Asia Pacific Journal of Corpus Research
Vol. 6, No. 1

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2025.6.1.25

"I Agree with This": A Multi-Modal Register Analysis of Lexical Bundles in the ICNALE

Trevor Sitler
(Kansai University)
Martin Spivey
(Akita University)

Sitler, T., & Spivey, M. (2025). "*I agree with this*": A multi-modal register analysis of lexical bundles in the ICNALE. *Asia Pacific Journal of Corpus Research, 6*(1), 25-45.

The purpose of this study was to examine the differences in lexical bundle use between the different registers of speaking and writing in a learner corpus. While various studies have examined multi-word items by foreign language learners, there is currently a lack of research on lexical bundle use in spoken corpora, particularly monologues. A comparative register analysis was designed to compare the lexical bundle output of L1 Japanese students of English in two corpora of argumentative written essays and spoken monologues, with two native speaker corpora being compared for reference. 4-word lexical bundles were extracted and analyzed through a Generalized Linear Mixed Model (GLMM) and categorized into two functional and structural taxonomies. Findings suggest bundle frequency decreases as proficiency level increases. Compositionally, it was found that both registers are structurally and functionally similar to each other, though the Japanese spoken monologue sub-corpus was the least varied. Furthermore, the use of prompt-specific language was found to be more influenced by register than L1 background, with both groups producing a significant amount of language that was dependent on the prompt. The register descriptions found in this study can inform pedagogical approaches to writing and monologue tasks.

Keywords: ICNALE, Learner Corpora, Lexical Bundles, Register Analysis, Sketch Engine

1. Introduction

In the past few decades, learner corpora have been utilized as tools to examine foreign and second language learners' productive output as well as for various purposes such as creating and compiling dictionaries, textbooks and course curricula, in addition to facilitating classroom instruction (Götz & Granger, 2024). Various studies have been conducted into learner language production including on the over/under-use of verbs in EFL student essay writing (Altenberg & Granger, 2001), collocations in EAP written project assignments (Durrant & Schmitt, 2009), and lexical verbs in academic writing (Granger & Paquot, 2009). In particular, there is a burgeoning number of investigations into formulaic language use in learner corpora (Biber et al., 2004; Chen & Baker, 2016; Hyland, 2008). A *formulaic sequence* is '... a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use…' (Wray & Perkins, 2000, p.1). Continuous word sequences have been given multiple labels but are often referred to as *lexical bundles*, which have been defined as '...sequences of words that were extracted from corpora on the basis of frequency information but do not constitute a complete structural unit...' (Jeong & Jiang, 2019, p.190). Some examples of frequent lexical bundles of differing lengths in written discourse include *as well as*, *the end of the* and *is one of the most*. In order to gain a deeper understanding of how language learners speak and write in L2 English, one should go beyond the production of individual words and instead take a more phraseological approach. By investigating the frequencies of lexical bundles in learner language, we can better comprehend how learners use their L2 and any common difficulties or errors that are typically made.

This present study seeks to develop our understanding of how L1 Japanese university students produce lexical bundles in L2 English monologues in contrast to written essays. First, we will offer some background on previous research into lexical bundle usage in learner corpora, as well as studies that have focused on bundles in spoken language production specifically. Following this, we outline our main research objectives to find the key similarities and differences between the two registers, with English native speaker production referred to for comparative purposes. To enable us to do this, four sub-corpora were created using the ICNALE corpora (Ishikawa, 2023) and 4-gram lexical bundles were extracted through the corpus software tool Sketch Engine (Kilgarriff et al., 2014). Data analysis was conducted with the aid of Biber et al.'s (2004) structural and functional taxonomies. A Generalized Linear Mixed Model (GLMM) was used to gain a deeper understanding of the degree to which the multiple variables involved (learner proficiency, register etc) had an influence on the text bundle frequency. Results showed that the two written corpora featured significantly more bundles than the spoken corpora. Additionally, as the English language proficiency of the Japanese students increased, there was an overall reduction in bundle production. Findings also indicated structural and functional similarities between the written essays and the monologues, with heavy usage of verb phrases and stance expressions among both L1 Japanese and English native speakers. Furthermore, there was a lack of evidence to suggest that L1 Japanese speakers are more reliant on the prompt than native speakers. However, there was an increased reliance by the former on discourse organizers to structure their responses. While recognizing the fact there are some limitations, our study findings have implications for Japan-based university English instruction and should be taken into consideration when designing language course curricula.

2. Literature Review

2.1. Learner Corpora

The next section will provide an outline of current research into lexical bundles in learner corpora, followed by a detailed look at studies involving spoken learner corpora.

2.1.1. Lexical Bundles in Written Learner Corpora

Several studies have utilized learner corpora for the analysis of lexical bundles to gain insights into key aspects of L2 learner language (e.g. Ädel & Erman, 2012; Chen & Baker, 2010, 2016). Chen & Baker (2010) investigated Chinese university students' use of lexical bundles alongside that of L1 English peers and L1 English expert writers. Comparing three sub-corpora created out of FLOB and BAWE[1], the authors examined the structural and functional categories of the lexical bundles produced, adopting the taxonomies designed by Biber et al. (2004). It was found that both learner groups relied somewhat on verb phrase-based (VP) lexical bundles over noun phrase- or prepositional phrase-based (NP/PP) bundles as they consisted of 55.8% and 52.5% of the L1 learners' and L2 learners' totals respectively. In addition, the Chinese writers showed an under-use of "passive verb + prepositional phrases" as well as stance expressions over referential and discourse-organizing bundles. In a later study, Chen and Baker (2016) examined the lexical bundles written by Chinese L2 English learners at three proficiency levels (CEFR B1, B2, C1) via 1,029 argumentative and descriptive essays taken from

[1] FLOB and BAWE refer to the Freiberg-Lancaster-Oslo/Bergen Corpus and British Academic Written English Corpus, respectively

the Longman Learner Corpus. The three sub-corpora were compared to native speaker academic prose and conversation from the Longman Grammar of Spoken and Written English corpus. Key findings include weaker Chinese learners' tendency to use VP bundles, similar to the conversational register, while higher-level students produced a greater proportion of NP/PP bundles, which more closely resembles the register of academic language. Moreover, the analysis revealed that Chinese students at all levels relied more on discourse-organizing bundles when compared to native English speakers. Bychkovska and Lee (2017) also compared 4-word lexical bundles in the argumentative writing of high proficiency Chinese L2 English and L1 English undergraduate students. While the authors discovered that the L2 English writers produced a greater proportion of VP-based bundles in comparison to the L1 English students, their findings contradicted Chen and Baker (2010) and Ädel and Erman (2012) to a certain degree as it was revealed that the Chinese students used a wider variety and higher frequency of bundles than the native speakers, though this may be due to setting different frequency and dispersion thresholds, in addition to a slight variation in functional classifications.

The argumentative essay writing of Chinese university students was also the focus of a study by Kim and Kessler (2022), who looked into the structural and functional constructions of lexical bundles (ranging from 3 to 5) in a small 18,000-word corpus of 120 essays. The essays were based on two topics (cellphone usage while driving and the safety of e-cigarettes) and handwritten under test conditions, with each essay subsequently graded for quality. The researchers extracted the three sets of bundles from the corpus and categorized them into prompt-dependent or non-prompt-dependent bundles. They extracted 22 4-word bundles compared to 66 and 21 3- and 5- word bundles, respectively. The majority of 4-word bundles were not prompt-dependent and roughly half expressed a stance. This differs from the 5-word bundles which featured more prompt-dependent sequences and proportionally more VP bundles. It was also discovered that higher-scoring students produced more prompt-dependent bundles than the lower-scoring students and also of note is that both groups tended to write more VP bundles.

Staples et al. (2013) investigated bundles at different proficiency levels through a 250,000-word corpus of TOEFL iBT responses. The data showed that the students at the lowest level produced a larger number of bundle tokens than the higher-level students, however there was a much greater reliance on prompt-based bundles. In addition, students at all three levels used much fewer referential bundles proportionally than discourse-organizing and stance expressions (less than 10% for each level). In a comparative study of advanced-level L1 Swedish and native British English linguistics university students, Ädel and Erman (2012) found that Swedish students produced a lower (and less varied) number of 4-word lexical bundles than their native English counterparts (60 vs 130). They also discovered that the native speakers used significantly more hedging and passive lexical bundles, while overall there were more NP/PP bundles, often featuring abstract nouns. This is apparently fairly typical of academic writing, with more phrasal lexical bundles being produced over clausal bundles (Biber, 2010).

Ruan (2017) investigated the development of lexical bundles by Chinese learners at an EMI university by comparing written assignments completed at four different points in their studies. Results showed that students in the early stages of their course frequently repeated a small number of bundles, however a larger range of types were used in later writing. Also, adopting the structural framework of Biber et al. (2004), it was found that students in the first year relied proportionally more on NP and VP bundles than PP bundles compared to later assignments which featured a greater proportion of PP bundles. This shows evidence of students' language development over the course of their university studies. As for the bundle functions, Biber et al.'s (2004) functional taxonomy was utilized to help discover that discourse organizers made up the greatest proportion of bundles (52.9% average) at every point, followed by referential expressions (34.3%) and then stance expressions (12.8%). An increase in the reliance of discourse organizing bundles from the second year onwards hints at the writing development undertaken by the Chinese students over the four years of their studies.

In Vo (2019), 4-word bundles produced in written tasks by ESL students in the US were examined through the English Placement Test Corpus. Three sub-corpora were created based on test scores in order to compare different written proficiency levels. A structural analysis was performed using Chen and Baker's (2010) taxonomy while functions were examined through Biber et al.'s (2004) taxonomy. The results highlighted that the two lower-level groups had a preference for using bundles based on the prompt words, while all three groups produced more NP/PP bundles than VP bundles. With regards to functions, only the highest-scoring group of learners displayed a greater preference for discourse organizers with the lowest level group much more reliant on referential expressions and stance bundles.

Using a sub-corpus of the ICNALE Written Essays corpus, Sawaguchi (2024) examined 4-word bundles produced by Japanese students in their argumentative written essays. 78 bundles were identified and analyzed for their functions and semantic transparency in order to create a list of target bundles for pedagogical purposes. It was found that Japanese learners used fewer referential bundles in their writing compared to native English speakers and also there was a tendency to rely on stance bundles. Furthermore, Japanese writers produced fewer of the more frequent transparent and opaque bundles than the native speakers and overused infrequent transparent bundles such as *a lot of money* and *think that it is*.

The aforementioned studies provide some useful insights into the lexical bundle production of L2 English learners. In spite of this, some issues require highlighting. First, the majority of research has been conducted on ESL learners or advanced-level EFL students. This means the output of such learners is likely to more closely resemble that of the L1 English writers that are used as a comparison. Secondly, these studies inform us about written production yet fail to tell us anything about how students produce spoken language. At present, there are considerably more corpora of student writing than spoken production due to the relative ease in which written texts can be collected. Spoken language initially requires audio recording and then manual (or automatic) transcription, which involves more expense, time and expert knowledge (Friginal, 2018; Gablasova & Bottini, 2022). The next section will briefly introduce some of the studies which have utilized corpora to examine L2 spoken English.

2.1.2. Lexical Bundles in Spoken Learner Corpora

While presently few in number, there are some investigations into spoken learner language in learner corpora research that can provide useful insights. As part of a study into the English vocabulary production of L1 Japanese speakers, Shirato and Stapleton (2007) examined a range of lexical bundles in a specialized 44,000-word corpus of adult EFL learner conversations, story descriptions and role-plays. The authors compared the corpus with a British National Corpus sub-corpus exclusively featuring roughly 4 million words of conversations. It was found that, compared to the native speakers, there was a distinct lack of 'vague' bundles such as *something like that* and *that sort of thing*, while there was a stark under-use of *sort of* for hedging purposes. While this study is informative, its focus is on spoken dialogues, combines bundles of various lengths, and fails to provide much qualitative analysis on the bundles produced. Sánchez Hernández (2013) investigated the structures and functions of 4-word lexical bundles created by L1 Spanish university majors in English. Students completed tasks in describing experiences, engaging in conversation, and story re-telling. Two corpora of first- and third-year students were created, as well as a reference corpus of British native speakers completing the same tasks. It was discovered that the native speakers produced the lowest number of bundle types and tokens. Also, all corpora featured a proportionally higher number of VP, with the L1 English corpus producing a greater proportion of NP/PP bundles than the two L2 English corpora. Regarding functional usage, referential expressions were comfortably the most frequent in all three corpora and discourse organizers barely appeared.

In Yan (2019), a study was carried out into the lexical bundles of L1 Chinese college English majors in spoken examinations. The data were extracted from a sub-corpus of the Spoken and Written English Corpus of Chinese Learners 2.0 (328,336 tokens) and it was found that all three proficiency levels under scrutiny produced a significantly greater proportion of topic-related bundles compared to non-topic-related ones, though reliance diminished as proficiency levels increased. Using Biber et al.'s (2004) taxonomy, the author also examined the different functional types and found that at all three proficiency levels, stance bundles were proportionally most frequent, followed by discourse organizing and lastly referential bundles. The production of discourse organizing bundles did proportionally increase slightly in correlation with the level, with referential bundles displaying the opposite effect. Finally, Lee and Zipagan (2018) analyzed a corpus of 58 Korean EFL students undertaking an Oral Proficiency Interview, which included speaking tasks describing their day, discussing how technology has changed, and talking about themselves. Their performance was also compared to the equivalent native speaker data. It was found that both groups predominantly used VP bundles and SE bundles, however the native speakers had a wider variety of bundles and also included more NP/PP bundles.

In summary, research to date (while limited) shows some evidence that L2 English learners tend to rely on verb phrases in their spoken production rather than more complex 'native-like' NP/PP phrases. There is also an over-reliance on the use of bundles directly related to the topic prompt, especially with lower proficiency levels, suggesting a lack of vocabulary or confidence when speaking. What is currently lacking is an in-depth understanding of how L2 English users produce lexical bundles while giving their opinions on the same topic(s) in both spoken and written registers. Moreover, little is known about the bundle usage of L1 Japanese students in the EFL context. The next section will outline the methodology of a research study designed to expand our knowledge on this area.

3. Methods

In this section, we first outline our key research objectives and the three questions we ask to guide our investigation. Following this is a comprehensive explanation of the research design. In particular, we provide details about the ICNALE corpora used in the study as well as Biber et al.'s (2004) structural and functional taxonomies.

3.1. Research Objectives

The objective of this investigation is to help us better comprehend how Japanese EFL learners produce lexical bundles in both spoken and written assessments. Specifically, we aim to learn about the key differences and similarities between the two registers and how these compare with the output of L1 English speakers. We also wish to develop our understanding of how students may or may not rely on the wording of the prompts when providing their written or spoken replies, and what this may tell us about Japanese students' ability to produce English and the difficulties they might face.

The following research questions have been created to help us achieve our objectives:

RQ1. What are the frequencies of the lexical bundle types and tokens produced by L1 Japanese students in English essay writing and spoken monologues?

RQ2. What are the structural and functional types of lexical bundles in L1 Japanese student English essay writing and spoken monologues?

RQ3. To what extent do L1 Japanese students rely on prompt-dependent lexical bundles when writing in English compared to speaking?

We will now provide a brief outline of the corpora used in the study, as well as the structural and functional taxonomies utilized for this investigation.

3.2. Data Collection

3.2.1. The ICNALE Corpora

The written and spoken data for this study were collected from the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2023). At present, the ICNALE collection of corpora consists of five corpora of both written and spoken texts. Its origins began with the ICNALE Written Essays corpus (Ishikawa, 2013), a 2.5-million-word corpus of argumentative essays written by EFL and ESL learners from several Asian countries, as well as L1 English speakers. Its main purpose is to contribute to the small number of corpora available that feature English language learners from Asia and to provide reliable corpus data for Contrastive Interlanguage Analysis between the numerous datasets. Essays of 200-300 words were written on two universally familiar topics (smoking in restaurants and student part-time work) under strict instructions, to control the conditions and ensure consistency across the data sets. The ICNALE Spoken Monologues corpus (Ishikawa, 2014) comprises 4,400 one-minute speeches from 1,100 Asian and native speaker contributors (500,000 words). The two topics replicate those in the essays and the speaker has two opportunities to talk on each subject. The data were collected via an answer phone system, with users following audio instructions for every stage of the process.

However, there were particular challenges for collecting data in the monologues that don't exist in the writing. First of all, in the monologues, students have a comparatively much shorter time, only 60 seconds, compared to the writing, in which they are given more time to formulate their thoughts. Second, participants in the monologue face the added pressure of performing quickly within a limited time frame. This caused many participants to become nervous and produce less speech.

In an attempt to remedy these challenges, Ishikawa (2014) decided to allow monologue participants two chances for each speech, noting that in the second chance, participants tended to produce more. While this may provide more linguistic output for analysis, this can pose additional challenges for data analysis. First, it is much easier for a particular participant's speech to be overrepresented, since they are speaking twice. Second, this violates the assumption of independence for any statistical analysis done. These issues will be addressed in section 3.3 below.

To support us in our research goals, we created two sub-corpora of L1 Japanese Spoken Monologues (J-SM) and L1 Japanese Written Essays (J-WE) in addition to two sub-corpora of L1 English Spoken Monologues (ENS-SM) and L1 English Written Essays (ENS-WE). Table 1 presents the relevant data. A decision was made to use the Spoken Monologues corpus to investigate the spoken language instead of Spoken Dialogues, as it was felt that this was closer to the Written Essays format in the sense that there was no interlocutor, the two topics were exactly the same, and the assessment was conducted under a strict time limit.

Table 1. ICNALE Sub-corpora

Contents	J-SM	J-WE	ENS-SM	ENS-WE
Total Tokens	48,469	198,731	105,100	96,976
Total Files	600	800	600	400
Average Tokens per File	80	248	175	242

3.2.2. Structural and Functional Taxonomies

This study employs the methodology and categorization scheme established in Biber et al. (2004). In the study, the authors used the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus), which includes samples from classroom teaching, textbooks, and other activities associated with university life. From this, structural and functional taxonomies were developed as a means to categorize lexical bundles. The three main categories in the structural taxonomy are: verb phrase (VP) bundles, dependent clause (DC) bundles, and noun/prepositional phrase (NP/PP) bundles.

In the same study, the authors sought to classify the different discourse functions of their bundles and developed a taxonomy that identified the main function of each one. These are stance expressions (SE), discourse organizers (DO), and referential expressions (RF). SE include epistemic stances such as *I don’t know if*, and phrases that express attitudes and modality as in *it is important to* and *is going to be*. DO introduce or focus on the topic (*if you look at*, *I would like to*, etc.) and elaborate on or clarify the topic (*on the other hand, as well as the*). Lastly, RF may identify or focus on something (*is one of the*), express imprecision (*or something like that*), name attributes (*have a lot of*), or refer to time, place etc. (*in the United States*, *at the same time*). A fourth category includes conversational functions such as *thank you very much* and *what are you doing*, however this is ignored in the present study as none of the corpora feature dialogic discourse or audience-facing presentations. As with the structural categorization, for the statistical analysis this study utilizes the three main functional categories.

Pan et al. (2020) found that the distribution of lexical bundles into these categories remains relatively consistent, even when comparing sub-corpora of different text and token counts. While the number of texts and tokens may have a large influence on the amount of lexical bundles, the proportion of bundles assigned to each sub-category in the above taxonomies remains consistent across sub-corpora.

3.3. Lexical Bundle Criteria

After extraction, there are two main criteria for choosing which lexical bundles to include for further analysis: length and dispersion. Lexical bundle length refers to how many words are included in the lexical bundles under scrutiny. Biber et al. (1999) reported that 3-grams tend to be too numerous to do a detailed study, and in this study, with the above criteria, a large number of 1521 3-gram lexical bundles were found. Likewise, 5-gram bundles tend to be too few, and in this study, 258 5-word n-grams were found. The 4-gram lexical bundles fell in a decent middle ground of 308, and it was decided to be the most appropriate number for the analysis. This would also allow ease of comparison to previous lexical bundles studies, as 4-grams are the predominant choice. (Yan, 2019; Lee and Zipagan, 2018; Chen and Baker, 2010). Furthermore, doing so would allow ease of comparison to previous studies.

The dispersion criteria for this study required more consideration. Dispersion refers to the number of different texts in which a single lexical bundle appears. Lexical bundles that appear in only one or two texts can be thought of as idiosyncratic of the particular speaker and not representative of the corpus as a whole. Previous studies used a dispersion criterion of three in a corpus in which only one individual made each text (Chen & Baker, 2016). However, in the monologue sub-corpora of the ICNALE (in this study represented by J-SM and ENS-SM), the original dataset contains two speeches from one individual for each topic as mentioned in section 3.2.1 above. In an attempt to capture a fuller range of linguistic possibilities in the monologue sub-corpora, both the first and second speeches for each individual were analyzed in this study. However, this can pose a major problem in overrepresenting a particular speaker’s idiosyncrasies and choice of lexical bundles. For this reason, a much stricter dispersion criterion of 12 texts was chosen in order to ensure that at least three different speakers used each bundle. This was done in order to provide a fuller range of linguistic possibilities while at the same time attempting to prevent individual speakers from overrepresenting the data. This criterion was also applied to the written essay sub-corpora for consistency.

Moreover, an additional criterion was established for analyzing prompt-based bundles. If a bundle contained two or more exact words taken from the prompt, it was tagged as a prompt-based bundle. Otherwise, it was tagged as a non-prompt-based bundle.

3.4. Data Analysis

The bundles were extracted using Sketch Engine (Kilgarriff et al., 2014) and then sorted by the above criteria. Subsequently, the bundles were manually coded by two individual coders according to the structural and functional taxonomies in Biber et al. (2004). The initial inter-rater reliability scores for the two taxonomies, respectively, were as follows: J-WE (59.5%, 65.5%), J-SM (68.1%, 82.6%), ENS-WE (77.2%, 65.8%), ENS-SM (81.2%, 70.9%). The coders then proceeded to discuss and come to an agreement on all discrepancies, resulting in 100% agreement for all eight data sets.

After the bundle extractions, statistical analysis and data visualization were conducted with R (version 4.3.3; R Core Team, 2024) using R Studio (version 2024.12.1) and the 'ggplot2' package (Wickham, 2016). First, to answer the first research question regarding bundle frequency, a Generalized Linear Mixed Model (GLMM) was used to estimate the influence of proficiency level, register, use of prompt-based bundles, structural, and functional type on bundle frequency per text. Since each speaker produced multiple texts, including a random intercept for Text ID enabled the study to account for within-speaker variation, and thus avoided violating the assumption of independence. GLMMs have been recommended for use in corpus linguistics and learner corpus research due to the hierarchical, clustered, and often non-normal nature of the data (Gries, 2021). The fixed effects were proficiency level (reference: native speaker), register (written), bundle type (non-prompt-based), structural type (VP), and functional type (SE). A random intercept for text ID was included to account for the clustering of observations for the multiple texts produced by speakers. The dependent variable was the count of lexical bundles per text.

Data analysis was conducted in R using the glmmTMB function from the glmmTMB package (McGillycuddy, 2025), which allows for the estimation of generalized linear mixed-effects models with truncated count distributions. As all texts contained at least one lexical bundle, a truncated negative binomial model was selected. This model accounts for both overdispersion and structural zero-truncation, making it appropriate for the distributional characteristics of the data.

Model fit was evaluated using the Akaike Information Criterion (AIC) and conditional and marginal R^2 values. The final model had an AIC of 39,165. The conditional R^2 (including both fixed and random effects) was 63%, while the marginal R^2 (fixed effects only) was 5.8%. This means that random effects accounted for 57.2% of the total variance. Model diagnostics were conducted using the DHARMa package (Hartig, 2024). Simulated residuals were plotted using QQ plots and residual vs. predicted plots. The residuals followed the expected pattern closely, with no clear patterns or outliers, indicating good model fit. No overdispersion was detected.

Next, the composition of each sub-corpora was analyzed in light of the above-mentioned lexical bundle taxonomy in Biber et al. (2004). To examine the structural, functional, and prompt-based distributions for each sub-corpora, stacked bar charts were made to visualize the distribution as a percentage. Subsequently, a chi-squared test of independence was performed, followed by a post hoc analysis using standardized residuals. The chi-squared test quantified variations shown in the visualization, with standardized residuals of |2| or more indicating over- or underuse. Effect sizes (Cramer's V) confirmed the practical significance of these trends. However, as mentioned in 3.2.1. above, the fact that the same speaker spoke twice for each topic violates the assumption of independence. It should be noted, however, that the main goal of this study was descriptive, and the chi-squared test was meant to quantify the observed differences seen in the visualization.

Finally, in order to provide a more detailed analysis of specific bundles used, tables for the top twenty non-prompt-based bundles and the top shared bundles between sub-corpora were created. The top twenty non-prompt-based bundles were analyzed to show the extent to which the participants could create original language independent of the prompt. The top twenty shared bundles show how common these original bundles were across sub-corpora. The creation of the tables was completed using the 'flextable' package (Gohel & Skintzos, 2024).

4. Results & Discussion

This section presents the results of the study by addressing each research question individually and discussing how the data relate to previous findings in the field. The section concludes with a consideration of the study limitations.

4.1. Frequencies of Lexical Bundle Types and Tokens

Table 2. Bundle Types and Total Raw Frequency

Sub-corpus	Bundle Type	Total Raw Frequency
J-WE	189	12,198
J-SM	64	1,574
ENS-WE	149	2,485
ENS-SM	117	3,386

First, overall frequency will be discussed. As shown in Table 2, there was a large amount of variation for the type and token of lexical bundles across the four sub-corpora. For the written register, Japanese learners produced a higher raw frequency and a higher token frequency than the native speakers, while for the spoken monologues, native speakers had a higher token and raw frequency. These mixed results are in contrast to the Ädel & Erman (2012) and Bychkovska & Lee (2017) studies on writing, which found that learners produced either a lower or higher amount of bundles respectively. For spoken corpora, this is also in contrast to Sánchez Hernández (2013) who found that native speakers produced fewer bundle types than learners.

Pan et al. (2020) noted that the raw frequency of lexical bundles is heavily tied to the text and token size of the corpus. In the written register, this also appears in this study: J-WE had the largest text number and token size and also had the largest amount of bundle tokens and frequencies. The ENS-WE had half the text size as well as a 132.3% difference in the raw frequency count. However, for the token count the percentage difference between the two groups was only 24.7% despite having such a large difference in the number of texts. This suggests that the J-WE contributors were much more repetitious in their bundle usage while the native speakers tended to use a larger variety of bundles.

For the monologue, the text size between the L1 Japanese sub-corpora and the native speakers was equal, however the token count for the ENS-SM was almost half as much as the J-SM, though the raw frequency count was quite close. These findings suggest that native speakers in the spoken monologue may be more repetitious and produce a greater variety of bundles.

Despite the fact that participants were allowed to speak twice for each topic in the monologue, J-SM still produced the least number of bundles in terms of both type and raw frequency. This highlights the overall lack of production in the monologic register for Japanese speakers. This may have a cultural influence, however, as it has been noted that Japanese learners in particular have a tendency to be more reticent and hesitant to speak (Ellis, 1991; Anderson, 1993).

Biber et al. (2004) found that lexical bundles overall were much more common in speech than in writing, however, in this case, it appears to be the opposite, with writing containing more lexical bundles than the spoken monologues for both corpora. This is especially true for L2 learners of English in this study.

Next, the results of the GLMM will be discussed. The results for the fixed effects can be seen in the table below. The model indicated .08 variation for the random effect of text ID, with a standard deviation of .29, indicating significant variation among different texts, even with the same speaker making multiple texts. As the coefficients were originally estimated on the log scale, they were exponentiated to produce rate ratios (RR). Rate ratios (RRs) greater than 1 indicate an increased frequency of lexical bundles relative to the reference level, while RRs less than 1 indicate a decrease.

Table 3. Fixed Effects from GLMM (Exponentiated Coefficients)

Fixed Effect	Estimate (RR)	Std. Error	Z-Score	95% CI Lower	95% CI Upper	*p*-value
(Intercept)	3.00	0.096	34.15	2.82	3.19	<0.001
A2	1.30	0.048	7.18	1.21	1.40	<0.001
B1-1	1.22	0.043	5.72	1.14	1.31	<0.001
B1-2	1.23	0.054	4.62	1.13	1.34	<0.001
B2+	1.00	0.060	0.04	0.89	1.13	0.966
Spoken Monologue	0.67	0.020	-13.36	0.63	0.71	<0.001
Prompt Based Bundle	2.96	0.070	45.65	2.82	3.10	<0.001
DC	0.44	0.013	-27.05	0.41	0.46	<0.001
NP/PP	0.59	0.017	-18.27	0.56	0.62	<0.001
DO	0.16	0.012	-24.21	0.13	0.18	<0.001
RF	0.87	0.023	-5.18	0.83	0.92	<0.001

As can be seen from the Table 3, the CEFR levels of A2, B1-1, and B1-2 were found to be significantly more frequent in their bundle use. Specifically, A2 were found to use bundles 30% more frequently, B1-1 used bundles 22% more frequently, B1-2 used bundles 23% more frequently. The level B2+, however, did not have a significant change from native speakers. This pattern suggests that lexical bundle frequency is higher at lower proficiency levels and progressively declines towards native speaker level.

It was found that spoken monologue had a 30% decrease in bundle frequency compared to the written register. This is in contrast to Biber et al. (2004), which found lexical bundles to be more frequent in spoken registers. Prompt-based bundles were also almost three times more frequent than non-prompt-based bundles, indicating the large influence of the prompt on bundle frequency.

For structure, the DC and NP/PP structural types were used much less frequently than the reference level VP. DC bundles showing a 56% decrease and NP/PP bundles showing 41% decrease compared to VP based bundles. As for functional types, SE was found to be by far the most frequent. DO bundles were used 84% less and RF bundles were used 13% less. This is consistent with native speakers (Biber et al., 2004) as well as other studies on learner corpora (Lee and Zipagan, 2018), which found that VP bundles and SE bundles were the most frequently used bundles.

Furthermore, the fact that random effects account for 57.2% of the total variance emphasizes the need to employ random effects in this model and potentially other models on learner corpora. Proficiency level, register, and bundle type were only able to account for 5.8% of the variance. This suggests that two texts by the same speaker varied significantly. This would make sense for the Spoken Monologues, as speakers were given two chances, one of which was done after warming up. This suggests that bundle use can vary significantly even among individuals at different points. However, this study incorporated both the written and spoken monologues together, and did not explore individual variation in the spoken monologues themselves. Future studies can explore how individual variation, especially when speakers are given multiple chances on the same topic.

To further explore these findings, the next section examines how structural and functional bundle types are distributed across the sub-corpora.

4.2. Structural Distribution of Bundles

Table 4. Structural Composition of Sub-Corpora in Raw Frequencies

Sub-Corpus	VP	NP/PP	DC
ENS-SM	1,987	790	609
ENS-WE	1,406	750	329
J-SM	1,216	185	173
J-WE	7,092	3,451	1,655

Next, the structural distribution of bundles across sub-corpora will be analyzed. The results of a chi-squared test of independence shown in Table 3 reveals that there were significant differences in the structural distribution between the groups, $\chi 2$ (6, N = 19,643) = 25.844, p < .001, V = .09. This indicates that while there is a significant difference in the structural distribution across sub-corpora, the effect size is quite small. To specify which structural categories contributed most to the differences, standardized residuals were examined for each sub-corpus. Residuals with an absolute value greater than |2| were considered significant.

The greatest source of variation appears to be the lack of use of NP/PP bundles in the J-SM corpus (standardized residual = -11), though ENS-SM also used fewer NP/PP bundles (standardized residual = -3) than both the J-WE (standardized residual = 4) and ENS-WE (standardized residual = 4). This indicates that within this study, NP/PP phrases were used much more often in the written sub-corpora than in the monologue. Another significant source of variation was the use of VP bundles, which was especially high in the J-SM (standardized residual = 9). VP bundles were used comparatively less often by J-WE (standardized residual = -2) and ENS-WE (standardized residual = -2). DC bundles also varied significantly, with ENS-SM using a significant amount (standardized residual = 6) and J-SM (standardized residual = -2) and J-WE (standardized residual = -2) containing relatively fewer of them.

The differences between the sub-corpora are shown visually through a stacked bar chart displaying the total proportion of bundles as expressed through raw frequency. It can be seen that J-SM is the most different, having the least variation and relying primarily on VP bundles. Overall, VP bundles were used the most among all four sub-corpora, making up more than half of the distribution. However, native speakers generally used more NP/PP bundles and more DC bundles, indicating an overall greater structural variety than the learner sub-corpora. This is consistent with Chen and Baker (2016), Bychkovska and Lee (2017), and Lee and Zipagan (2018), which found that lower-level students tended to use more VP bundles in the written register, while native speakers more often used NP/PP based bundles. Many of the VP bundles contained verbs that indicated their opinion on the topic: *I agree this opinion*, *I think smoking is*, *I do believe that*. Learners in J-SM in particular overused the construction "I + agree + with + this/the", and even frequently used it within a non-grammatical construction, e.g. *I agree this opinion.* Native speakers, on the other hand, used this construction much less frequently, and when it was used, it was often combined with a more complex relative that-clause e.g. *I agree that smoking.*

NP/PP were made up mostly of the nouns and phrases that were mentioned in the question: *at all the restaurants*, *for a college student*, *smoking in the restaurant*. Many DC were used to modify nouns that were mentioned in the prompt: *who want to smoke, to have a job, idea that smoking should.*

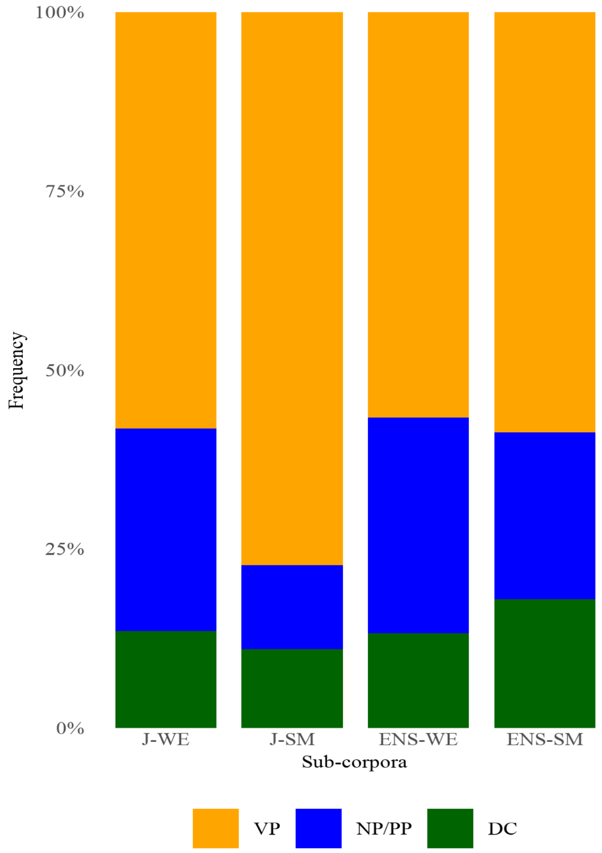

Figure 1. Structural Distribution of Lexical Bundles Across Sub-Corpora

In Biber et al. (2004), it was found that the writing language of native speakers and expert-level writers is characterized by the extensive use of NP/PP fragments, while the use of VP bundles was characteristic of conversation. Biber et al. (2011) argues that the tendency for learners to rely on VP bundles indicates that their writing is more characteristic of conversational registers than written registers. In light of that study, the prevalence of VP bundles makes both registers in this study more similar to speaking than the written register, though native speaker corpora were able to incorporate more NP/PP based bundles than the L2 learners.

While structurally the written and spoken modes among native speakers are quite similar, the monologues of the L2 English learners had a notable absence of NP/PP bundles and a large use of VP based bundles, making up 75% of the total structural composition as can be seen in Figure 1. If the native speaker corpora are used in a reference, progression in monologues may mean the integration of more NP/PP based bundles.

4.3. Functional Distribution of Bundles

Table 5. Functional Composition of Sub-Corpora in Raw Frequencies

Sub-Corpus	SE	RF	DO
ENS-SM	2,099	1,271	16
ENS-WE	1,392	1,025	68
J-SM	1,158	274	142
J-WE	6,450	5,168	580

Next, the functional distribution of lexical bundles will be analyzed. The results of a chi-squared test of independence reveal that there were significant differences in the functional distribution between the groups, χ^2 (6, N = 19,643) = 312.14, p < .001, V = 0.12. This indicates a significant difference between sub-corpora with a medium effect size, much larger than the effect size for structural distribution above. To specify which functional categories contributed most to the differences, standardized residuals were examined for each cell. Residuals with an absolute value greater than |2| were considered significant. The most striking difference comes from the J-SM sub-corpora, which used significantly fewer RF bundles (standardized residuals = -14) and a significant overuse of DO bundles (standardized residual = 10). In contrast, ENS-SM used far fewer DO bundles (standardized residual = -10), indicating an overreliance on DO bundles by L1 Japanese speakers for monologues. The native speaker and Japanese L1 monologues also contained significantly more SE bundles (standardized residuals = 4 and 9, respectively) than the two written sub-corpora. The J-WE sub-corpora contained significantly fewer SE bundles (standardized residual = -5), though significantly more RF bundles and DO bundles (standardized residuals = -5 and 4 respectively).

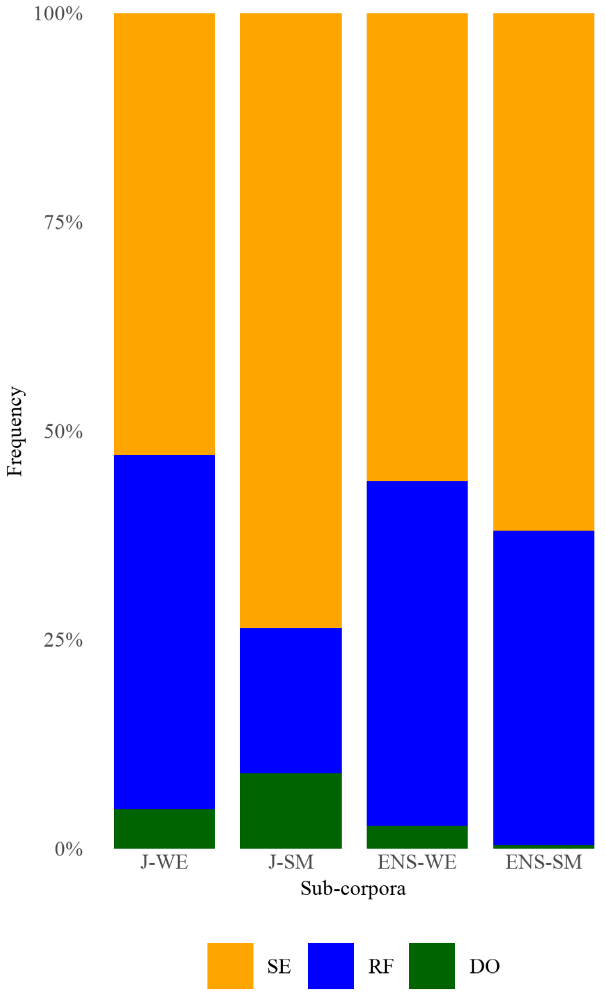

Figure 2. Functional Distribution of Lexical Bundles Across Sub-Corpora

The results of the analysis are visualized in Figure 2. One of the most striking differences between native and L1 Japanese sub-corpora is the absence of DO bundles among the native corpora. These results are similar to Chen and Baker (2016), which also found that Chinese L2 learners relied more heavily on DO than native speakers. A common structure used by the Japanese students was “Verb + (number) reasons”, e.g. *I have two reasons*, *I have three reasons*, *There are two reasons*. This is a structure that native speakers did not use but which Japanese speakers often relied on to structure their arguments. To the extent that native speakers used DO, it was often in the written section and included phrases such as *on the other hand*, *the fact of the matter*, and *when it comes to*. Overall, it appears that L1 Japanese students relied on explicit discourse organizers to structure their output, especially so in the monologue register.

However, the sub-corpora were similar in that SE bundles made up the largest percentage of the distribution. Bundles such as *I agree with this, I do not think, and I think that* were common across all four sub-corpora, however they tended to be more numerous in the written sub-corpora. However, genre may have had a decisive influence here, as making both written and spoken arguments requires frequent expressions of one's opinion and stance.

The written sub-corpora included more RF bundles, most of which referenced the subject matter of the question e.g. *with the statement that, having a part-time job*. For the spoken registers, ENS-SM in particular was composed of many referential expressions that tied their speech to the topic, something that was notably absent in J-SM, which was comprised predominantly of SE, with the above-mentioned *I agree with this* being the most frequent.

Chen and Baker (2010) found that lower-level students relied more heavily on SE bundles in essays, noting that this created a more personal, argumentative tone, which is usually not found in the academic writing studied. Biber et al. (2004) likewise found that SE bundles were more common in conversation than in academic writing. However, in the case of these registers, argumentative essays and monologues are not strictly academic and may be permitted to have a different tone than the objective, clinical tone needed in academic writing. Argumentative essays and monologues, then, may require more of a personal style than the types of academic writing previously studied.

4.4. Prompt-Based Bundle Distribution

Table 6. Chi-Square Analysis of Prompt vs. Non-Prompt Frequencies

Sub-Corpus	Non-Prompt-Based Bundle	Prompt-Based Bundle
ENS-SM	1,331	2,055
ENS-WE	1,135	1,350
J-SM	689	885
J-WE	4,300	7,898

Next, the distribution of prompt-based bundles was analyzed. The results of a chi-squared test of independence reveals that there were significant differences in the use of prompt-based bundles between the groups, χ^2 (3, N = 19,643) = 126, p < .001, V = .08. This indicates a significant difference between sub-corpora, though the effect size is quite weak. To specify how often prompt or non-prompt bundles were used for each category, standardized residuals were examined for each cell. Residuals with an absolute value greater than |2| were considered significant. ENS-WE used significantly fewer prompt-based bundles (standardized residual = -5) and more non-prompt-based bundles (standardized residual = 5), whereas J-WE used more prompt-based bundles (standardized residual = 4) than non-prompt-based bundles (standardized residual = -5). This shows that L2 Japanese learners used more prompt-based bundles in the written register than native speakers. However, for the spoken monologues, the opposite result appears, with J-SM using significantly fewer (standardized residual = -3) prompt-based bundles than ENS-SM.

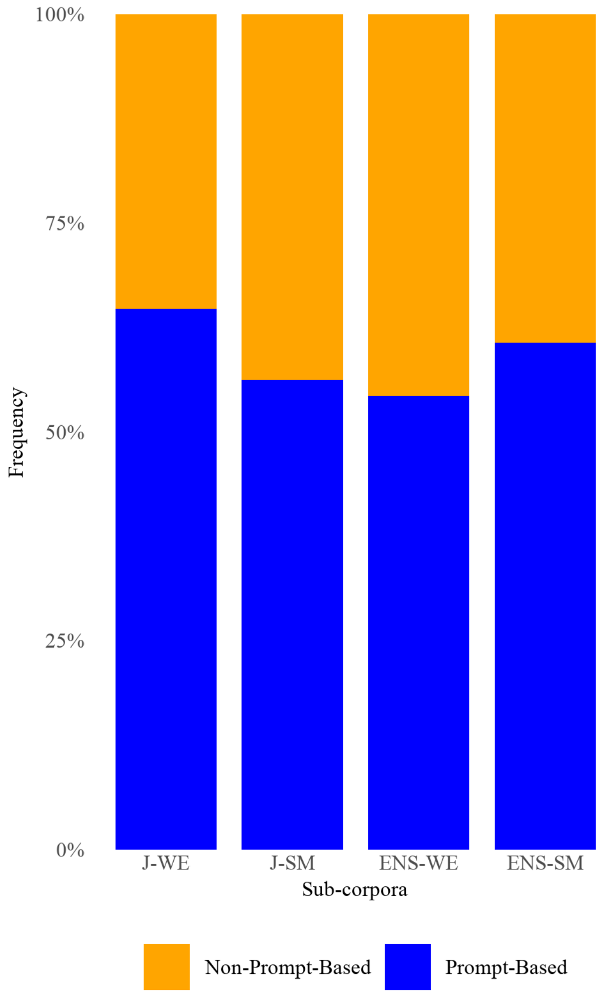

Figure 3. Distribution of Prompt and Non-prompt-based Lexical Bundles Across Sub-corpora

The results are visualized in the stacked bar chart shown in Figure 3. From this data it cannot be said that L1 Japanese students on the whole relied more on prompt-based bundles than native speakers. This may be true for the written sub-corpora, but the opposite results were found in the spoken monologue, indicating that register may be a more important variable in the use of prompt-based language. While this study did not measure proficiency as a variable, these results are in contrast to Staples et al. (2013) and Vo (2019), which found that lower-level students relied more heavily on prompt-based bundles. If native speakers are taken as being more proficient than L2 learners, then the results of this study imply that the use of prompt-based bundles may not be directly tied to proficiency in monologues. This is also in agreement with Kim and Kessler (2022), which found that higher-scoring students used more prompt-based bundles than lower-scoring students. Furthermore, prompt-based bundle use was over half of total bundles for all four groups. This is in contrast to Kim and Kessler (2022), which found that the majority of 4-word bundles were not prompt-dependent.

It should perhaps come as little surprise that prompt-based bundles are frequently used in both argumentative writing and speaking produced under exam conditions. Teachers often instruct students to use the question words in their responses as a form of signposting that will offer text coherence and increase the potential of gaining a higher score. Furthermore, reference to and inclusion of prompt-based expression is necessary in an argumentative context in order to fully express and reason one's opinion on the topic. It is reasonable to suggest that in the J-WE, Japanese learners were more reliant on the prompts in their written responses because the questions were available for reference throughout the entire examination whereas for the J-SM the speakers only heard the question and instructions prior to giving a reply.

4.5. Most Frequently Used Bundles

The next logical step is to remove the prompt-based bundles and examine the most frequent non-prompt-based bundles for each corpus. This may give us a better understanding of how language is produced in each register. As a reminder, prompt-based bundles are bundles with at least three question words in any particular order. Non-prompt-based bundles may refer to the question topic but cannot be said to be directly copied from the prompt. Table 5 (see Appendix 1) presents the top 20 non-prompt-based bundles for the four corpora.

Initial observations highlight the fact that there are several personal stance expressions in the lists of most frequent bundles (e.g. *I think it is*, *I don't like*), which one would expect in the mode of argumentation. There appears to be overlap with some of the bundles in the table. Take, for instance, the top two bundles in J-SM; *I think it is* and *is not good for*. The former has a raw frequency of 47 while the latter features 34 times. One may assume that these two overlap, however *I think it is not good for* appears only once in the J-SM corpus. Therefore, we have chosen to leave the bundles as is to avoid skewing the data.

Another interesting point to make is the number of highly frequent bundles shared between the different corpora. Table 7 shows the shared bundles from the top 20 in alphabetical order. The corpora in which the bundles appear in the most frequent list are in bold and the other corpora have their relative frequencies in chevrons for comparison.

Table 7. Top Shared Bundles

Bundle	J-SM	J-WE	ENS-SM	ENS-WE
a lot of money	**371**	**392**	<48>	<62>
a lot of people	<309>	**332**	**247**	<113>
Don't want to	**454**	**252**	**324**	<72>
I don't know	<62>	<55>	**343**	**196**
I don't think	**371**	**211**	**923**	**371**
I think it is	**970**	**1,001**	**285**	**320**
I think smoking should	<289>	**367**	**238**	<31>
I think that it	<124>	**720**	**276**	**423**
I think that smoking	<62>	**357**	**295**	**134**
people who don't	**433**	**468**	<181>	<31>
should have a part-time	<62>	**357**	**371**	**237**
think that it is	<144>	**684**	**114**	**371**
think that smoking should	<0>	**367**	<333>	**175**

Perhaps unsurprisingly, many of the shared bundles are compounds of stance expressions such as *don't want to* and *should have a part-time*. Compared to L1 English speakers, Japanese learners are more likely to refer to *a lot of money* and *people who don't*. For reference, these two bundles feature 5 and 20 times respectively in the ENS-SM, and 6 and 3 times respectively in the ENS-WE corpus. When discussing money, ENS-WE is more likely to refer to *the value of money*, ranked the 15th most frequent non-prompt-based bundle. This expression is not used once in the J-SM but features 16 times in the J-WE.

The results of this study suggest that argumentative essays are structurally and functionally similar to each other, with the exception of the J-SM sub-corpora, which contained far more VP and SE bundles and less NP/PP and RF bundles than the other sub-corpora. Putting aside the J-SM, it would seem that the written essays and spoken monologues have similar functional goals, which is to convince a large audience of a certain point of view. As such, their structural and functional distribution is similar, with both including a large amount of VP and SE bundles.

While the Japanese sub-corpora did not rely more on the prompt than the native speaker sub-corpora, they were more reliant on formulaic DO bundles to structure their argument, e.g. *I have three reasons*. Japanese students also seemed to struggle more with argumentative monologues than with the written essay, relying too much on VP based bundles and having difficulty incorporating NP/PP and RF bundles.

5. Limitations & Conclusion

This study used lexical bundles as defined by Biber et al. (2004) to compare and contrast four sub-corpora of spoken monologues and argumentative essays produced by L1 Japanese speakers and English native speakers. The sub-corpora were investigated for the overall frequency of lexical bundles, structural and functional composition, and use of bundles related to the prompt. In terms of frequency, results show that while L1 Japanese used a higher frequency of the same bundles, native speakers used a wider variety of bundles, each of which were used less frequently. It was also found that raw bundle frequency tends to decline as proficiency level increases up to the native level. As for composition, the two sub-corpora were structurally and functionally similar to each other, with a predominance of verb phrase-based bundles and stance expressions. Native speakers tended to use more NP/PP based bundles and L1 Japanese speakers relied more on DO. Finally, the results of this study do not show that L1 Japanese learners produced significantly more prompt-based bundles than the native speakers, with native speakers using a great deal more prompt-based bundles in the spoken monologues than the Japanese speakers.

The key findings from this investigation have implications for classroom pedagogy in the L1 Japanese EFL context. Firstly, instructors can take the most frequent NP/PP bundles from the ENS corpora and explicitly teach them to their learners to help them become more proficient language users. Also, concordance lines featuring key bundles can be utilized in class activities as part of data-driven learning instruction that allows students to see firsthand how bundles are produced in authentic language by speakers of various linguistic backgrounds.

While this study yielded valuable insights into bundle usage, it was not without its limitations. This study focused exclusively on L1 Japanese speakers. Future studies can examine speakers of different L1 backgrounds to examine if this influences bundle frequency, bundle composition, or prompt usage.

Likewise, the choice to focus on the argument genre may have influenced the structural and functional composition of the sub-corpora. As noted above, the use of stance expressions in particular is necessary when arguing one's opinion on a topic, and other genre types, such as expository or personal narrative, may have a very different structural and functional composition.

This study also highlighted the difficulty of comparing different modes, even on the same topic. The different modes posed different challenges in data collection, which likewise posed challenges in data analysis. Future comparisons of the two modes can limit speeches to only once per participant. This would also allow a lower threshold for bundle inclusion, which may or may not have an impact on the linguistic diversity of the monologue register.

It would also be useful to expand on recent work done by Sawaguchi (2024) and examine how frequently Japanese learners produce the 78 target 4-word bundles when speaking in monologue and dialogue tasks, in addition to the written argumentative essay.

Developing speaking and writing skills in second language instruction requires two varied approaches in order for learners to improve their proficiency and produce target-level English in both respective registers. This study has sought to gain a deeper understanding of how L1 Japanese students speak and write in L2 English through an investigation of 4-word lexical bundles. Despite its limitations, it provides food for thought for instructors seeking to support their students in developing their English language capabilities and can assist syllabus designers when creating tasks to work on formulaic language at the academic level.

Acknowledgements

Special thanks to Atsushi Mizumoto for assistance with the Generalized Linear Mixed Model.

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

Anderson, F. (1993). The enigma of the college classroom: Nails that don't stick up. In Wadden, P. (Ed.) *A Handbook for Teaching English at Japanese Colleges and Universities.* (pp. 101-110). New York, NY: Oxford University Press.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195. https://doi.org/10.1093/applin/22.2.173

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In Heine, B., & Narrog, H. (Eds.), *The Oxford Handbook of Linguistic Analysis* (1st ed., pp. 159–191). New York, NY: Oxford University Press.

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*. London: Longman.

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., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Academic Writing*, *14*(2), 30–49.

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*, *47*(2). https://doi.org/10.1515/iral.2009.007

Ellis, R. (1991). Communicative competence and the Japanese learner. *JALT Journal, 13*, 103-127.

Friginal, E. (2018). *Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities*. New York, NY: Routledge. https://doi.org/10.4324/9781315649054

Gablasova, D., & Bottini, R. (2022). Spoken learner corpora for language teaching. In Jablonkai, R. R., & Csomay, E. (Eds.), *The Routledge Handbook of Corpora and English Language Teaching and Learning* (pp. 296–310). New York, NY: Routledge.

Gohel, D., & Skintzos, P. (2024). flextable: Functions for tabular reporting. R package version 0.9.7, https://davidgohel.github.io/flextable/, https://ardata-fr.github.io/flextable-book/.

Götz, S., & Granger, S. (2024). Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research*, *10*(1), 1–38.

https://doi.org/10.1075/ijlcr.00039.got

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In Charles, M., Hunston, S., & Pecorari, D. (Eds.), *Academic Writing: At the Interface of Corpus and Discourse*. London: Bloomsbury Academic. https://doi.org/10.5040/9781474211703

Hartig, F. (2024). DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.4.7, https://github.com/florianhartig/dharma.

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

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Ishikawa, S. (2014). Design of the ICNALE spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner Corpus Studies in Asia and the World*, *2*, 63–76.

Ishikawa, S. (2023). *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. New York, NY: Routledge.

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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Kim, S., & Kessler, M. (2022). Examining L2 english university students' uses of lexical bundles and their relationship to writing quality. *Assessing Writing*, *51*, 100589. https://doi.org/10.1016/j.asw.2021.100589

Lee, K. R., & Zipagan, M. N. (2018). Korean english learners' use of lexical bundles in speaking. *The Journal of Asia TEFL*, *15*(2), 276–291. https://doi.org/10.18823/asiatefl.2018.15.2.2.276

McGillycuddy, M., Warton, D.I., Popovic, G., & Bolker, B.M. (2025). Parsimoniously fitting large multivariate random effects in glmmTMB. *Journal of Statistical Software*, *112*(1), 1–19. doi:10.18637/jss.v112.i01.

Pan, F., Reppen, R., & Biber, D. (2020). Methodological issues in contrastive lexical bundle research: The influence of corpus design on bundle identification. *International Journal of Corpus Linguistics*, *25*(2), 216–230. https://doi.org/10.1075/ijcl.19063.pan

R Core Team. (2024). R: A language and environment for statistical computing (Version 4.3.3) Computer software. R Foundation for Statistical Computing. https://www.R-project.org/

Ruan, Z. (2017). Lexical bundles in chinese undergraduate academic writing at an english medium university. *RELC Journal*, *48*(3), 327–340. https://doi.org/10.1177/0033688216631218

Sánchez Hernández, P. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, *12*(S1), 187–209. https://doi.org/10.35360/njes.281

Sawaguchi, R. (2024). Potential of L1 and L2 corpora to identify target lexical bundles for argumentative essay writing. *Asia Pacific Journal of Corpus Research*, *5*(1), 1–21. https://doi.org/10.22925/APJCR.2024.5.1.1

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, *11*(4), 393–412. https://doi.org/10.1177/1362168807080960

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

Vo, S. (2019). Use of lexical features in non-native academic writing. *Journal of Second Language Writing*, *44*, 1–12. https://doi.org/10.1016/j.jslw.2018.11.002

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

Yan, H. (2019). I think we should…: Investigating lexical bundle use in the speech of english learners across proficiency levels. *International Journal of Translation, Interpretation, and Applied Linguistics*, *1*(2), 1–16. https://doi.org/10.4018/IJTIAL.2019070105

Appendix

Appendix 1. Top 20 Non-Prompt Bundles for each Sub-Corpus

Rank	J-SM	J-WE	ENS-SM	ENS-WE
1	I agree with this	have a part-time job	smoking should be banned	have a part-time job
2	agree with this statement	at all the restaurants	have a part-time job	to have a part-time
3	I agree with the	to have a part-time	a part time job	for college students to
4	have a part-time job	banned at all the	that smoking should be	students to have a
5	I think it is	completely banned at all	should be banned in	having a part-time job
6	I disagree with this	all the restaurants in	I don’t think	important for college students
7	with this statement because	the restaurants in the	to have a part-time	college students to have
8	agree with the statement	restaurants in the country	have a part time	smoking should be banned
9	is not good for	important for college students	students to have a	I think that it
10	I don’t agree	be completely banned at	to be able to	that smoking should be
11	it is important for	for college students to	be banned in all	I don’t think
12	I don’t like	is important for college	having a part-time job	think that it is
13	smoking is bad for	it is important for	college students to have	a part-time job is
14	I have two reasons	should be completely banned	for college students to	is important for college
15	disagree with this statement	college students to have	banned in all restaurants	the restaurants in Japan
16	agree with this opinion	students to have a	do n’t think it	I think it is
17	I agree this statement	smoking should be completely	important for college students	it is important for
18	do n’t like smoking	I think it is	to have a part	all the restaurants in
19	should be completely banned	that smoking should be	should have a part-time	do n’t think that
20	do n’t agree with	I agree with the	a part-time job while	should be banned in

THE AUTHORS

Trevor Sitler is currently pursuing a Ph.D. in foreign language education at Kansai University. He holds an MA in TESOL from the University of Birmingham and has been teaching and living in the Kansai region of Japan for a decade. He is currently working as an adjunct lecturer at Kindai University, Ryukoku University, and Ritsumeikan University.

Martin Spivey is an adjunct lecturer of EAP at Akita University and Akita International University in Tohoku, Japan. He holds an MA in TESOL from the University of Birmingham and has two decades of experience as an English foreign language instructor. His research interests include learner corpora, data-driven learning, and corpus-assisted discourse analysis.

THE AUTHORS' ADDRESSES

First and Corresponding Author
Trevor Sitler
Ph.D. Student
Kansai University
3-chōme-3-35 Yamatecho, Suita, Osaka, JAPAN
Email: sitlert@yahoo.com

Co-author
Martin Spivey
Assistant Professor
Akita University
1-1 Tegatagakuenmachi, Akita, JAPAN
Email: mart.spiv@gmail.com

Received: 25 May 2025
Received in Revised Form: 5 August 2025
Accepted: 17 August 2025