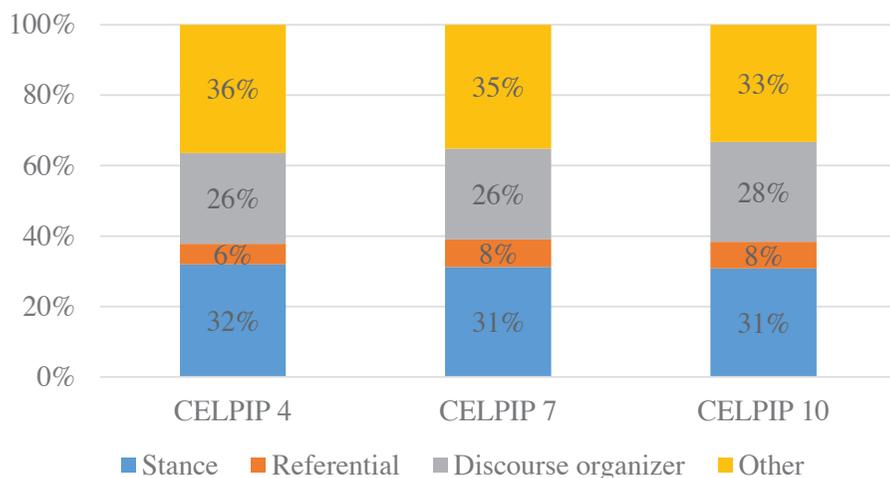


Figure 1: Distribution of lexical bundle types across proficiency levels (percentage of tokens)

CELPiP Level 7: 31%, and CELPiP Level 10: 31%). With regard to the referential bundles, CELPiP Levels 7 and 10 shared the same percentage (8%), which is slightly higher than the counterpart for CELPiP Level 4 (6%). The variation of the percentages of discourse organizing bundles appears to be small, too (26% vs. 28% at CELPiP Level 10). A slight decreasing trend was observed in the bundles of other functions, with lower-proficiency-level writing samples containing a relatively larger percentage (CELPiP Level 4: 36% vs. CELPiP Level 10: 33%).

The similarity of distributional patterns among the different proficiency levels was also observed in other studies comparing bundles used at different proficiency levels (Ädel & Erman, 2012; Chen & Baker, 2016; Staples et al., 2013). However, the specific proportions of bundle functions in our study appear to be rather different from the findings in other studies that shared a focus of bundle functions. For example, Ädel and Erman (2012) reported a large proportion of referential bundles (45–47%) in their analysis of academic writing samples from L1 Swedish writers and native English speakers, as opposed to the remarkably smaller proportions (6–8%) found in our study. Chen and Baker (2016), on the other hand, identified about 40% of the bundles serving as discourse organizers and 20% as referential bundles in their study of L1 Chinese learners of English, while Staples et al. (2013) found that more than 50% of the bundles were discourse organizers and fewer than 10% were referential bundles in a corpus of graded TOEFL iBT writing samples. These distributional differences may be attributed partially to the natures of the writing tasks used in the studies. The language samples elicited by the email writing task in the CELPiP-General test are likely very different from the academic writing as analyzed in Ädel and Erman (2012) or the exam essays collected in Chen and Baker (2016) and Staples et al. (2013). The different proportional features of the function types suggest that formal email writing may be a special genre constrained with highly formulaic language and established conventions (Crystal, 2001).

Stance bundles

Stance bundles are used to express epistemic stance or certainty, desire, intention, obligation/directive, or ability (Biber et al., 2004). Stance bundles exhibited some variations in the proportion of the different subfunctions of stance across proficiency levels (see Figure 2). For example, the bundles expressing desire or intention constituted 62% of the stance bundle tokens at CELPiP Level 4, while they made up 52% and 46% at CELPiP Levels 7 and 10, respectively. Likewise, CELPiP Level 4 had a relatively larger proportion of the stance bundles about obligation or directive (29%), compared with CELPiP Levels 7 (25%) and 10 (25%). In other words, the CELPiP Level 4 writers used a higher percentage of stance bundles expressing their own desire or intention as well as obligations or directive to others than did more proficient writers.

Another salient feature in the use of stance bundles was that the ones used by lower-level writers tended to be more direct and informal, while the ones employed by the higher-level writers, especially the CELPIP Level 10 writers, seemed more formal and polite, as shown in the following selected examples with corresponding frequency information.

CELPIP Level 4: Desire *I would like to* (430), and *I want to* (59), *I just want to* (55); Ability *be able to* (61); Obligation/Directive *I hope you will* (60), *I would like you* (26), *I need your help* (21); Certainty *I don't know* (60).

CELPIP Level 7: Desire *I would like to* (783), *I just want to* (36), *I am hoping for* (32); Ability *be able to* (243); Obligation/Directive *I hope you will* (81), *I would like you* (70), *I would really appreciate* (50); Certainty *I am sure that* (31)

CELPIP Level 10: Desire *I would like to* (741), *I would love to* (41); Ability *be able to* (345), Obligation/Directive *I would like you* (56), *I would really appreciate* (49), *do not hesitate to* (46), *I hope you will* (39); Certainty *I am sure that* (27)

This observation is somewhat in line with findings from the studies on requestive strategies in email writing (Leopold, 2015; Zhu, 2012), which demonstrated that less-proficient English learners tended to use more direct requests and fewer mitigation devices. As shown in the examples above, nearly all the bundles at CELPIP Level 4 are either need-statements or want-statements, while the examples from CELPIP Levels 7 and 10 have slightly more expectation statements (*I hope you will* or *I am hoping for*).

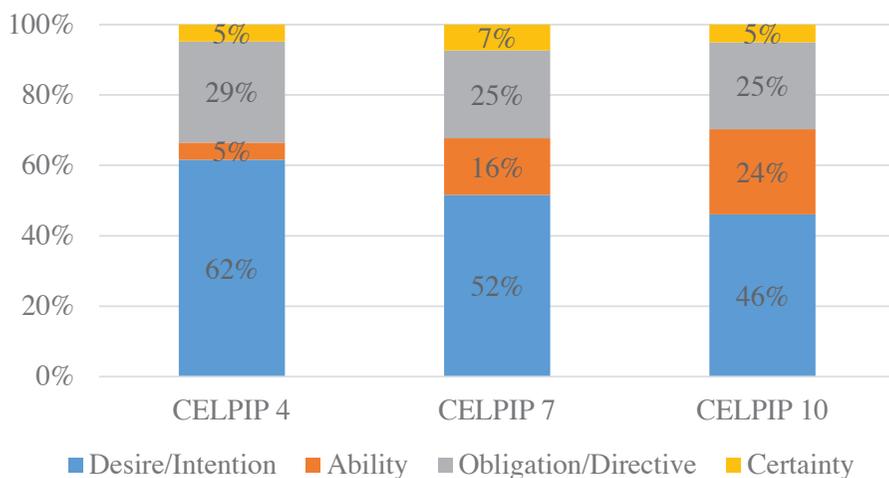


Figure 2: Distribution of stance bundles across proficiency levels (percentage of tokens)

Referential bundles

As described earlier, referential bundles only composed a small part of the extracted bundles at all three proficiency levels (6–8%). These bundles can be further analyzed for their subfunctions, that is, reference to time, place, or textual information, framing an entity, and quantifying an entity.

As shown in Figure 3, the writing samples at CELPIP Levels 7 and 10 looked alike in terms of the configuration of referential bundle subfunctions. That is, the majority of the referential bundles were used to refer to time, place, or textual information (82% in CELPIP Level 7 and 80% in CELPIP Level 10), and only a fraction or none of the bundles were used to render quantifying information (4% in CELPIP 7 and 0% in CELPIP 10). On the other hand, the majority of referential bundles at CELPIP Level 4 were quantifying bundles (70%), and in actuality all the quantifying bundles were similar in structure, for example, *there's a lot, have a lot of, a lot of people*. It is also noteworthy that the responses at CELPIP Levels 4 and 7 used the same percentage of framing bundles (13%), while the responses at CELPIP Level 10 contained a relatively larger proportion of the framing bundles (19%).

The proportional distributions of referential bundles suggest that the time/place/text reference bundles and framing bundles used at CELPIP Levels 7 and 10 helped package more details and attributes of the entity of interest to the email messages. These patterns are similar to the findings in Chen and Baker (2016) regarding the use of these subfunctions by writers at three different proficiency levels. In addition, some of the quantifying bundles identified in our analysis are also found in the writing of lower-proficiency-

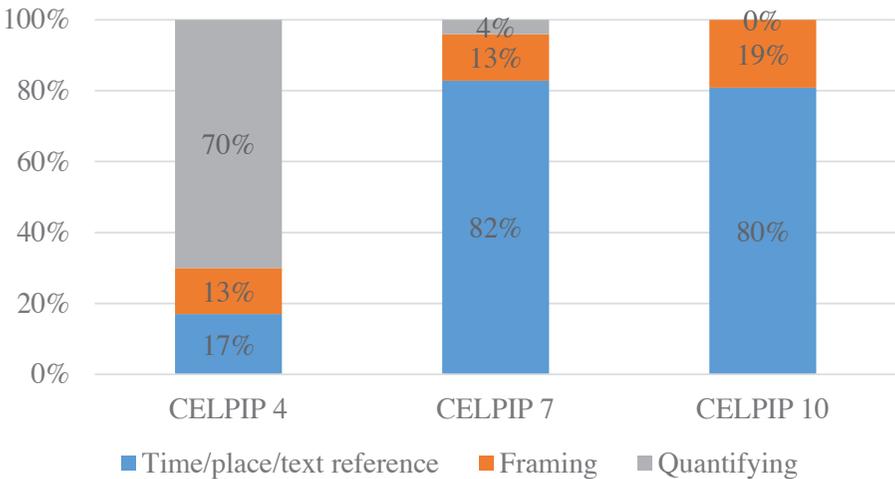


Figure 3: Distribution of referential bundles across proficiency levels (percentage of tokens)

level writers (B2), and they are typically used in oral communication. Some selected examples of referential bundles are listed below.

CELPPIP Level 4: Time/place/text reference as soon as possible (103), at the same time (21); Framing is one of the (17); Quantifying have a lot of (39),

CELPPIP Level 7: Time/place/text reference as soon as possible (192), at the same time (58); Framing the reason why I (41); Quantifying a lot of people (22)

CELPPIP Level 10: Time/place/text reference as soon as possible (127), the end of the (44); Framing in regards to the (42), as a result of (34)

Discourse organizing bundles

An analysis of the subfunctions of the discourse organizing bundles revealed all three CELPIP levels shared a similar distributional pattern of the discourse organizing bundles, while the differences in the overall proportion of the bundles in the corpus were also similar (26–28%, as shown in Table 1). Specifically, about three quarters of the discourse organizing bundles were used to introduce topics (73–78%), while only a quarter were devoted to elaborating or clarifying topics (22–27%) (see Figure 4).

Some examples of discourse organizing bundles are listed below. It seems that all three proficiency levels featured the discourse organizing bundles that introduce writers' purpose of email-writing as with *I am writing to* and *to inform you of*. We speculate that the similarity of the proportional distribution of these two subfunctions may be related to the length of the writing samples.

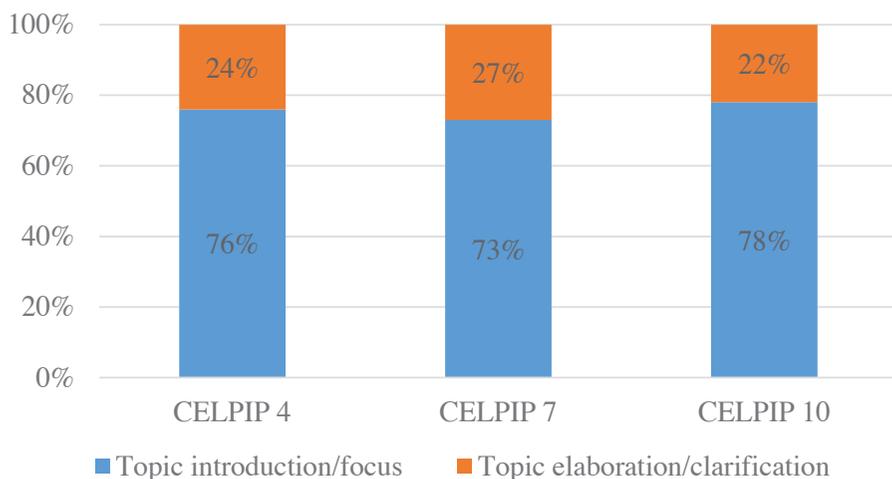


Figure 4: Distribution of discourse organizing bundles across proficiency levels (percentage of tokens)

CELPPIP Level 4: Topic introduction/focus *I am writing to* (89), *I am writing this* (86), *I don't have* (71), *because I have a* (46); Topic elaboration/clarification *to let you know* (51), *first of all I* (47), *is very important to* (33), *don't have a* (21)

CELPPIP Level 7: Topic introduction/focus *I am writing this* (321), *I am writing to* (233), *please let me know* (81), *I am writing you* (76); Topic elaboration/clarification *to inform you that* (99), *to let you know* (86), *first of all I* (60), *to inform you about* (43)

CELPPIP Level 10: Topic introduction/focus *I am writing to* (441), *I am writing this* (136), *if you have any* (90); Topic elaboration/clarification *to bring to your* (56), *to inform you that* (54), *to let you know* (50), *to inform you of* (37)

Bundles with other functions

In this study, about one third (33–36%) of the extracted bundles were labeled as other functions. These bundles included expressions of politeness as well as certain elements that are unique to email writing and other types of correspondence forms such as salutation, greetings, self-identification in the opening section, and complimentary closes in the closing section.

Figure 5 shows the proportional distributions of bundles with other functions in three different proficiency levels. It appears that the CELPIP Level 4 writing contained the largest proportion of bundles for opening purposes (61%) while having the smallest proportions of bundles for closing (17%) and expressing politeness (22%). The CELPIP Level 7 writing had a smaller proportion of bundles for opening purposes (37%) but used more bundles for closing (42%) and politeness (20%). The CELPIP Level 10 writing had a smaller proportion of bundles for opening purposes (32%) but used more bundles for closing (50%) and politeness (18%).

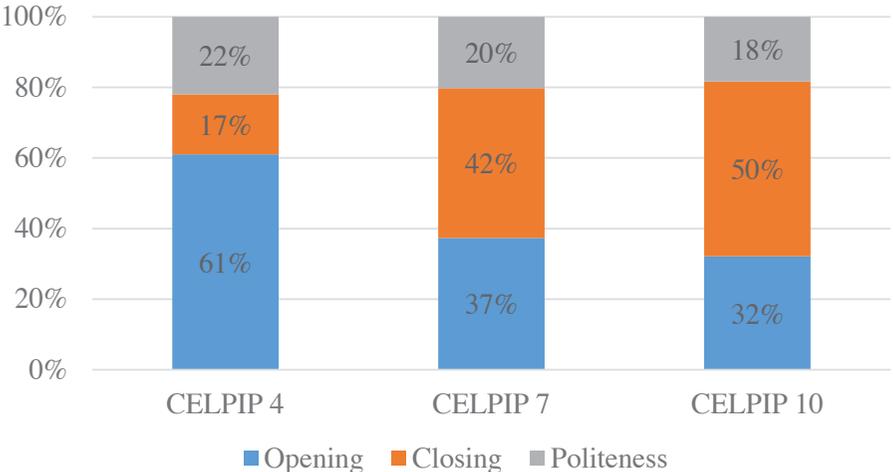


Figure 5: Distribution of the bundles with other functions across proficiency levels (percentage of tokens)

closing (42%) and a similar proportion for politeness (20%), compared with the CELPIP Level 4 writing. Half of the bundles with other functions used in CELPIP Level 10 writing were related to the closing part of email writing (50%) and the rest were divided between politeness expressions (18%) and opening expressions (32%).

Some examples of the bundles with other functions are listed below. It is obvious that some of the bundles used in the CELPIP Level 4 writing are absent in the bundles from the higher-level writing. For example, *how are you* as a part of the opening section of emails was used 72 times in CELPIP Level 4. This bundle is highly colloquial and does not appear as a bundle in the samples of higher proficiency levels. Similar cases include *good day* in the opening and *have a good/nice/great day* in the closing part in CELPIP Level 4. On the other hand, the email responses written at CELPIP Levels 7 and 10 used more formal expressions such as *to whom it may concern* (63 in CELPIP Level 4 vs. 197 in CELPIP Level 7 and 318 in CELPIP Level 10) and *best regards* (88 in CELPIP Level 4 vs. 236 in CELPIP Level 10). In addition, there were more variations in the use of politeness-expressing bundles employed by higher-proficiency-level writers in terms of bundle type (6 in CELPIP Level 4 vs. 9 and 11 in CELPIP Levels 7 and 10, respectively).

These proportional differences in the subfunctions of the bundles of other functions may be explained using the continuum of formality in relation to second language proficiency. The findings from Chen and Baker (2016) suggest that the writing samples from lower-proficiency-level writers tended to exhibit more features of oral language and thus appear more informal. In our study, the lexical bundles used by CELPIP Level 4 writers are perceived in a similar way. Some examples of bundles of other functions are listed below.

CELPIP Level 4: **Opening** *my name is* (309), *dear sir/madam* (275), *good day* (160), *how are you* (72), *to whom it may concern* (63); **Closing** *best regards* (88), *have a good day* (56), *I am looking forward* (28), *best regard* (22); **Politeness** *thank you very much* (104), *thank you so much* (94), *thank you for your* (52)

CELPIP Level 7: **Opening** *my name is* (341), *dear sir/madam* (219), *to whom it may concern* (197), *dear sir or madam* (72); **Closing** *kind regards* (324), *best regards* (236), *to hear from you* (117), *to hearing from you* (91); **Politeness** *thank you for your* (122), *thank you very much* (111), *thank you in advance* (53),

CELPIP Level 10: **Opening** *to whom it may concern* (318), *my name is* (316), *dear sir or madam* (59), *hope this email finds you well* (20); **Closing** *kind regards* (324), *best regards* (236), *I look forward to* (280), *to hearing from you* (186); **Politeness** *thank you for your* (222), *thank you in advance* (59)

The analysis of bundles of other functions leads to an observation of grammatical correctness of the bundles. Some bundles used in CELPIP Levels 4 and 7 are problematic. For example, in CELPIP Level 4, 22 occurrences of *best regard* appeared, instead of *best regards*. Another common error is [I

look forward] *to hear from you*, which should be written as *to hearing from you*. CELPIP Level 7 has more erroneous cases of that bundle (117) than the correct one (91), while CELPIP Level 10 showed a dominating use of the correct form (186 correct uses vs. 51 errors).

Conclusions and Implications

This study investigated the discourse functions of lexical bundles used by test-takers of different English proficiency levels in email writing tasks as a part of a high-stakes English proficiency test. The use of lexical bundles varied across the three proficiency levels with CELPIP Levels 7 and 10 using more lexical bundles than CELPIP Level 4, in terms of both lexical bundle types and normalized counts of tokens. It was also observed that there were different numbers of unique bundles among the top 40 lexical bundles at the three proficiency levels while about one third of the top 40 were shared across the proficiency levels. Overall, the proportional distributions of bundle function types were similar across the three proficiency levels and only small variations were observed. Nevertheless, more salient differences were found across the three proficiency levels in the use of the subfunctions of stance bundles, referential bundles, and the bundles of other functions.

Some limitations in this study should be acknowledged before we discuss the implications of these findings. This study employed only proficiency level as the sole background variable. Previous studies have established that other factors may affect linguistic choices in email writing, such as gender (McKeown & Zhang, 2015), task types (Li, 2000), cultural background (Björge, 2007), and age on arrival or length of acculturation (Leo, 2012). In addition, different writing prompts included in the current corpus may have elicited different types of speech acts. Future studies on lexical bundles in email writing tasks may include some of these factors to shed light on their effects. There are also some methodological concerns in this study. One is about the internal structure of lexical bundles. This study investigated lexical bundles as continuous fixed word sequences only. Renouf and Sinclair (1991) remind us that formulaic expressions can also be discontinuous, as witnessed in the recent studies on formulaic expressions with variable slots, which are known as phrase frame or concgrams (Cheng, 2007). Another concern is about the decisions of setting cut-off criteria for identifying lexical bundles. One of our challenges in setting the cut-off values was due to the special structure of the email writing corpus—a large number of short writing samples. This characteristic to some extent turns the criteria of frequency and range into one, as observed in our data. This is because, unlike in longer responses, it is uncommon to see a lexical bundle being used more than once in a short email sample. The current criterion (40 times per million words) is in the middle of cut-off values in other studies. Different criteria may be experimented with in future studies on a similar type of corpus.

The findings of this study have implications for both teaching email writing and testing email writing performances. Given the variety of discourse functions fulfilled by the lexical bundles, a list of lexical bundles will be informative for language teachers to teach email writing. Activities such as cross-proficiency comparisons of lexical bundles can serve as awareness-raisers for English language learners to notice these differences in the use of lexical bundles and to learn to use lexical bundles more appropriately depending on the contextual factors. In this regard, useful guidance and discussions on lexical-bundle related pedagogy can be found in Byrd and Coxhead (2010), Meunier (2012), and Cortes (2006). Meanwhile, information on lexical bundles can be useful for developing language testing projects such as revisiting scoring rubrics in light of the criterial features of lexical bundles. For example, the lexical bundles used by test-takers of different proficiency levels exhibited different levels of formality as well as politeness in writing. If pragmatic competence or appropriateness of language is a part of the writing construct to be measured, these distinctive features in lexical bundles can be used as validity evidence to support the interpretation of the scores on that particular aspect. The lexical bundle lists can also be incorporated in rater training materials, together with some corresponding concordance lines, to highlight the differential uses of lexical bundles by test-takers of different proficiency levels.

Authors

Zhi Li is a Language Assessment Specialist and research lead at Paragon Testing Enterprises, BC, Canada. He holds a PhD degree in the applied linguistics and technology from Iowa State University, USA. Zhi's research interests include language assessment, academic writing, computer-assisted language learning, corpus linguistics, and systemic functional linguistics.

Alex Volkov is a Content Development Lead at Paragon Testing Enterprises, BC, Canada. After receiving a Master's degree in Applied Linguistics from Carleton University, Alex has been working at Paragon on content development, scoring, test design, item bank management, and test validation.

References

- Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81–92. <https://doi.org/10.1016/j.esp.2011.08.004>
- Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University.
- Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly*, 13(1), 55–71. <https://doi.org/10.1080/15434303.2015.1126718>
- Baron, N. S. (1998). Letters by phone or speech by other means: the linguistics of email. *Language & Communication*, 18(2), 133–170. [https://doi.org/10.1016/S0271-5309\(98\)00005-6](https://doi.org/10.1016/S0271-5309(98)00005-6)
- Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275–311. <https://doi.org/10.1075/ijcl.14.3.08bib>
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, 263–286. <https://doi.org/10.1016/j.esp.2006.08.003>

- Biber, D., Conrad, S. M., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371–405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., Johansson, S., Leech, G., Conrad, S. M., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Longman.
- Biesenbach-Lucas, S. (2007). Students writing emails to faculty: An examination of e-politeness among native and non-native speakers of English. *Language Learning & Technology*, 11(2), 59–81. <https://doi.org/10.125/44104>
- Bjorge, A. K. (2007). Power distance in English lingua franca email communication. *International Journal of Applied Linguistics*, 17(1), 60–80. <https://doi.org/10.1111/j.1473-4192.2007.00133.x>
- Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, 5(5), 31–64.
- 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>
- Chen, Y.-S. (2016). Understanding the development of Chinese EFL learners email literacy through exploratory practice. *Language Teaching Research*, 20(2), 165–180. <https://doi.org/10.1177/1362168814562014>
- Cheng, W. (2007). Congramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning*, 1(1), 22–35.
- Conrad, S. M., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, 20, 56–71. <https://doi.org/10.1515/9783484604674.56>
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423. <https://doi.org/10.1016/j.esp.2003.12.001>
- Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17(4), 391–406. <https://doi.org/10.1016/j.linged.2007.02.001>
- Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12, 33–43. <https://doi.org/10.1016/j.jeap.2012.11.002>
- Crystal, D. (2001). *Language and the Internet*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511487002>
- Gains, J. (1999). Electronic mail—A new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes*, 18(1), 81–101. [https://doi.org/10.1016/S0889-4906\(97\)00051-3](https://doi.org/10.1016/S0889-4906(97)00051-3)
- Gray, B. E. (2016). Lexical bundles. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research* (Routledge Advances in Corpus Linguistics) (pp. 33–56). New York, NY: Routledge.
- Hyland, K. (2008). 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>
- Lan, L. (2000). Email: A challenge to standard English? *English Today*, 16(4), 23–29. <https://doi.org/10.1017/S0266078400000511>
- Leo, K. (2012). Investigating cohesion and coherence discourse strategies of Chinese students with varied lengths of residence in Canada. *TESL Canada Journal*, 29(6), 157–178. <https://doi.org/10.18806/tesl.v29i0.1115>
- Leopold, L. (2015). Request strategies in professional e-mail correspondence: Insights from the United States workplace. *TESL Canada Journal*, 32(2), 1–29. <https://doi.org/10.18806/tesl.v32i2.1205>
- Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System*, 28(2), 229–245. [https://doi.org/10.1016/S0346-251X\(00\)00009-9](https://doi.org/10.1016/S0346-251X(00)00009-9)
- McKeown, J., & Zhang, Q. (2015). Socio-pragmatic influence on opening salutation and closing valediction of British workplace email. *Journal of Pragmatics*, 85, 92–107. <https://doi.org/10.1016/j.pragma.2015.06.012>

- Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics*, 32, 111–129. <https://doi.org/10.1017/S0267190512000128>
- 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>
- Paragon Testing Enterprises. (2015). *CELPiP study guide: Reading and writing*. Vancouver, BC: Author.
- Renouf, A., & Sinclair, J. M. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 128–143). London, UK: Routledge.
- Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, 12(3), 214–225. <https://doi.org/10.1016/j.jeap.2013.05.002>
- Zhu, W. (2012). Polite requestive strategies in emails: An investigation of pragmatic competence of Chinese EFL learners. *RELC Journal*, 43(2), 217–238. <https://doi.org/10.1177/0033688212449936>

Appendix. Sample Email Writing Task in the CELPIP-General Test

Writing Task 1: Writing an Email

You recently made reservations for dinner at a very famous and expensive restaurant in town. However, the meal and the service were terrible. The restaurant manager was not available to solve the problem, so you left without a resolution.

Write an email to the restaurant's manager in about 150–200 words. Your email should do the following things:

State what problems you had with the food you ordered.

Complain about the service.

Describe how you want the restaurant to resolve the problem to your satisfaction.