




Article

The Impact of Lexical Bundle Length on L2 Oral Proficiency

Dan Hougham ^{1,*} , Jon Clenton ² , Takumi Uchihara ³ and George Higginbotham ⁴ 

¹ Institute for Foreign Language Research and Education, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Hiroshima, Japan

² Graduate School of Humanities and Social Sciences, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8521, Hiroshima, Japan; jclenton@hiroshima-u.ac.jp

³ Graduate School of International Cultural Studies (GSICS), Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576, Miyagi, Japan; takumi.uchihara.a2@tohoku.ac.jp

⁴ Department of Social System Design, Eikei University of Hiroshima, 1-5, Nobori-cho, Naka-ku, Hiroshima 730-0016, Hiroshima, Japan; george@eikei.ac.jp

* Correspondence: hougham@hiroshima-u.ac.jp

Abstract: Lexical bundles (LBs) are crucial in L2 oral proficiency, yet their complexity in terms of length is under-researched. This study therefore examines the relationship between longer and shorter LBs and oral proficiency among 150 L2 learners of varying proficiency levels at a UK university. Through the analysis of oral presentation data (scores ranging from intermediate to advanced) and employing a combined text-internal and text-external approach (two- to five-word bundles), this study advances an innovative text-internal LB refinement procedure, thus isolating the unique contribution of LB length. Robust regression, dominance analysis, and random forest statistical techniques reveal the predictive power of bigram mutual information (MI) and longer three-to-five-word sequences on higher proficiency scores. Our results show that learners using higher MI score bigrams tend to perform better in their presentations, with a strong positive impact on scores ($b = 14.38$, 95% CI [8.01, 20.76], $t = 4.42$; dominance weight = 58.63%). Additionally, the use of longer three-to-five-word phrases also contributes to better performance, though to a lesser extent (dominance weight = 18.80%). These findings highlight the pedagogical potential of a nuanced approach to the strategic deployment of LBs, particularly bigram MI, to foster oral proficiency. Suggestions for future LB proficiency research are discussed in relation to L2 speech production models.



Citation: Hougham, Dan, Jon Clenton, Takumi Uchihara, and George Higginbotham. 2024. The Impact of Lexical Bundle Length on L2 Oral Proficiency. *Languages* 9: 232. <https://doi.org/10.3390/languages9070232>

Academic Editor: Claudine Bowyer-Crane

Received: 13 March 2024

Revised: 3 June 2024

Accepted: 18 June 2024

Published: 26 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: oral proficiency; multiword sequences; lexical bundles; learner corpus research; dominance analysis; random forests analysis; multiple regression analysis

1. Introduction

In recent years, there has been a growing interest in studying multiword sequences (MWSs) in second language (L2) oral proficiency. Foundational research by [Pawley and Syder \(1983\)](#) emphasized the critical role of lexical phrases, or “chunks”, in achieving native-like fluency. They argued that a large part of native speakers’ fluency stems from their use of these prefabricated chunks, which reduce the cognitive load during speech production by allowing speakers to retrieve whole phrases from memory rather than constructing sentences word-by-word. This process, known as chunking, enables more fluent and efficient language use and is crucial for effective communication.

Building on these insights, subsequent research has underscored the significance of MWSs in L2 learning and speaking proficiency ([Schmidt 1992](#); [Nation 2013](#)). Lexical bundles (LBs), defined as frequent, contiguous MWSs, have gained particular attention for their role in speech fluency ([McGuire and Larson-Hall 2021](#); [Tavakoli and Uchihara 2020](#)) and oral proficiency ([Garner and Crossley 2018](#); [Kyle and Crossley 2015](#); [Zhang et al. 2021](#)). The findings from these studies suggest that incorporating MWS-focused activities in language teaching can significantly enhance students’ speaking skills. Despite this, previous studies have predominantly focused on shorter LBs (typically two- and three-word sequences),

thus potentially limiting our understanding of learners' phraseological knowledge and its impact on oral proficiency. Exploring this gap by examining both longer (e.g., *as you can see*) and shorter (e.g., *there are*) LBs is critical, as longer LBs could offer enhanced proficiency and show processing efficiencies that are essential for fluent speech production. Responding to calls from recent literature (e.g., Hougham et al. 2024; Tavakoli and Uchihara 2020), our study addresses this gap by investigating the predictive power of both shorter and longer LB usage on oral presentation scores among learners with varying proficiency levels. We aim to isolate the unique contributions of longer LBs using an innovative text-internal and text-external approach, thus providing a more comprehensive understanding of how LB length influences L2 oral proficiency. Addressing this gap is crucial for advancing our theoretical understanding of phraseological knowledge in L2 learners and for developing practical teaching strategies that leverage the benefits of various types of LBs to enhance oral proficiency.

2. Literature Review

2.1. Multiword Sequences and Lexical Bundles

Multiword sequences include various combinations of words that commonly co-occur, such as idioms (e.g., *hit the nail on the head*), collocations (e.g., *heavy rain*), phrasal verbs (e.g., *look up*), proverbs (e.g., *make hay while the sun shines*) and LBs (e.g., *in terms of*) (Garner and Crossley 2018). These sequences can be identified through either a phraseological approach or a frequency-based approach (Granger and Paquot 2008; Nesselhauf 2004). The phraseological approach categorizes MWSs using linguistic criteria (Cowie 1981), which are often based on the intuitions of first language (L1) speakers. Previous studies employing this approach (e.g., Boers et al. 2006) frequently rely on human raters to assess the formulaic nature of MWSs, but this can result in low inter-rater reliability. For example, Boers et al. (2006) reported a reliability coefficient of less than 0.60, which is significantly lower than the median inter-rater reliability of 0.92 reported in SLA research by Plonsky and Derrick (2016).

On the other hand, the frequency-based approach, developed by Sinclair (1991) and Nesselhauf (2004), uses corpus-based automated extraction techniques to identify frequent word combinations based on quantitative criteria such as frequency and range. These are often termed lexical bundles or n-grams and can include both structurally complete (e.g., *at the end of the day*) and incomplete sequences (e.g., *in the middle of*), regardless of their idiomaticity or structural status (Biber et al. 1999). The objective nature of the frequency-based approach has made it popular in learner corpus research (Paquot and Granger 2012). For instance, Ebeling and Hasselgård (2015) emphasized the benefits of this approach, such as accessing a large quantity of data for quantitative analysis.

In our study, we use the term LB to refer to sequences identified using automatic extraction software based on criteria like minimum frequency, range, and mutual information (MI) scores, which measure the strength of association between word pairs. Our aim is to distinguish LBs from other MWSs, even though there may be some conceptual overlap (e.g., high MI score bigrams like *global warming* can also be categorized as collocations).

In learner corpus research, LB production can be analyzed using either text-internal or text-external approaches. The text-internal approach (e.g., Biber and Gray 2013) focuses on data within learner corpora. In contrast, the text-external approach (e.g., Tavakoli and Uchihara 2020) examines LBs in terms of criteria like the frequency of co-occurrence in external reference corpora consisting of L1 speaker language. Both approaches have their limitations such as the arbitrary frequency cutoffs in the text-internal approach (Myles and Cordier 2017) and the challenges in measuring longer bundles with text-external analysis tools (e.g., some software programs such as the Tool for the Automatic Analysis of Lexical Sophistication: TAALES (Kyle et al. 2018) can only analyze bigrams and trigrams). Additionally, text-external techniques may not ensure that frequently occurring word combinations in external corpora are psycholinguistically relevant to the learners being

studied (Ellis et al. 2009; Myles and Cordier 2017). Our study employs both approaches to assess the contributions of shorter and longer LBs with respect to L2 oral proficiency.

2.2. From Theory to Empirical Findings

Pawley and Syder (1983) lay foundational insights into the importance of multiword sequences in language proficiency through their exploration of lexical phrases. They propose that a speaker's command of fixed and semifixed phrases, or "chunks" of language, is crucial for achieving native-like fluency. Such lexical phrases ease the cognitive load during speech production and serve as essential building blocks for fluent and idiomatic language use. This early perspective is significant, as it suggests that the repertoire of lexical phrases is a key factor distinguishing more proficient speakers from less proficient ones, thereby enabling more efficient processing and production of language. Building upon the understanding of lexical phrases, Levelt's (1989) speech production model provides further theoretical justification for the link between LBs and proficiency. The model describes three stages in speech production: conceptualization, formulation, and articulation. During the conceptualization stage, speakers plan the content of their speech. In the formulation stage, they encode lexical and grammatical items in the mental lexicon, activating appropriate lemmas and constructing syntactic structures. Finally, the articulation stage involves implementing the phonetic plan, resulting in speech production. Originally designed for L1 speakers, Kormos (2006) adapted this model to address the specific challenges of L2 speakers, such as a smaller and less structured mental lexicon with fewer formulaic expressions. The model suggests that speakers with a larger repertoire of MWSs can retrieve these sequences as easily as single words during the formulation stage, which reduces cognitive load and provides a processing advantage. This efficiency allows L2 learners to allocate cognitive resources to other aspects of speech production, such as lexical and grammatical accuracy or complexity (Kormos 2006; Skehan 2014). Utilizing MWSs can boost oral fluency, particularly during lexical selection at the formulation stage (Kormos 2006; Levelt 1992). In contrast, speakers with a limited MWS repertoire may struggle, as they expend more cognitive resources when retrieving individual lexical items during the formulation stage. Alongside these frameworks, usage-based theory, as articulated by Ellis (2002) and supported by Bybee (2006), further complements our understanding of how language proficiency develops from the experience and frequency of usage. Ellis (2002) introduces the notion that linguistic competence is significantly influenced by the accumulation of exposure to MWSs, which reinforces their retrieval and production. This perspective aligns with the proceduralization of language use, where proficiency emerges from the frequent and familiar use of language structures. Bybee suggests that repeated exposure to MWSs not only facilitates their entrenchment in the learner's mental lexicon but also highlights the importance of frequency and familiarity for L2 learning. Together, these theoretical perspectives—Pawley and Syder's lexical phrases theory, Levelt's speech production model, and the usage-based theories proposed by Ellis and Bybee—form a multifaceted view of the mechanisms underlying L2 oral proficiency. The current study aims to bridge these theoretical insights with empirical evidence, exploring how the length, frequency, and association strength of LBs contribute to oral proficiency. We seek to illuminate the relationship between LB use and language proficiency, grounded in a broad spectrum of linguistic theory.

To date, most learner corpus-based studies have focused on LBs used in writing rather than speaking (e.g., Appel and Wood 2016; Siyanova-Chanturia and Spina 2020; Staples et al. 2013). This emphasis can be attributed to two main reasons. First, natural speech within L2 classes/tests is often not recorded, whereas L2 students regularly submit written work, thus making written corpora more readily available for analysis. Second, spoken corpora require more time to analyze, as there are additional steps in processing. Specifically, the speech first needs to be transcribed into a text form, and it may need to be divided into speech units for ease of analysis, thus making the analysis of spoken language more labor-intensive and time-consuming than written texts. Given these constraints, most research

has naturally gravitated towards written corpora. However, the findings from studies focusing on written language—highlighting the complexity of the MWS–L2 proficiency link and trends regarding the use and proficiency-related evolution of LBs—underscore a rich area for exploration within spoken language as well. Specifically, the research exploring written LBs has revealed general trends, including the following: (a) the quantity of LBs produced by learners decreases as proficiency increases (Appel and Wood 2016; Staples et al. 2013) or with time spent in an English-speaking country (Groom 2009). Less proficient learners overuse bundle tokens and under-use bundle types, which is a trend resembling the use of “phraseological teddy bears” (i.e., overusing high-frequency phrases with which one feels comfortable, as demonstrated in Hasselgård 2019). Research has shown that proficient learners have a firmer and more creative command of lower-frequency bundles whose constituent words are nonassociated (Siyanova-Chanturia and Spina 2020); and (b) less proficient learners rely more on LBs copied from writing prompts or source texts due to their relatively limited lexical repertoires (e.g., Appel and Wood 2016; Staples et al. 2013). Overall, these studies suggest that much remains to be understood about the relationship between LB use and general writing proficiency. While most existing learner corpus-based studies have focused on the use of LBs in written contexts, the insights they provide into the relationship between MWS–L2 proficiency potentially apply to the spoken domain as well. The understanding that learners’ reliance on LBs decreases with proficiency or extended exposure to an English-speaking environment, and that proficiency dictates a learner’s command over lower-frequency bundles, offers a compelling framework to consider for spoken data. The current study seeks to bridge this gap by exploring how these patterns manifest in speech proficiency, which has been less frequently examined. There is potential that the dynamics of LB use in writing, as detailed in these studies, could find parallels in oral production, thereby providing richer insights into speech proficiency and its intricacies.

Several recent studies have examined the link between LB use and general speaking proficiency from a text-external perspective, focusing on the extent to which L2 learners use L1 target-like two- and three-word (bi- and trigram) measures in terms of quantitative indices (frequency, proportion, and association) (e.g., Garner and Crossley 2018; Kyle and Crossley 2015; Zhang et al. 2021). The findings from these studies differ from the typical trend of writing-centered studies (which proposed that less-proficient L2 writers (over)use a greater number of high-frequency bundles). Kyle and Crossley (2015) found significant correlations between (human-rated) oral proficiency scores and several n-gram scores, of which the strongest predictor of speaking proficiency came from high-frequency trigrams, thus suggesting that more skillful L2 speakers use a larger number of highly frequent trigrams. Garner and Crossley (2018) found that beginning-level L2 learners showed the greatest increase in oral production of high-frequency bigrams over the course of their four-month longitudinal study. Zhang et al. (2021) reported that several n-gram measures (e.g., bigram proportion and association: MI and t scores) significantly correlated with (human-rated) oral proficiency scores on story retelling and monologic tasks. Such studies highlight the important role of proficiency in the development of LB use but suggest that further research is required to bring clarity to this research area.

Relatively few studies (e.g., Biber and Gray 2013; De Cock 2004) have examined LB use in spoken corpora from a text-internal perspective. De Cock (2004) examined two- to six-word bundle use among advanced EFL learners compared to L1 speakers. She found that learners’ preferred bundles were less interactional and included relatively few vagueness markers (e.g., *or something, kind of*) compared to L1 speakers. Biber and Gray’s (2013) study of spoken and written responses to the TOEFL iBT showed a slightly more complex pattern than other n-gram studies. They reported that intermediate-level participants produced a greater number of bundles (four-word units) than their lower- and higher-proficiency groups, thus suggesting a general developmental progression in which lower-level participants use a smaller number of bundles, intermediate-level participants overuse a larger number of bundles, and high-scoring participants show greater control and creativity in using the bundles they have acquired (p. 37). To summarize, these studies

suggest a complex perspective and a need for further research on how language use varies among individuals from diverse backgrounds.

While such studies have offered insights into the patterns of LB use in spoken corpora, a broader context emerges when we consider the significant relationship between MWS use and speech fluency, and it is clear across various teaching contexts. Studies that show significant and positive relationships between MWS use and speech fluency (including L2 proficiency in the broader sense) come from a range of different teaching contexts (e.g., Boers et al. 2006; Hougham et al. 2024; McGuire and Larson-Hall 2017, 2021; Stengers et al. 2011; Suzuki et al. 2022; Tavakoli 2011; Tavakoli and Uchihara 2020; Uchihara et al. 2021; Wood 2009, 2010). Boers et al. (2006) and Stengers et al. (2011) found strong links between the number of MWSs used (in story retelling tasks) and (perceived) oral ability scores in a Belgian EFL context. Wood (2009) examined the effect of MWS-focused teaching on MWS use and oral fluency in a case study ($N = 1$) in a Canadian ESL context. Wood found that MWS-focused instruction can lead to increased MWS use and increased spoken fluency over a short period (six weeks). Wood (2010) also found similar results with a slightly larger sample size ($N = 11$) in a similar context over a longer period (six months). Tavakoli (2011) compared the pausing patterns of L1 versus L2 speakers' performance in a UK university context. She found that L2 learners rarely paused in the middle of multi-word units, thus providing further corroborating evidence that lexical chunks facilitate fluency. Similarly, Uchihara et al. (2021) found that speakers who provided more low-frequency MWS (collocational type) responses to a word association task (Lex30) spoke more rapidly with fewer silent pauses. McGuire and Larson-Hall (2017) replicated Wood's (2009) study in an American ESL study abroad context. They reported a moderately strong relationship between all participants' MWS use and fluency measures. Tavakoli and Uchihara's (2020) study, reporting the link between two- and three-word LBs and one objective measure from each aspect of utterance fluency (speed, breakdown, and repair) across assessed proficiency levels in a UK university context, represents the first systematic study of its kind. Tavakoli and Uchihara reported that greater LB use (a larger proportion of frequent LBs and more frequent LBs) was positively and significantly related to higher speaking ability scores and with some fluency aspects (faster articulation rate and fewer pauses within clauses). Suzuki et al.'s (2022) task repetition intervention study examined the use of single words and trigrams on speed, breakdown, and repair fluency aspects. They found that the recycling of more complex MWSs through task repetitions seemed to facilitate proceduralization (i.e., more efficient retrieval of MWSs), but they also found that such reuse had both positive and negative influences on midclause pauses specifically, as well as fewer but longer pauses within clauses, which may show that learner encoding systems were in the process of restructuring. Hougham et al. (2024) examined the relationship between both shorter (bi- and trigram) and longer (four-to-five-word) LBs and three dimensions of fluency (speed, breakdown, and repair) using both text-internal and text-external techniques. They found that using longer LBs, specifically four-to-five-word sequences of high collocational quality (those with high MI scores), significantly enhanced speech fluency by reducing the frequency of pauses and repairs. Moreover, they uncovered a correlation between frequent combinations of two words and a faster rate of speech, whereas complex combinations (those with high mutual information scores) were found to slow down speech.

2.3. Gaps and Unexplored Areas in Previous Research

While the studies reviewed here support the hypothesis of a positive relationship between MWS use and proficiency, they might be limited in at least six important ways. Foremost, studies focusing on relatively short (two- and/or three-word) sequences might not fully capture learners' actual phraseological knowledge and how it relates to oral fluency units. This is exemplified by multiple studies (e.g., Garner and Crossley 2018; Kyle and Crossley 2015; Kyle et al. 2018; McGuire and Larson-Hall 2021; Suzuki et al. 2022; Tavakoli and Uchihara 2020; Zhang et al. 2021). For example, using longer LBs might be more beneficial for improving aspects of oral fluency and increasing high-stakes assess-

ment scores. Emphasizing the potential importance of longer LBs, Tremblay et al. (2011) demonstrated that longer (four- and five-word) LBs offer online processing advantages over non-LBs in receptive tasks. Hougham et al. (2024) also showed that longer LBs of high collocational quality enhanced various aspects of fluency. Despite this understanding, the effects of longer linguistic units on achieving fluency in speech production have not been fully explored. Given what we know from Pawley and Syder's insights into the significance of lexical phrases, the comprehensive framework provided by Levelt's speech production model, as well as the principles of usage-based theory by researchers such as Bybee and Ellis, it is reasonable to hypothesize that employing longer LBs can lead to enhanced processing efficiency. Second, many previous studies have had methodological or contextual limitations, such as measuring MWSs subjectively using a criteria checklist and L1 speaker intuition (e.g., McGuire and Larson-Hall 2017; Wood 2009). Third, most previous research is restricted to investigating the MWS proficiency link with a learner-external approach (i.e., examining learners' use of selected sequences that are thought to be formulaic in L1 speaker English and identified in advance as formulaic or quantifying a text's formulaicity by checking the frequency of all of its constituent word sequences against an external reference corpus) (e.g., Garner and Crossley 2018; Tavakoli and Uchihara 2020; Zhang et al. 2021). Only one study by Hougham et al. (2024) has systematically attempted to employ both text-external and text-internal methods to analyze learner-produced LBs in relation to aspects of oral fluency, but their study did not examine oral proficiency scores. Fourth, some previous studies have suffered from small sample sizes (e.g., $N = 19$ in McGuire and Larson-Hall 2017; $N = 1$ in Wood 2009; $N = 11$ in Wood 2010). Fifth, the LBs examined in studies based on learner corpora frequently have varying lengths, typically ranging from two to six words. Additionally, the criteria for extracting these LBs, such as frequency and dispersion, differ significantly across different studies. Therefore, it is important to view the aforementioned findings and general patterns as hypotheses that require further testing using alternative corpus data within diverse contexts across various proficiency levels.

3. The Current Study

Informed by prior theoretical insights (e.g., Levelt 1989; Pawley and Syder 1983) and responding to the call for a more comprehensive approach to MWSs in EFL research (Hougham et al. 2024; Tavakoli and Uchihara 2020), our study explores the under-investigated area of longer LB usage, hypothesizing that these can offer significant processing advantages for L2 speakers, which is a notion previously suggested but not empirically tested across a range of proficiency levels. Our research question is designed to directly address the identified need for more comprehensive analyses of LB usage in relation to proficiency scores. This is done by building on and extending the work of Tavakoli and Uchihara (2020), as well as Hougham et al. (2024). We investigate the relationship between the use of shorter (bi- and trigrams) and longer (three-to-five-word) LBs and oral proficiency scores. By comparing and contrasting findings across different learner populations and proficiency scores, the current study seeks to contribute to a more detailed understanding of how LBs function in L2 speech production models and inform future LB proficiency research directions. By examining a broader dataset (150 L2 learners at varying proficiency levels from a UK university's preessional course) and using an innovative text-internal LB refinement procedure that identifies more structurally complete and useful LBs, the current study allows us to explore the relationship between LB usage and oral proficiency, aiming to provide a more comprehensive understanding of the relationship. Additionally, the current study seeks to explore the extent to which both shorter and longer LBs can predict speaking proficiency scores. By employing robust regression, dominance analysis, and random forest techniques, we not only aim to validate previous findings but also uncover new patterns, potentially leading to more effective pedagogical strategies for enhancing oral proficiency in diverse L2 learning contexts.

The current study addresses the following research question: To what extent do longer lexical bundles (three-to-five-word units) predict oral proficiency scores compared to shorter lexical bundles (bi- and trigrams) in L2 learners?

Based on the findings in [Tavakoli and Uchihara \(2020\)](#) and the theoretical frameworks of speech production (e.g., [Kormos 2006](#)), the current study hypothesizes that longer LBs will significantly contribute to higher oral proficiency scores, potentially more so than shorter LBs, due to their ability to enhance processing efficiency and fluency.

4. Materials and Methods

4.1. Participants

The participants were 150 language learners taking a preessional course at a UK University. They were from 20 different L1 backgrounds (outlined in [Table 1](#)), with most of the participants being L1 Chinese ($n = 101$), L1 Saudi Arabian ($n = 9$), or L1 Turkish ($n = 9$). Participants’ raw presentation scores (described in the next section) were first converted into IELTS bands ranging from 6.5 to 7.5. Next, they were categorized into three groups depending on these bands. From a larger pool, we randomly selected 50 individuals for each level, ensuring balanced representation across the three bands. There were 50 participants at IELTS level 6.5, 50 at IELTS level 7, and 50 at IELTS level 7.5 (see [Table 2](#)).

Table 1. Participants according to L1 backgrounds ($N = 150$).

Nationality	<i>n</i>	Nationality	<i>n</i>	Nationality	<i>n</i>
Chinese	101	Indonesian	2	Taiwanese	1
Saudi Arabian	9	Bangladeshi	1	Chilean	1
Turkish	9	German	1	Egyptian	1
Japanese	7	Bahraini	1	Georgian	1
Thai	6	Kuwaiti	1	Pakistani	1
Indian	2	Ghanian	1	Russian	1
Colombian	2	Italian	1		

Table 2. Participants’ IELTS-rated speaking proficiency levels.

IELTS speaking level	6.5	7	7.5	Total
<i>n</i>	50	50	50	150

4.2. Oral Presentation Tasks and Proficiency Scores

All participants completed oral presentations comprising a 7 min presentation (monologue) in small groups with the aid of PowerPoint slides as part of the preessional course. It is important to note that these presentations were not conducted specifically for the purposes of the current study but were completed as a requirement of the preessional course at Queen Mary University of London. Their outputs were video recorded online and assessed by teachers using grading descriptors (in equal measure): presentation content, presentation structure, seminar leadership, language fluency, and language accuracy (see the full descriptors in [Appendix A](#)). The oral presentation tasks were designed to cover a wide range of subtopics under the overarching theme of globalization, which had been the focal point of the participants’ 5-week academic English course. While globalization served as the central theme, participants were encouraged to explore this broad topic through various lenses, ranging from economic and legal perspectives (e.g., how globalization affects development and international trade law) to more specialized areas (e.g., acoustic telemetry in fisheries). This diverse range of subtopics allowed participants to delve into areas aligned with their academic interests and expertise. The students were from a wide range of disciplines: humanities, law, science, technology, engineering, and mathematics majors. From the video recordings, we took 3 min speech samples starting at the 30 s timestamp. As the 3 min sections of speech analyzed came from the “presentation” part of

their seminar, we are specifically analyzing monologic oral presentations rather than freer dialogic speech. These samples were transcribed using Sonix.ai web-based software, and the transcriptions were checked for accuracy by a research assistant and double-checked by the first researcher. We used the transcripts in the lexical analyses (described in detail below), and we used the raw scores and the banded IELTS scores as the measures of oral proficiency in the current study.

4.3. Measuring Lexical Bundles: A Two-Pronged Approach

For the current study, we adopt a frequency-based approach using both text-internal and text-external techniques to isolate the unique contribution of shorter versus longer LBs.

4.3.1. Text-External Lexical Bundle Analysis and Measures

Following previous studies (Garner and Crossley 2018; Hougham et al. 2024; Tavakoli and Uchihara 2020), we used three n-gram indices (proportion, frequency, and association) to objectively measure the use of shorter LBs, specifically two- and three-word contiguous sequences (i.e., bi- and trigram tokens) in our learner corpus. TAALES version 2.0 was used to calculate three kinds of n-gram scores, producing six score indices (two proportion, two frequency, and two association indices). As our external reference corpus, we chose the spoken subsection of the Corpus of Contemporary American English (COCA Davies 2009), which comprises 79 million words from transcriptions of a wide range of TV and radio programs. Our choice of this spoken corpora was in alignment with research findings showing a gap in L2 learners' spoken and written vocabulary sizes (Uchihara and Harada 2018) and differences in lexical profiles between spoken and written modes (Dang et al. 2017). We maintained consistency between the modality in which L2 words were elicited and the modality of the reference corpus based on the practices used in previous studies (e.g., Uchihara and Clenton 2020; Uchihara et al. 2021).

In our current study, we use proportion score indices to measure the occurrence of bi- and trigrams in our learner speech sample data. These bi- and trigrams are also among the 30,000 most frequent ones in the external reference corpus (COCA). Higher proportion scores show that participants in the sample produced a higher percentage of high-frequency, target-like bi- and trigrams. Higher frequency scores show that participants in our sample produced a larger number of high-frequency target-like bi- and trigrams. Logarithmic bi- and trigram scores, instead of raw frequency scores, were used to control for Zipfian effects common in word frequency lists (Kyle and Crossley 2015; Tavakoli and Uchihara 2020). Association score indices measure the association strength between individual words within bigrams and trigrams. Of the five association measures available in TAALES, the one association measure we used was mutual information score.¹ MI score measures the strength of association between two words. MI scores show the strength of word associations, with higher scores suggesting stronger associations. However, MI also focuses on word pairs that are not commonly found together (Schmitt 2010, p. 130). Before n-gram analysis using TAALES, all the transcripts were cleaned by correcting any misspellings and mispronunciations and removing any markings of filled pausing (i.e., *ums*, *uhs*, etc.). The resulting transcripts ranged between 216 and 436 words ($M = 310.78$, $SD = 47.32$).

4.3.2. Text-Internal Lexical Bundle Identification and Refinement Procedures

We adopted a text-internal approach to isolate and measure the unique contribution of longer LBs to proficiency. As a first step, we conducted frequency analyses using AntConc (Anthony 2022) to generate lists of the most frequently used four-word LBs in the learner corpus. The frequency and dispersion thresholds used to identify lexical bundles vary from study to study. Figures used for "frequency cut offs are somewhat arbitrary" (Hyland 2008, p. 8) depending on both the size and specificity of the corpus. For relatively small spoken corpora like the one in this study, a raw cutoff frequency has often been used, ranging from two to ten occurrences (e.g., Altenberg 1998; Biber and Barbieri 2007; De Cock 1998). Given the small size of the spoken corpus in the current study

(46,617 words), for four-word combinations to qualify as lexical bundles, we used a cutoff point of three or more occurrences in at least three texts, following [Biber and Barbieri \(2007\)](#). These minimum figures help to ensure that the identified bundles are not idiosyncrasies confined to occurrences produced by an individual speaker. This resulted in 447 instances of four-word LBs that met these criteria. To deal with the issues of overlap and structural incompleteness among these instances, we used a refinement procedure developed by previous researchers ([Wood and Appel 2014](#)) aiming to identify more structurally complete and useful LBs. We split each four-word sequence (e.g., *I will give you*) into two constituent three-word clusters (e.g., *I will give*, *will give you*). The frequency of the two three-word clusters in the corpora were identified and compared. If the frequency of one three-word cluster was at least double the frequency of the other, the more frequent cluster was classified as the root structure and the fourth word was considered as a word that commonly occurred with that structure and was put in parentheses. For example, *I will give* occurred over two times (freq = 41) than *will give you* (freq = 20) in the current data set. Therefore, the final resultant structure is in this case: *I will give (you)*. Another example from the current data set is the sequence *the first part is*. Since *the first part* (freq = 41) occurred over two times more frequently than *first part is* (freq = 17), the final resultant structure is *the first part (is)*. The refinement process produced a list of 119 multiword structures, primarily comprising core three-word phrases and four-word structures, with a smaller number of longer five-word structures included. Table 3 shows the number of three-word (55), four-word (58) and five-word structures (6) identified. The examination of the extensive list provided in Appendix B shows that many of the resulting three-word structures are self-contained units in terms of semantics or structure. For instance, *with the development of* forms a complete unit. This observation suggests that the refinement procedure successfully pinpointed additional core structures.

Table 3. Numbers and examples of structures identified at different lengths.

Lexical bundle length	three-word units	four-word units	five-word units	Total
Number of units	55	58	6	119
Examples	<i>I'm going to the first part I will give</i>	<i>I would like to at the end of at the same time</i>	<i>if you have any questions the presentation will be brief in a socially responsible way</i>	

Note: Contracted forms (e.g., *I'm*) were counted as a single word; for example, *I'm going to* was considered to be a trigram rather than a quadgram.

It is crucial to differentiate the newly identified structures from traditional LBs (and other multiword structures) when describing them, as this refinement procedure goes beyond the usual LB approach. Recall that the traditional LB approach strictly identifies multiword sequences within two parameters: frequency and range. For this reason, when making modifications or refinements to traditional LB methods, previous researchers (e.g., [Simpson-Vlach and Ellis 2010](#); [Wood and Appel 2014](#)) have used different terminology to refer to such refined or modified LBs. [Simpson-Vlach and Ellis \(2010\)](#), for instance, started with an LB approach and then applied additional criteria (e.g., human ratings of formula teaching worth combined with MI score as a measure of collocation strength), thus aiming to identify more useful multiword units for teaching purposes. [Simpson-Vlach and Ellis \(2010\)](#) adopted the general term “formulaic language” to describe the word combinations identified in their study. Another example is [Wood and Appel \(2014\)](#), who developed the LB refinement procedure used in the current study. Wood and Appel adopted the general term “multiword constructions” to refer to the refined LBs identified in their study. “Formulaic language” and “multiword constructions” have also been used as umbrella terms in other studies such as [Liu \(2012\)](#). Because of the fuzzy nature of boundaries between many types of semifixed multiword combinations, and because each study’s identification and refinement methods are different, there is no consensus in the literature as to which terms apply in all cases. Although it is challenging to pin down a consistent definition of

MWSs across all studies, most researchers agree that it is important to distinguish between different types of MWSs where possible and to report clearly how MWSs are identified in each study to facilitate comparisons of research findings across studies. In the current study, to avoid terminological confusion, we use the term “refined LBs” to refer to the refined LB list (see Appendix B) produced by the above-described refinement procedure.

4.4. Scoring the Use of Text-Internal Bundles

It is important to quantify usage of the list of refined LBs so that we can run various quantitative analyses (e.g., multiple regression) and compare the relationships between different types of LBs (unrefined text-external vs. refined text-internal; shorter vs. longer) with oral presentation scores. To do so, we awarded one point for each identified refined LB used by each participant. We tallied up the total number of points for each participant, arriving at a three-, four-, and five-word usage score for each participant. As for text-internal MI scores, we extracted MI scores for sequences of various lengths using the Collocate 2.0 software program (Barlow 2015). MI scores for three-to-five-word sequences have been used in several studies (e.g., Ellis et al. 2008; Simpson-Vlach and Ellis 2010), as they appear to offer a reliable indication of phrasal coherence. For each refined bundle used by each participant, we awarded the corresponding MI score. We then tallied all MI scores and gave each participant a total MI score.

4.5. Statistical Analyses

We analyzed six text-external n-gram (two- and three-word) measures and two text-internal refined n-gram (three-, four-, and five-word) measures. To examine our research question, that the use of LBs of varying length can predict oral presentation raw scores, we selected regression analysis. Initially, we examined the assumptions underlying regression models. The presentation raw scores variable yielded a Shapiro–Wilk p -value of 0.04, thus showing a significant deviation from a normal distribution. To identify influential outliers within the data, Cook’s distance was employed with a conservative threshold of $4/n$, thus facilitating the assessment of the impact of individual data points on the regression model. Several data points emerged as influential outliers, surpassing the Cook’s distance threshold. Their presence implies a potential influence on the estimated regression coefficients (see Supplementary Materials for detailed results of these checks). Subsequently, guided by Larson-Hall (2015, p. 264), we conducted a robust regression using MM estimation with the “rlm” function in the MASS package in R (R Development Core Team 2019). We chose the “rlm” function because it aptly accommodates data sets with non-normal distributions and outliers.

In an effort to make the results more robust, we decided it was relevant to conduct additional analyses on the relative importance of each predictor variable in explaining the variance in the model. We pursued this inquiry through a dominance analysis (DA) using the “calc.relimp” function from the “relaimpo” package in R (Grömping 2006). DA can effectively address correlations among predictor variables and can help in better understanding the unique contribution of each PV to the criterion variable in multiple regression analysis as opposed to relying solely on possibly misleading standardized beta coefficients (Mizumoto 2022). DA facilitates comprehension by computing dominance weights for each predictor, which show the mean impact of a variable on the predictability of all potential subsets of predictors, consequently presenting a thorough picture of the influence of each predictor on the outcome.

To achieve a more precise estimation of the importance of each variable in the multiple regression model, it is important to conduct dominance analysis in combination with random forests analysis (Mizumoto 2022). The random forests approach is a nonparametric machine learning model, meaning it can offer more precise outcomes when multiple regression assumptions are violated (Liakhovitski et al. 2010). Using random forests allows researchers to acquire a nuanced perspective on variable importance. Hence, following the guidelines by Mizumoto (2022), we integrated the random forests analysis using

the Boruta package in R (Kursa and Rudnicki 2010). The Boruta algorithm, specifically designed for feature ranking based on random forests, runs the random forests multiple times. It labels features (or predictors) as “confirmed”, “rejected”, or “tentative” based on their significance compared to randomized shadow features. “Confirmed” predictors are deemed significant, “rejected” ones are considered unimportant, and “tentative” labels are reserved for predictors whose importance remains uncertain. We show these using boxplots in the figures presented in the following section.

Integrating Boruta with robust regression and DA allows for a comprehensive analysis, where each method compensates for the limitations of the others. Robust regression ensures that our model is not unduly influenced by outliers, providing reliable coefficient estimates even when the data distribution is non-normal. DA helps clarify the relative importance of each predictor by showing their unique contributions to the model while addressing correlations among predictors. Meanwhile, Boruta offers a rigorous feature selection process that ranks predictors based on their importance and independently of the distributional assumptions. By using Boruta, we can validate which predictors are truly significant, providing a layer of verification to the results obtained from robust regression and DA. This combination of methods allows for triangulation, where the results from one method can support and validate the findings of the others, thus possibly enhancing the overall reliability and robustness of our conclusions. In what follows, we present the descriptive statistics first, we then report the robust regression model in combination with DA and random forests to address our research question.

5. Results

Table 4 shows the descriptive statistics for the presentation raw scores and different LB measures used in this study. The table includes the mean (*M*), standard deviation (*SD*), median, minimum, and maximum values for each measure. These statistics provide an overview of the distribution and central tendency of the presentation scores and n-gram measures, highlighting the variability and range of the data.

Table 4. Descriptive statistics for presentation scores and n-gram measures.

Measure	<i>M</i>	<i>SD</i>	Median	Minimum	Maximum
Presentation raw scores	62.71	5.53	63.00	53.00	78.00
Bigram log frequency	1.32	0.12	1.31	1.04	1.73
Bigram proportion	0.44	0.07	0.44	0.26	0.63
Bigram association MI	1.54	0.17	1.54	1.13	2.09
Trigram log frequency	0.68	0.13	0.67	0.41	1.07
Trigram proportion	0.12	0.04	0.12	0.04	0.26
Trigram association MI	2.53	0.25	2.54	1.91	3.21
Three-to-five-word usage	10.29	5.10	10.00	1.00	27.00
Three-to-five-word MI	132.39	73.57	117.97	19.87	399.56

Note: *M* = mean. *SD* = standard deviation. MI = mutual information.

A boxplot was created to visualize the distribution of presentation scores across different L1 groups (Figure 1). The boxplot provides a clear representation of the central tendency and variability within each group. Although a Kruskal–Wallis test did not find statistically significant differences in the presentation scores across L1 groups (Kruskal–Wallis chi-squared = 29.825, *df* = 19, *p*-value = 0.054), the boxplot offers valuable insights into the data distribution. It shows that most L1 groups have similar median scores, but there is considerable variability within some groups. The blue dots represent individual data points, highlighting the spread of scores within each group.

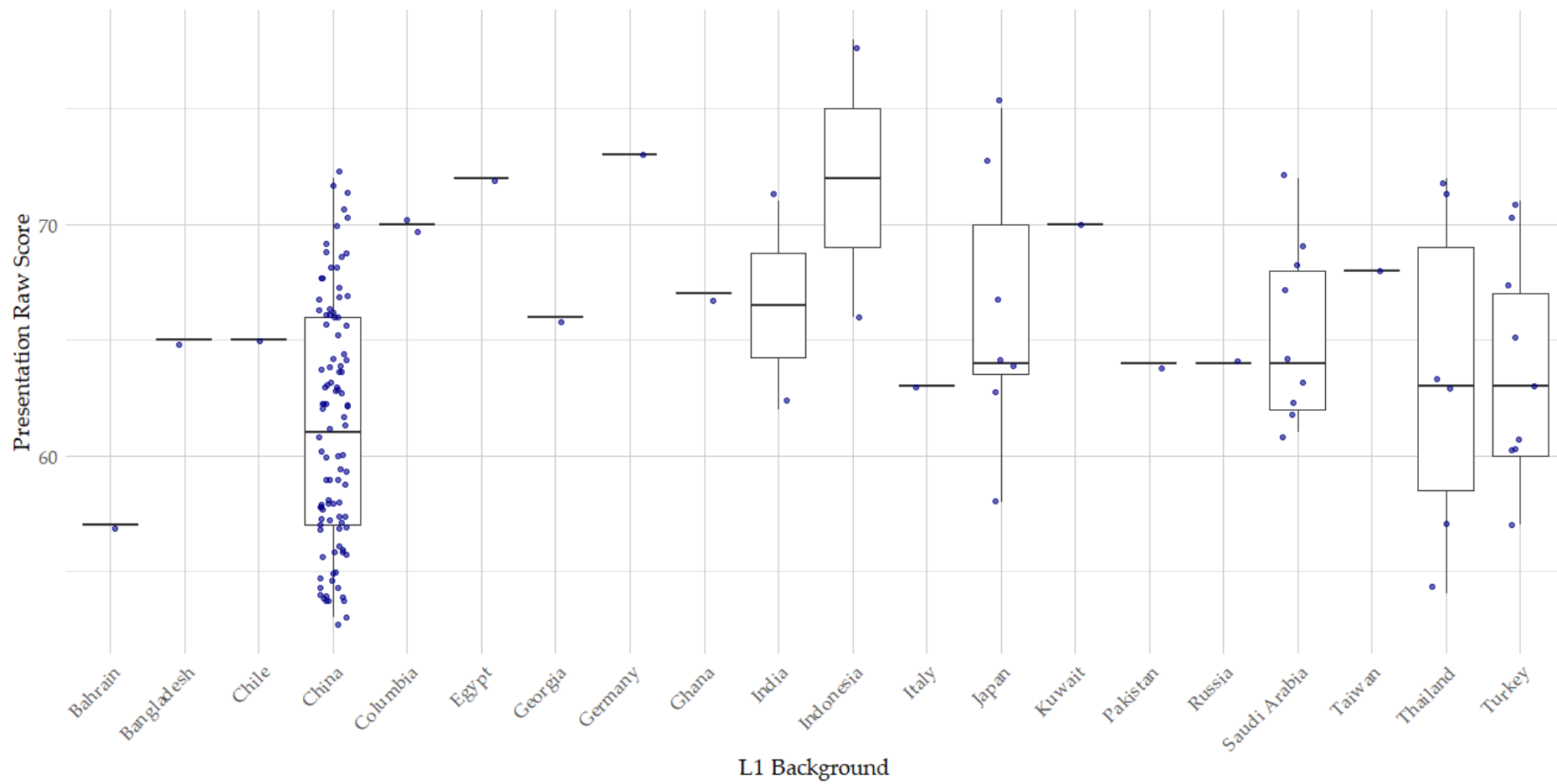


Figure 1. Presentation scores according to L1 background.

Table 5 shows the findings from the robust regression and dominance analysis with the criterion variable being presentation raw scores. The robust regression analysis unveils multiple predictors impacting the presentation raw scores. Bigram MI displayed a strong positive association ($b = 14.38$, 95% CI [8.01, 20.76], $t = 4.42$), accounting for a dominant 58.63% of the variance in presentation raw scores. This implies that learners using more bigrams with higher MI scores tend to score higher on their presentations, holding other variables constant. Examples of bigrams with high MI scores produced by high-scoring participants include *carbon dioxide*, *Kyoto Protocol*, and *global warming* (for more examples, see Appendix C). Three-to-five-word usage also showed a significant positive influence on the model ($b = 0.53$, 95% CI [0.06, 0.99], $t = 2.22$), contributing to 18.80% of the variance in the presentation raw scores. This suggests that individuals employing specific three-to-five-word phrases achieve higher presentation scores. Other predictors like trigram frequency, trigram proportion, and trigram MI indicated lesser influence, with each contributing less than 4% to the total dominance weight.

Table 5. Robust regression and dominance analysis (criterion: presentation raw scores).

Predictor	B	95% CI	SE	T	Dominance Weight (%)
Intercept	39.02	16.45, 61.58	11.51	3.39	
Bigram frequency	8.10	−4.17, 20.37	6.26	1.29	0.008 (3.41%)
Bigram proportion	−12.09	−39.57, 15.39	14.02	−0.86	0.006 (2.64%)
Bigram MI	14.38 *	8.01, 20.76	3.25	4.42	0.129 (58.63%)
Trigram frequency	−3.07	−13.10, 6.95	5.11	−0.60	0.003 (1.23%)
Trigram proportion	4.15	−58.81, 67.11	32.12	0.13	0.007 (3.23%)
Trigram MI	−2.05	−6.24, 2.14	2.14	−0.96	0.007 (3.28%)
Three-to-five-word usage	0.53*	0.06, 0.99	0.24	2.22	0.041 (18.80%)
Three-to-five-word MI	−0.02	−0.05, 0.01	0.02	−1.25	0.019 (8.78%)
Total					0.220 (100%)

Note: MI = mutual information. * Results are marked significant if the 95% confidence interval excludes zero.

In Figure 2, the dominance weights are presented in descending order starting from the predictor variable with the highest weight (bigram MI) and ending with the one with the lowest weight (trigram frequency). Figure 1 helps us visualize each predictor’s relative importance, highlighting bigram MI as the most important predictor among all variables in our study.

Figure 3 shows a variable importance plot derived from random forests using the Boruta algorithm. The Boruta results confirm the importance of five attributes: Bigram MI, three-to-five-word usage, trigram proportion, three-to-five-word MI, and bigram proportion. Two attributes, trigram frequency and trigram MI, were confirmed as unimportant, while bigram frequency remained tentative. Detailed results and the R code are available in Supplementary Materials.

Overall, the Boruta analysis, along with dominance and robust regression analyses, together highlight the substantial influence of bigram MI and three-to-five-word usage on presentation scores. These consistent findings across different analytical techniques underscore the importance of specific lexical choices, especially bigram MI, in determining speaking performance, thus offering a triangulated insight into how LBs impact speaking proficiency.

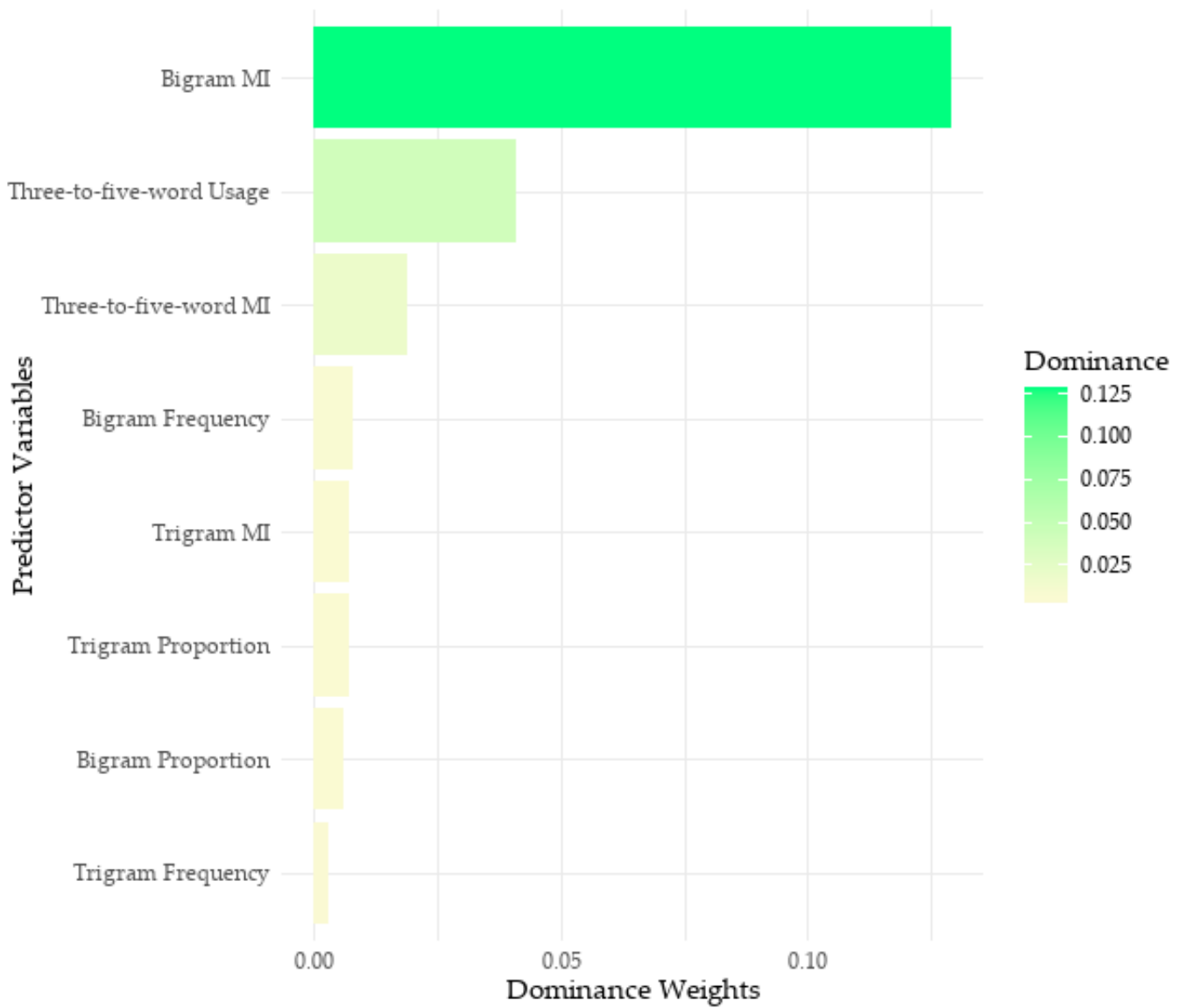


Figure 2. Dominance weights in descending order (criterion: presentation raw scores).

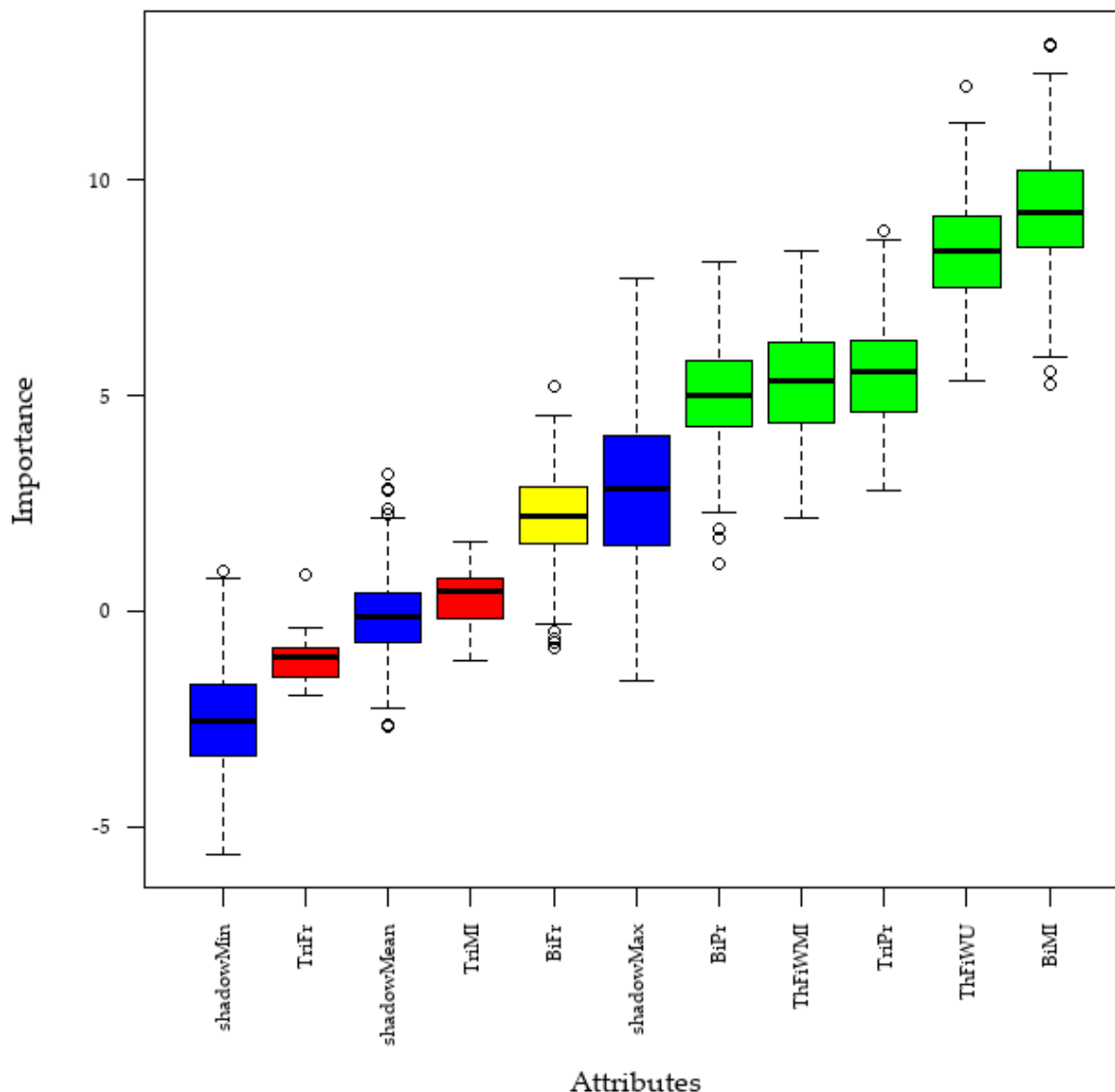


Figure 3. Variable importance plot from random forests using the Boruta algorithm (criterion: presentation raw scores). *Note:* In the Boruta algorithm box plots, green indicates “confirmed” variables, red indicates “rejected” variables, yellow indicates “tentative” variables, and blue indicates “randomized shadow” variables, which serve as a reference to assess the importance of the original variables against random chance.

6. Discussion

The overarching aim of this study was to explore the impact of LB usage on oral proficiency across a broad dataset and to employ a novel LB refinement technique for a more detailed analysis. Our research question asked to what extent LB usage of varying lengths could predict raw presentation scores. The following discussion has been structured to address this research question in relation to theoretical frameworks, previous findings, and our own hypothesis.

The robust regression and dominance analysis identified bigram MI as a significant positive and dominant predictor of raw presentation scores ($b = 14.38$, 95% CI [8.01, 20.76], $t = 4.42$; dominance weight = 58.63%). This was complemented by the finding that three-to-five-word LB usage was also found to be a significant positive and powerful predictor of presentation success ($b = 0.53$, 95% CI [0.06, 0.99], $t = 2.22$; dominance weight = 18.80%). These two LB measures’ significances were reinforced by the variable importance plot

generated by the Boruta algorithm, showing consistency across different approaches. Our discussion will thus focus on these two LB measures.

Our finding that learners who effectively use text-external high MI bigrams achieve higher proficiency levels are consistent with [Pawley and Syder's \(1983\)](#) lexical phrases theory. Pawley and Syder argued in their seminal work that the native-like selection of expressions, which includes collocations and idiomatic phrases captured by measures like bigram MI, is an integral aspect of speaking a language fluently. Their lexical phrases theory posits that a speaker's proficiency is marked by the ability to produce sequences of words that native speakers recognize as familiar, suggesting that native-like proficiency is to some extent a function of memory for lexically stored sequences rather than just rules for combining words. Our findings support and extend this notion by suggesting that not just the presence of LBs, but their "rare exclusivity", which is the main practical effect of the MI score ([Gablasova et al. 2017](#), p. 10), differentiates more proficient speakers from their less proficient counterparts. The empirical evidence presented in this paper supports the view that language proficiency, especially in productive skills such as speaking, involves the use of language patterns that are characteristic of L1 speaker usage. Our findings suggest that as L2 proficiency increases, so does the use of high MI score bigrams, indicating a narrowing gap in collocational usage between L2 and L1 speakers. These findings highlight the importance of integrating specific types of LBs, particularly those with high MI, into language assessment and instruction, supporting the idea that proficiency is not merely a matter of lexical range but also involves the strategic use of linguistically sophisticated and exclusive lexical patterns.

Our results also align with [Kormos's \(2006\)](#) extension of Levelt's model to L2 speakers by illustrating the formulation stage's critical role in speech production. Specifically, the increase in text-external bigram MI and longer text-internal LB usage with proficiency suggests a more efficient lexical selection process among higher proficiency learners, echoing the notion that advanced speakers can activate and employ appropriate lemmas with greater ease. This efficiency likely contributes to the freeing up of cognitive resources for other processing needs, which is essential for achieving fluency. It provides empirical support for the theory that a larger MWS repertoire enables L2 learners to enjoy a processing advantage by reducing the demands on cognitive resources, thereby facilitating more fluent and sophisticated language production. In addition, our findings that bigram MI and longer (three-to-five-word) LB usage predicts higher scores echoes the tenets of usage-based theory, specifically as it relates to language proficiency. Usage-based theory suggests that language learning is exemplified through the increased use and understanding of recurrent patterns or constructions in language, which are acquired through exposure and use ([Bybee 2006](#); [Ellis 2002](#)). Bigram MI within the usage-based framework can be seen as a proxy for the type of patterned use that is a hallmark of language proficiency. Bigram MI measures the exclusivity and specificity of the co-occurrence between two words, reflecting how their joint appearance significantly exceeds what would be anticipated based on their independent distribution across texts. This measure not only highlights word pairs that share a unique connection but also underscores the meaningful combinations that are preferentially utilized by more proficient speakers. The high bigram MI scores in our study suggest that more proficient language users are more likely to employ exclusive and information-rich lexical sequences. In a sense, bigrams with high MI scores are those that are entrenched in the linguistic repertoire of proficient speakers, thus reflecting common usage patterns.

Our findings, highlighting the predictive power of bigram MI for proficiency levels, are also consistent with key insights from corpus-based SLA studies that have explored collocational usage differences between L1 and L2 writers. These findings, while encouraging, are for writing, while our paper considers speaking. Notably, the research in this area has consistently found that L1 users tend to produce collocations with higher MI score values compared to L2 users ([Durrant and Schmitt 2009](#); [Ellis et al. 2015](#); [Schmitt 2012](#)). The work of [Schmitt \(2012, p. 6\)](#) emphasizes this disparity, concluding that the absence of high MI collocations is a distinctive marker of non-native versus native production. Our

findings also echo the findings of previous research, which compared bigram usage among L2 writers across different proficiency levels. Granger and Bestgen (2014) compared the use of collocations by intermediate and advanced L2 writers, finding that essays scoring in the advanced range of the CEFR had higher proportions of bigrams with high MI scores than essays scoring in the intermediate range. These convergent findings highlight the significance of the MI score as a measure for distinguishing between the collocational choices of L1 and L2 users, as well as between intermediate and advanced L2 users, thus highlighting its utility and growing importance in language research. The differential use of high MI score collocations between L1 and L2 users and between intermediate- and higher-proficiency L2 users highlights a key aspect of language proficiency: the ability to employ exclusive, less frequent word combinations beyond common collocations.

However, a notable divergence arose concerning our bigram MI finding and the other existing literature focusing on spoken production. Notably, Tavakoli and Uchihara (2020) found a general decrease in MI with rising proficiency, suggesting a broadening between LB combinations as learners become more proficient. Conversely, our results showed an increase in bigram MI with proficiency levels (except for a slight decrease in trigram MI). Although both studies used TAALES software to measure bigram MI through a text-external approach, the difference in findings could be due to the distinct data sets, methodologies, or proficiency level groupings employed in the two studies. While Tavakoli and Uchihara used a Kruskal–Wallis H test because of violations of the homogeneity of variance, our study applied robust regression to account for outliers, which may have contributed to the contrasting outcomes in the MI trends. While both studies corroborate the trend of increased LB usage with proficiency, the current analysis contributes a novel perspective by documenting the pattern of bigram MI increase and by bringing to light the intricate usage of longer LBs as learners progress. Such contrasting results underscore the need for further research to explore the intricate dynamics of bigram MI use and proficiency levels, hopefully enriching our understanding of effective language use by L2 speakers.

Regarding our hypothesis that the use of LBs of all lengths would positively correlate with raw oral presentation scores, the results show a mixed but insightful picture. The robust regression and dominance analysis, corroborated by the Boruta algorithm, strongly support the hypothesis in the case of bigram MI and three-to-five-word usage. Bigram MI, which emerged as the most powerful and dominant factor, accounted for a significant portion of the variance in presentation scores. This likely reflects the importance of conciseness in effective oral communication, where the use of meaningful and compact word pairs (bigrams with high mutual information) enables speakers to convey their points succinctly, thereby engaging the audience effectively. The positive influence of three-to-five-word usage also highlights the value of specificity. These slightly longer LBs, which we measured through the text-internal approach, are critical for articulating complex or specific ideas in a clear and focused manner, enhancing the speech's informativeness without becoming overly wordy. However, our hypothesis found less support with other types of LBs. Bigram proportion, for instance, showed a negative nonsignificant association with presentation scores, and trigram-related measures (frequency and MI) had lesser influence. This suggests that, while certain LBs positively correlate with presentation performance, not all LB types show the same level of predictive power. These findings partially validate our hypothesis, thus underscoring the significant predictive power of specific types of LBs, particularly bigram MI, in relation to raw presentation scores.

In summary, while our hypothesis that all lengths of LBs would positively predict oral proficiency scores was only partially supported, the current study contributes insights into the specific lexical features (bigram MI in particular) that are most predictive of higher proficiency scores. The current study's findings enhance the existing literature by providing a nuanced picture of how LB usage evolves at higher levels of language learning. Our findings also contribute to research methodology by showing the effectiveness of combining robust statistical techniques with machine learning algorithms like Boruta to strengthen the robustness of educational research. The convergence of evidence from

different analytic techniques strengthens the reliability of the current study's findings. Using robust regression helped to mitigate the influence of outliers, dominance analysis provided insights into the relative importance of predictors, and the Boruta algorithm offered an additional layer of confirmation about which LB attributes are truly influential. Such methodological triangulation enhances the study's credibility, allowing for more confident conclusions regarding the predictive power of LB usage for speaking proficiency.

7. Limitations and Suggestions for Future Research

Although the results are encouraging, there is still room for future research to overcome the current study's limitations. One notable area that was not explored in the current study is the potential influence of different presentation topics on LB usage and presentation scores. This unexplored area might shed further light on findings such as the increased use of high MI score bigrams among higher-scoring learners. For example, [Gablasova et al. \(2017\)](#) emphasize that the MI score favors less frequent and more specialized combinations, such as technical terms, which can vary significantly with the nature of the presentation topic. They note that "the technical nature of a specific topic can influence the strength of collocations as measured by the MI score", often revealing hidden patterns when only generalized MI score rankings are considered across whole corpora. ([Gablasova et al. 2017](#), p. 20). This observation could be highly pertinent to our research, indicating that the frequency and variety of high MI score word pairs we have observed might not only reflect the speakers' language proficiency but also the specific vocabulary requirements of their chosen topics. Technical presentations, for instance, might necessitate the use of specialized vocabulary, thereby increasing the MI scores of the collocations used. Future studies could benefit from incorporating an analysis of the presentation topics, examining how the choice of topic influences the use of high MI score bigrams across proficiency levels.

At least five other potential limitations warrant consideration. First, the proficiency level of the participants was confined to intermediate- and higher-proficiency learners (IELTS bands 6.5 to 7.5), thus excluding lower-proficiency speakers and potentially limiting generalizability. Second, the study's cross-sectional design limits the ability to trace language proficiency development over time or establish causality between variables. A longitudinal design could be useful in future studies, monitoring LB usage and proficiency over time. Third, the frequency-based approach using corpus analysis software has certain limitations, especially when dealing with spoken corpora. The current study used a minimum frequency of three and a minimum range of three to identify LBs in the learner corpus in order to keep the analysis manageable. These minimum frequency cutoffs mean that this study did not identify all multiword units in a comprehensive way. Some multiword units were used only once or twice or used idiosyncratically or in a nonstandard way by the L2 learners in our corpus. Many multiword sequences tend to blend into the linguistic context in transcripts, and many are frames or have larger fillable slots, which present real challenges for automatic extraction techniques. Such infrequent and/or semifixed units were not detected in the current study's approach. The current study focused on bi- and trigrams using TAALES and three-to-five-word phrases using AntConc, possibly leaving out other multiword structures that may affect speaking proficiency. Future research could expand the scope of MWS analysis for more comprehensive insights. Fourth, the operationalization of LB usage through MI and frequency does not account for the qualitative aspects of contextual appropriateness in conversation. Lastly, since MI scores were primarily designed for two-word collocations, and since they do not consider the order of the words ([Biber 2009](#); [Hyland 2012](#)), the MI scores might not reliably measure longer lexical strings. Future research should aim to address these limitations.

8. Pedagogical Implications and Suggestions

The findings support including bigram MI and longer LBs in assessing and teaching English proficiency, especially in speaking. They highlight the importance of focusing on the quality of LBs, particularly high MI bigrams, rather than just their quantity. This insight

can inform instructional design to enhance learners' speaking proficiency. Here are a few practical steps for implementing these findings in the classroom:

1. Select relevant texts: Choose texts that align with students' interests and proficiency levels.
2. Identify high MI bigrams: Use Tom Cobb's "Phrase Extractor" tool (<https://www.lextutor.ca/multiwords/phrase/>, accessed 3 February 2024) to extract high MI bigrams from the selected texts. These bigrams often serve as the foundation for longer LBs and expressions, thus making them a useful starting point.
3. Practice and raise awareness:
 - a. Develop exercises: Create exercises that target high MI bigrams and longer LBs.
 - b. Highlight during activities: Have students notice and highlight high MI bigrams and longer LBs during reading and listening activities.
 - c. Assessment criteria: Include criteria related to use of high-quality LBs in speaking (and writing) grading rubrics.

By integrating and building on these steps, educators can effectively enhance students' proficiency and awareness of high-quality LBs in English.

9. Conclusions

The current study has provided detailed insights into the patterns of LB usage that correspond to higher oral presentation scores. By extending beyond the scope of [Tavakoli and Uchihara \(2020\)](#), it has illuminated the intricate relationship between proficiency and both the frequency and complexity of LBs. Through extensive statistical analysis, it has established the link between linguistic sophistication and presentation scores, emphasizing the importance of both shorter and longer LBs in speaking proficiency. The current study's findings make a persuasive case for the critical role of LB usage, particularly bigram MI, in predicting English language proficiency and presentation performance. These findings enhance our understanding of the relationship between lexical choice and speaking performance, thus offering practical insights for language assessment and teaching. The current research has highlighted bigram MI (and to some extent three-to-five-word usage) as a reliable indicator of English language scores on oral academic presentations. This is consistent with linguistic theories first put forward in the 1980s (e.g., [Pawley and Syder 1983](#); [Levelt 1989](#)) that emphasize the efficient use of preformed chunks of language as a requirement of fluent speech and therefore a hallmark of proficiency. The implications for language teaching are considerable, suggesting that educators should include a focus on teaching strategies that improve learners' awareness and command of high MI bigrams.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/languages9070232/s1>; Table S1. Assumption checks for regression models; Table S2. Variance inflation factor (VIF) values; Figure S1. Cook's distance plots for detecting influential observations in the regression model of presentation raw scores; Table S3. R code and detailed results of the dominance analysis for presentation raw scores; Table S4. R code and detailed results of random forests and Boruta analysis for presentation raw scores.

Author Contributions: Conceptualization, D.H., J.C., T.U. and G.H.; Data curation, D.H. and G.H.; Formal analysis, D.H.; Funding acquisition, D.H., J.C., T.U. and G.H.; Investigation, D.H., J.C., T.U. and G.H.; Methodology, D.H., J.C., T.U. and G.H.; Project administration, D.H., J.C. and G.H.; Resources, D.H., J.C., T.U. and G.H.; Supervision, J.C.; Validation, D.H.; Visualization, D.H.; Writing—original draft, D.H.; Writing—review & editing, D.H., J.C., T.U. and G.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by two Grants-in-Aid for Scientific Research (No. 21K00669 and No. 22K00700) from the Japan Society for the Promotion of Science. The authors are very grateful for this support.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and APA ethical standards. It was approved by the Ethics Committee of Queen

Mary University of London where the data collection took place (research ethics approval number: QMREC2414a).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We extend our sincere gratitude to the two anonymous reviewers and the *Languages* production team for their valuable comments and suggestions. Their insightful feedback has greatly improved the quality of our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Presentation grading descriptors.

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Presentation Content [20]	<ul style="list-style-type: none"> • Purpose of presentation is clear, appropriate, and fully achieved. • Presentation is clearly focused, and only relevant issues presented. • Excellent research, which is clearly demonstrated through illustrations and examples. • All source material cited. • Visual aids are designed to a professional standard in terms of layout, bibliography, and contents. • Very good analysis, synthesis, and application of research. 	<ul style="list-style-type: none"> • Purpose of presentation is clear, appropriate, and fully achieved. • Presentation is focused, and only relevant issues presented. • Appropriate research is clearly demonstrated through illustrations and examples. • All source material cited. • Clear, well-designed visual aids; effectively proofread. • Good analysis, synthesis, and application of research. 	<ul style="list-style-type: none"> • Purpose of presentation is clear, appropriate, and largely achieved. • Presentation focused; issues presented are mainly relevant issues. • Appropriate research is demonstrated through illustrations and examples. • All source material cited, despite minor errors. • Generally clear and well-designed visual aids. Some evidence of proofreading, but some errors may persist despite this. • Some evidence of ability to analyze, synthesize, and apply research. 	<ul style="list-style-type: none"> • Appropriate and adequately achieved purpose, though may lack clarity. • May be occasional loss of focus and irrelevancies in parts. • Presentation shows some evidence of research and an understanding of the topic. • All source material is cited, though with some errors. • Generally satisfactory design of visual aids. Some lack of proofreading may result in careless mistakes. • Presentation may be more descriptive than analytical. 	<ul style="list-style-type: none"> • Purpose of presentation is appropriate but may not be entirely achieved. • Some loss of focus and some irrelevancies may be evident. • Presentation demonstrates evidence of adequate research and some understanding of the topic • Most source material is cited, though with frequent errors. • Adequately designed visual aids. Inadequate proofreading may lead to careless mistakes. • Presentation may be rather descriptive. 	<ul style="list-style-type: none"> • Purpose of presentation may be unclear or inappropriate. • Presentation is generally unfocused with many irrelevancies. • Presentation demonstrates little evidence of research and weak understanding of the topic. • Some citation of source material. • Visual aids may provide inadequate support for the presentation. Inadequate proofreading or lack of proofreading may result in careless errors. • Presentation may be largely descriptive. 	<ul style="list-style-type: none"> • Purpose of presentation unclear or inappropriate. • Presentation is unfocused and contains many irrelevancies. • Presentation demonstrates no evidence of research and limited understanding of the topic. • Little or no citation of source material. • Visual aids nonexistent or inadequate. Lack of proofreading results in incomprehension. • Presentation may be entirely descriptive.

Table A1. Cont.

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Presentation Structure [20]	<ul style="list-style-type: none"> • Excellent flow; causes no difficulties for listener. • Logical sequencing of ideas and very good organization of presentation. Excellent introduction and conclusion. Questions invited. • Very good organization within sections. • Excellent use of signposting expressions to create cohesion and coherence. 	<ul style="list-style-type: none"> • Good flow; causes no difficulties for listener. • Logical sequencing of ideas good organization of presentation. Very clear introduction and conclusion. Questions invited. • Good organization within sections. • Effective use of signposting expressions to create cohesion and coherence. 	<ul style="list-style-type: none"> • Reasonably good flow; causes few difficulties for listener. • Good sequencing of ideas, which enables the message to be followed clearly. Good introduction and conclusion. • Reasonably good organization within sections, although some room for improvement. Questions invited. • Fairly good use of signposting language to create cohesion and coherence. 	<ul style="list-style-type: none"> • Reasonable flow causes occasional difficulties for listener. • Some ability to sequence ideas, but overall structure may contain flaws. Reasonable introduction and conclusion. • Fairly good attempt to organize sections into main and supporting ideas; some use of examples but insufficient. Questions invited. • Attempt at signposting language; sometimes inappropriate or inaccurate; parts may lack cohesion. 	<ul style="list-style-type: none"> • Adequate flow, but causes some difficulties for listener. • Limited ability to sequence ideas, and overall organization may be flawed, but the message can be followed adequately. • Introduction and conclusion may be simplistic, overlong, or rushed. Questions not immediately invited. • Sections may lack unity, but may show an attempt to use topic and supporting sentences. • Attempt at signposting language, but it may be inappropriate; there may be some lack of cohesion. 	<ul style="list-style-type: none"> • Lack of flow but causes strain for listener. • Flawed overall sequence of ideas, but message can be followed in places. Introduction and conclusion: simplistic, weak, do not correspond to body. Questions not invited. • Section structure may be weak and disconnected—sections short and disjointed; little use of examples and illustrations. • Limited use of signposting language and often inappropriate; some lack of cohesion. 	<ul style="list-style-type: none"> • Absence of flow, which often puts strain on listener. • Ideas are poorly sequenced and organized, and the message is difficult to follow. Lacks clear organization structure. Little understanding of the purpose of introductions and conclusions. Questions not invited. • Sections poorly organized and show little understanding of the purpose of structure. • Very limited or inaccurate use of signposting language; lack of cohesion.
Seminar Leadership [20]	<ul style="list-style-type: none"> • There is a totally clear task for seminar participants, and the content is all highly focused and relevant. • The student clearly demonstrates a very high level of awareness of his/her audience. • The discussion is excellently controlled throughout. • The student gives a highly lucid summary of the discussion at its conclusion. 	<ul style="list-style-type: none"> • There is a clear task for seminar participants, and the content is focused and relevant. • The student demonstrates very good awareness of his/her audience. • The discussion is very well controlled. • The student gives a very good, lucid summary of the discussion at its conclusion. 	<ul style="list-style-type: none"> • There is a fairly clear task for seminar participants, and the content is mostly focused and relevant. • The student demonstrates good awareness of his/her audience. • The discussion is well controlled. • The student gives a good summary of the discussion at its conclusion. 	<ul style="list-style-type: none"> • There is a task for seminar participants, and the content is mostly relevant, but there may be some lack of clarity. • The student has satisfactory awareness of his/her audience. • An acceptable attempt is made to control the discussion. • The student gives a satisfactory summary of the discussion at its conclusion. 	<ul style="list-style-type: none"> • There is a task for seminar participants, but it may not be presented clearly. Some of the content may lack focus and relevance. • The student may lack awareness of his/her audience. • The discussion may not be well controlled. • The student gives a summary of the discussion at its conclusion, but this may lack clarity. 	<ul style="list-style-type: none"> • There may be some confusion about the task for seminar participants. The content lacks focus and relevance. • The student lacks awareness of his/her audience. • The discussion is only just controlled. • The student gives a summary of the discussion at its conclusion, but this lacks clarity. 	<ul style="list-style-type: none"> • The task for seminar participants may be inappropriate or unclear and is poorly explained. The content is unfocused and irrelevant. • The student has little or no awareness of his/her audience. • The discussion is not controlled. • The student fails to give a summary of the discussion at its conclusion or does this very poorly.

Table A1. Cont.

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Language Fluency [20]	<ul style="list-style-type: none"> • Clear pronunciation all the time. • Very good, fluent command of language, with almost no hesitations and excellent control of speed. • Excellent use of intonation and stress to convey stance and topic changes • Register always appropriate for this type of interaction. • Script independent; very confident and effective use of nonverbal communication (e.g., facial expressions, appropriate appearance); 	<ul style="list-style-type: none"> • Clear pronunciation most of the time. • Good, fluent command of language, with few hesitations and very good control of speed. • Very good use of intonation and stress to convey stance and topic changes • Register always appropriate for type of interaction. • Script independent; confident and effective use of nonverba communication (e.g., facial expressions, appropriate appearance). 	<ul style="list-style-type: none"> • Generally clear pronunciation. • Good, fluent production, with some hesitations but good control of speed. • Generally good use of intonation and stress to convey stance and topic changes. • Register generally appropriate for type of interaction. • Generally, script independent; effective use of nonverbal communication (e.g., facial expression, appropriate appearance). 	<ul style="list-style-type: none"> • Pronunciation is clear, but there are some mispronunciations. • Speaks with a degree of fluency, but with limited control of speed and some hesitations. • Reasonable use of intonation and stress to convey topic changes, but stance may not always be evident. • Register reasonably appropriate for type of interaction. • Often script independent; often effective use of non-verbal communication (e.g., facial expressions) and acceptably appropriate appearance. 	<ul style="list-style-type: none"> • Pronunciation is generally clear enough to be understood despite a noticeable accent. • Can speak, but with significant hesitation. May require a ‘sympathetic’ interlocutor. • Intonation and stress may only occasionally be used to convey stance or topic change. • Register is just appropriate; may sometimes be inappropriate for type of interaction. • Partly script independent; some limited awareness of nonverbal communication (e.g., facial expressions used effectively on occasion, fairly appropriate appearance). 	<ul style="list-style-type: none"> • Mispronunciation sometimes makes communication difficult. • Hesitations can make communication difficult. Speed may be too fast or too slow. Often requires a ‘sympathetic’ interlocutor. • Stance and topic change not signalled with intonation and stress. • Register is often inappropriate for interaction. • Script dependent; little awareness of nonverbal communication (e.g., facial expressions sometimes inappropriate, fairly inappropriate appearance). 	<ul style="list-style-type: none"> • Mispronunciation severely impedes communication. • Frequent hesitation or lack of control over speed severely impedes communication. Requires a ‘sympathetic’ and active interlocutor.’ • Little control of intonation and stress. • Register is inappropriate for interactions. • Script dependent; poor awareness of nonverbal communication (e.g., inappropriate facial expressions and/or inappropriate appearance).

Table A1. Cont.

	80–100% CEFR: C2	70–79%+ CEFR: HIGH C1	60–69% CEFR: LOW C1	50–59% CEFR: HIGH B2	40–49% CEFR: LOW B2	30–39% CEFR: B1	1–29% CEFR: A1–A2
Language Accuracy [20]	<ul style="list-style-type: none"> • Student demonstrates mastery of the grammar required for the task; excellent ability to manipulate complex structures. • Excellent use of vocabulary, which is appropriate to the task. • Excellent academic style, with totally appropriate use of register, very good ability to express caution and to avoid overgeneralizing. • Clear evidence of proofreading (in visuals) and practice in presentation. 	<ul style="list-style-type: none"> • Student demonstrates an authoritative use of the grammar required for the task; good ability to manipulate complex structures. • Good use of vocabulary, which is appropriate to the task. • Very good academic style with appropriate use of register, good ability to express caution and to avoid overgeneralizing. Clear evidence of proofreading (in visuals) and practice in presentation. 	<ul style="list-style-type: none"> • Student shows an above average level of use of grammar required for the task; some use of complex structures but perhaps incorrect use. • Good range of appropriate vocabulary. • Good awareness of academic style (register, expression of caution, few overgeneralizations). • Good evidence of proofreading (in visuals) and practice in presentation, but some errors may persist despite this. 	<ul style="list-style-type: none"> • Student shows a reasonable use of grammar with some ability to manipulate complex structures. There may be a limited number of grammatical errors, but these do not interfere with meaning. • Vocabulary generally appropriate to the task. • Awareness of academic style, but some inappropriate register; expression of caution may be weak, and overgeneralizations may be evident. • Some lack of proofreading (in visuals) and practice in presentation may result in careless mistakes. 	<ul style="list-style-type: none"> • Student shows a basic grasp of grammar, but limited ability to manipulate complex structures. Errors may interfere with meaning. • Adequate range of appropriate vocabulary: a narrow range of simple language. • Some awareness of academic style, but there are likely to be several overgeneralizations and limited ability to express caution. • Inadequate proofreading (in visuals) and practice in presentation may lead to careless mistakes. 	<ul style="list-style-type: none"> • There may be recurrent grammatical errors and limited ability to manipulate complex structures. • Some inappropriate use of vocabulary. • Choice of style and register is often inappropriate. • Inadequate proofreading and practice may result in careless errors. 	<ul style="list-style-type: none"> • Significant, recurrent grammatical errors. Very limited ability to manipulate structures appropriately and frequent errors in basic grammatical structures. • Range of vocabulary is inadequate for the task; errors make the meaning difficult to discern and cause strain for the reader. • Limited or no ability to use academic style. • Lack of proofreading and practice results in incomprehension.

Note. These descriptors were used by teachers who assessed the preessional presentations at Queen Mary University of London. Shared with permission from Queen Mary University of London.

Appendix B

Table A2. The most frequent three-, four-, and five-word recurrent word combinations used by all speakers (*N* = 150): Ranked according to MI score.

Rank (FREQ)	Rank (MI)	Bundles	Freq	Range	FREQ as 5-Gram	MI Score
21	1	if you have any questions	22/18 *	22/18	18	33.14
29	2	market forces and free competition (are)	18/18	9/9	18	32.59
73	3	in a socially responsible way	6/5	5/4	5	32.17
98	4	the possibility of armed conflict	4/4	3/3	4	28.58
111	5	Let's start by taking	3	3		25.07
63	6	I shall(will) only take	7(3)	7(3)		24.89
56	7	the presentation will be brief	10/10	10/10	10	23.62
34	8	(seven) minutes of your time	14	14		23.40
101	9	a wide range of	3	3		22.34
78	10	have a seminar discussion	5	5		21.27
77	11	does globalization lead to	5	5		20.87
37	12	at the same time	12	12		20.78
47	13	as we all know	11	11		20.65
95	14	can be defined as	4	3		20.50
69	15	will last about seven (to eight minutes)	6	6		20.43
86	16	(have/has) a positive impact on	5	4		20.38
97	17	for a long time	4	3		20.32
108	18	it can be seen	3	3		20.10
99	19	a large amount of	3	3		20.06
76	20	the rich and the poor	4/4	4/4	4	19.99
24	21	as you(we) can see (from)	20(7)	17(7)		19.87
80	22	(to) help you understand the	5	5		19.72
102	23	advantage and disadvantage of	3	3		19.71
74	24	on the other hand	6	5		19.55
2	25	I would like to (talk about the/ discuss/explain/introduce)	50	39		19.50
61	26	all over the world	8	6		19.31
75	27	we all know that	6	4		19.26
96	28	developed and developing countries	4	3		18.79
100	29	a large number of	3	3		18.72
5	30	(firstly/ then) I will talk about	37	24		18.49
118	31	to deal with this	3	3		18.26
85	32	(my) presentation is(will be) divided into	5(4)	5(4)		18.25
50	33	I will focus on	10	10		18.24
117	34	This talk will last	3	3		17.58
13	35	corporate social responsibility (is)	26	12		17.51
105	36	in the United States	3	3		17.08
43	37	we are going to (talk)	12	6		16.76
109	38	it is easy to	3	3		16.32
110	39	my presentation today is	3	3		16.21
9	40	(is) my thesis statement (is/and)	28	22		16.04
67	41	Let's turn to the	7	5		15.88
25	42	at the end of (my/ the talk/ presentation)	20	19		15.86
88	43	as a result of	4	4		15.58
92	44	I will make a	4	4		15.27
40	45	(I will be) glad(happy) to answer (them)	12(5)	12(5)		15.18
106	46	is a form of	3	3		14.89
58	47	the last one is (the)	8	8		14.79
23	48	and then I(we) will (show)	21(4)	19(4)		14.70
116	49	there will be a	3	3		14.60
107	50	is the most important	3	3		14.39
119	51	will move on to	3	3		14.20
82	52	(I) will start with the	5	5		14.16
87	53	moving on to the	5	4		14.15
10	54	Let's move on (to the)	28	22		14.10
81	55	I will analyze the	5	5		14.00
89	56	in the context of	4	4		13.95
112	57	Let's talk about the	3	3		13.90
1	58	(today/ firstly/ then) I'm going to (talk about/ present/ analyze)	52	34		13.87
45	59	Let's start (begin) with (the)	11(5)	11(5)		13.70
114	60	the last part is	3	3		13.24
113	61	So in this presentation	3	3		13.09
30	62	we can see (the/that)	18	11		12.87
14	63	(I have divided) my presentation(talk) into (four/ five parts)	25(4)	25(4)		12.78

Table A2. Cont.

Rank (FREQ)	Rank (MI)	Bundles	Freq	Range	FREQ as 5-Gram	MI Score
94	64	So what are the	4	4		12.64
103	65	in the era of	3	3		12.53
104	66	in the field of	3	3		12.40
66	67	a brief introduction (of)	7	7		12.39
18	68	(is/about) the relationship between (a and b)	24	14		12.30
8	69	(first/then/second) I will introduce (the/my/some)	29	18		12.13
4	70	(firstly/and finally) I will give (you/the/a/some)	41	31		11.83
91	71	we move to the	4	4		11.81
93	72	(be) seen as a	4	4		11.66
90	73	that globalization is a	4	4		11.32
65	74	as well as (the)	7	7		11.15
6	75	(Let's) look at the (first part)	34	28		10.97
12	76	I want to (talk about/show)	26	20		10.88
39	77	(this/here) is my outline (and)	12	12		10.82
33	78	(with/about) the background information	15	10		10.78
52	79	(then) I will discuss (the)	10	10		10.75
31	80	We need to (look at)	17	9		10.74
41	81	Let's move to (the)	12	10		10.74
84	82	(the) third one is	5	5		10.51
20	83	first of all (I/the)	23	19		10.38
62	84	in this slide (I)	8	6		10.33
46	85	(and) after that I (will)	11	11		10.24
38	86	(becoming) more and more	12	12		10.12
16	87	(in/and) the second part (is/I)	24	22		10.00
26	88	So what is (globalization/the)	20	19		9.91
42	89	to give you (some/a)	12	10		9.80
51	90	(and) the last one (is the)	10	10		9.69
72	91	to start with (I)	6	6		9.62
115	92	the needs of the	3	3		9.48
3	93	(on to/at) the first part(one) (is/the introduction)	41(12) **	31(10)		9.39
59	94	(and) the third part (is)	8	8		9.33
48	95	we will have (a)	11	10		9.30
7	96	(the outline/focus/part) of my presentation (is)	31(4)	29(4)		9.27
79	97	(be) responsible for the	5	5		9.06
70	98	(will) show you the	6	6		8.79
32	99	in this presentation (I will)	15	13		8.43
54	100	in my presentation (I)	10	10		8.39
64	101	(and) the second one (is)	7	7		8.35
35	102	this presentation is (the)	14	13		8.08
60	103	in this part (I will)	8	7		7.99
49	104	(that) I think the	11	9		7.78
53	105	The purpose of (this presentation is)	10	10		7.68
71	106	the effect of (globalization)	6	6		7.38
28	107	(the) impact of globalization (on)	22	12		7.30
17	108	(go/move) to the next (slide/part/point)	24	20		7.04
27	109	(in) the process of (globalization)	19	14		6.97
15	110	the definition of (the/globalization/CSR)	25	19		6.81
44	111	the protection of (women's labor rights)	12	8		6.58
83	112	(presentation) is going to	5	5		6.45
68	113	(and) the third is	7	6		6.39
19	114	(is) one of the (main/most)	24	20		5.98
11	115	(lead to/with) the (rapid) development of	27	19		5.81
55	116	(about) the introduction of (globalization)	10	7		5.28
36	117	the first is (introduction)	13	8		4.98
57	118	of the world (economy)	9	8		4.92
22	119	This is the (outline)	22	19		4.72

Note: Contracted forms (e.g., *I'm*) were counted as a single word; for example, *I'm going to* was considered to be a trigram rather than a quadgram. * Multiple-frequency figures listed in column 3 represent the individual frequencies of the four-word sequences that make up the longer five-word structure. ** Frequency of the word combination containing the word in parentheses in the fillable slot.

Appendix C

Table A3. The top-20 MI bigrams produced by the highest-scoring 50 participants.

Rank (MI)	Bigram	MI Score	Freq	Range
1	carbon dioxide	11.36	2	1
2	Kyoto protocol	11.19	2	2
3	Hong Kong	11.16	1	1
4	Saudi Arabia	9.77	1	1
5	infectious diseases	9.75	1	1
6	global warming	9.57	2	1
7	21st century	9.20	1	1
8	sexual harassment	9.06	7	1
9	19th century	8.84	1	1
10	breast cancer	8.65	8	8
11	human beings	8.59	1	1
12	greenhouse gas	8.16	3	2
13	intellectual property	8.16	16	5
14	environmental protection	7.84	1	1
15	wide range	7.81	3	3
16	pharmaceutical companies	7.67	1	1
17	virtual reality	7.58	2	1
18	substance abuse	7.57	1	1
19	gas emissions	7.48	1	1
20	17th century	7.41	1	1

Notes

- ¹ Unlike Tavakoli and Uchihara (2020), we did not use t-scores, because it has recently been indicated that they do not measure association very reliably (Gries 2022).

References

- Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent Word combinations. In *Phraseology: Theory, Analysis, and Applications*. Edited by Cowie Anthony Paul. Oxford: Oxford University Press, pp. 101–22.
- Anthony, Laurence. 2022. AntConc (Version 4.0.5) [Computer Software]. Waseda University. Available online: <https://www.laurenceanthony.net/software/antconc/> (accessed on 3 February 2024).
- Appel, Randy, and David Wood. 2016. Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels. *Language Assessment Quarterly* 13: 55–71. [CrossRef]
- Barlow, Michael. 2015. Collocate (Version 2.0) [Computer Software]. Athelstan. Available online: <https://athel.com/> (accessed on 3 February 2024).
- Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14: 275–311. [CrossRef]
- Biber, Douglas, and Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the *Toefl iBT*[®] test: A lexico-grammatical analysis. *ETS Research Report Series* 2013: i–128. Available online: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2013.tb02311.x> (accessed on 3 February 2024).
- Biber, Douglas, and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–86. [CrossRef]
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Boers, Frank, June Eyckmans, Jenny Kappel, Helene Stengers, and Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10: 245–61. [CrossRef]
- Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 711–33. Available online: <http://www.jstor.org/stable/4490266> (accessed on 3 February 2024). [CrossRef]
- Cowie, Anthony Paul. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2: 223–35. [CrossRef]
- Dang, Thi Ngoc Yen, Averil Coxhead, and Stuart Webb. 2017. The academic spoken word list. *Language Learning* 67: 959–97. [CrossRef]
- Davies, Mark. 2009. The 385+ million word *Corpus of Contemporary American English* (1990–2008+). Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14: 159–90. [CrossRef]
- De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3: 59–80. [CrossRef]

- De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, New Series* 2: 225–46. Available online: https://dial.uclouvain.be/downloader/downloader.php?pid=boreal:75157&datastream=PDF_01 (accessed on 3 February 2024).
- Durrant, Phil, and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47: 157–77. [CrossRef]
- Ebeling, Signe, and Hilde Hasselgård. 2015. Learner corpora and phraseology. In *The Cambridge Handbook of Learner Corpus Research*. Edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. Cambridge: Cambridge University Press, pp. 207–29.
- Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24: 143–88. Available online: <https://psycnet.apa.org/doi/10.1017/S0272263102002024> (accessed on 3 February 2024). [CrossRef]
- Ellis, Nick C., Eric Frey, and Isaac Jalkanen. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In *Exploring the Lexis-Grammar Interface*. Edited by Ute Römer-Barron and Rainer Schulze. Amsterdam: John Benjamins, pp. 89–114.
- Ellis, Nick C., Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42: 375–96. [CrossRef]
- Ellis, Nick C., Rita Simpson-Vlach, Ute Römer, Matthew B. O'Donnell, and Stefanie Wulff. 2015. Learner corpora and formulaic language in second language acquisition research. In *The Cambridge Handbook of Learner Corpus Research*, 1st ed. Edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. Cambridge: Cambridge University Press, pp. 357–78. [CrossRef]
- Gablasova, Dana, Vaclav Brezina, and Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67: 155–79. [CrossRef]
- Garner, Jamie, and Scott Crossley. 2018. A latent curve model approach to studying L2 N-Gram development. *The Modern Language Journal* 102: 494–511. [CrossRef]
- Granger, Sylviane, and Magali Paquot. 2008. Disentangling the phraseological web. In *Phraseology: An Interdisciplinary Perspective*. Edited by Sylviane Granger and Fanny Meunier. Amsterdam: John Benjamins, pp. 27–49.
- Granger, Sylviane, and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching* 52: 229–52. [CrossRef]
- Gries, Stefan Th. 2022. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5: 1–33. [CrossRef]
- Groom, Nicholas. 2009. Effects of second language immersion on second language collocational development. In *Researching Collocations in Another Language*. Edited by Andy Barfield and Henrik Gyllstad. London: Palgrave Macmillan, pp. 21–33.
- Grömping, Ulrike. 2006. Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software* 17: 1–27. [CrossRef]
- Hasselgård, Hilde. 2019. Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In *Corpus Linguistics, Context and Culture*. Edited by Michaela Mahlberg and Viola Wiegand. Berlin: De Gruyter, pp. 339–62.
- Hougham, Dan, Jon Clenton, and Takumi Uchihara. 2024. Disentangling the contributions of shorter vs. longer lexical bundles to L2 oral fluency. *System* 121: 103243. [CrossRef]
- Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21. [CrossRef]
- Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32: 150–69. [CrossRef]
- Kormos, Judit. 2006. *Speech Production and Second Language Acquisition*. London: Routledge.
- Kursa, Miron B., and Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36: 1–13. [CrossRef]
- Kyle, Kristopher, and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49: 757–86. [CrossRef]
- Kyle, Kristopher, Scott A. Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50: 1030–46. [CrossRef] [PubMed]
- Larson-Hall, Jenifer. 2015. *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. London: Routledge.
- Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- Levelt, Willem J. M. 1992. Accessing words in speech production: Stages, processes and representations. *Cognition* 42: 1–22. [CrossRef] [PubMed]
- Liakhovitski, Dmitri, Yegor Bryukhov, and Michael Conklin. 2010. Relative importance of predictors: Comparison of random forests with Johnson's relative weights. *Model Assisted Statistics and Applications* 5: 235–49. [CrossRef]
- Liu, Dilin. 2012. The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes* 31: 25–35. [CrossRef]
- McGuire, Michael, and Jenifer Larson-Hall. 2017. Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal* 34: 1–25. Available online: <https://teslcanadajournal.ca/index.php/tesl/article/download/1271/1106> (accessed on 3 February 2024). [CrossRef]

- McGuire, Michael, and Jenifer Larson-Hall. 2021. The contribution of high-frequency multi-word sequences to speech rate and listening perception among EFL learners. *Vocabulary Learning and Instruction* 10: 18–29. Available online: https://vli-journal.org/wp/wp-content/uploads/2022/02/VLI_10_2.pdf#page=22 (accessed on 3 February 2024). [CrossRef]
- Mizumoto, Atsushi. 2022. Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning* 73: 161–96. Available online: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12518> (accessed on 3 February 2024). [CrossRef]
- Myles, Florence, and Caroline Cordier. 2017. Formulaic sequence (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition* 39: 3–28. [CrossRef]
- Nation, Paul. 2013. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nesselhauf, Nadja. 2004. What are collocations? In *Phraseological Units: Basic Concepts and Their Application*. Edited by David John Allerton, Nadja Nesselhauf and Paul Skandera. Karlsruhe: Schwabe, pp. 1–21.
- Paquot, Magali, and Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32: 130–49. [CrossRef]
- Pawley, Andrew, and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and Communication*. Edited by Jack C. Richards and Richard W. Schmidt. London: Longman, pp. 29–59.
- Plonsky, Luke, and Deirdre J. Derrick. 2016. A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal* 100: 538–53. [CrossRef]
- R Development Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: <http://www.r-project.org/> (accessed on 3 February 2024).
- Schmidt, Richard. 1992. Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition* 14: 357–85. [CrossRef]
- Schmitt, Norbert. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. London: Palgrave Macmillan.
- Schmitt, Norbert. 2012. Formulaic language and collocation. In *The Encyclopedia of Applied Linguistics*, 1st ed. Edited by Carol A. Chapelle. Hoboken: Wiley.
- Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An academic formulas list (AFL). *Applied Linguistics* 31: 487–512. [CrossRef]
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Siyanova-Chanturia, Anna, and Stefania Spina. 2020. Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning* 70: 420–63. [CrossRef]
- Skehan, Peter. 2014. *Processing Perspectives on Task Performance*. Amsterdam: John Benjamins.
- Staples, Shelley, Jesse Egbert, Douglas Biber, and Alyson McClair. 2013. Formulaic sequences and EAP development: Lexical bundles in the TOEFL iBT writing section. *English for Specific Purposes* 12: 214–25. [CrossRef]
- Stengers, Helene, Frank Boers, Alex Housen, and June Eyckmans. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics* 49: 321–43. [CrossRef]
- Suzuki, Yuichi, Masaki Eguchi, and Nivja de Jong. 2022. Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective. *TESOL Quarterly* 56: 1290–319. [CrossRef]
- Tavakoli, Parvaneh. 2011. Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal* 65: 71–79. [CrossRef]
- Tavakoli, Parvaneh, and Takumi Uchihara. 2020. To what extent are multiword sequences associated with oral fluency? *Language Learning* 70: 506–47. [CrossRef]
- Tremblay, Antoine, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61: 569–613. [CrossRef]
- Uchihara, Takumi, and Jon Clenton. 2020. Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research* 24: 540–56. [CrossRef]
- Uchihara, Takumi, and Tetsuo Harada. 2018. Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly* 52: 564–87. Available online: <https://www.jstor.org/stable/44987081> (accessed on 3 February 2024). [CrossRef]
- Uchihara, Takumi, Masaki Eguchi, Jon Clenton, Kristopher Kyle, and Kazuya Saito. 2021. To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech* 65: 311–36. [CrossRef] [PubMed]
- Wood, David. 2009. Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics* 12: 39–57. Available online: <https://journals.lib.unb.ca/index.php/CJAL/article/download/19898/21737/> (accessed on 3 February 2024).
- Wood, David. 2010. *Formulaic Language and Second Language Speech Fluency: Background, Evidence, and Classroom Applications*. Dallas: Continuum.

-
- Wood, David, and Randy Appel. 2014. Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes* 15: 1–13. [[CrossRef](#)]
- Zhang, Xiaopeng, Baoshan Zhao, and Wenwen Li. 2021. N-gram use in EFL learners' retelling and monologic tasks. *International Review of Applied Linguistics in Language Teaching* 61: 939–65. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.