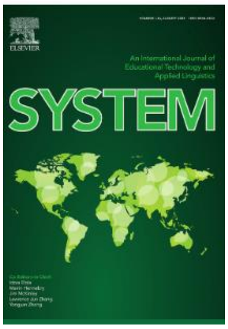
Accepted for publication in *System* (Elsevier)

Disentangling the contributions of shorter vs. longer lexical bundles to L2 oral fluency



Daniel Hougham (Hiroshima University)
Jon Clenton (Hiroshima University)
Takumi Uchihara (Tohoku University)

ABSTRACT

Shorter lexical bundles (LBs) have been the central point of focus in L2 oral fluency studies, with longer LBs often being neglected. The current study examined the extent to which longer LBs vs. shorter LBs relate to aspects of oral fluency. Data were collected from 50 undergraduate L2 English learners performing three speaking tasks. We analyzed speaking performances in terms of speed, breakdown, and repair fluency, while LB (2- to 5-word) usage was measured using a combined text-internal and text-external approach. Utilizing robust multiple regression, dominance analysis, and random forest techniques, our study found a marginal positive effect of longer LB use on speed fluency, a potential negative association with the frequency of mid- and end-clause pauses, and a strong negative association with the frequency of total repair. Furthermore, the analysis uncovers the significant impacts of shorter LBs (bigrams and trigrams) on various aspects of fluency. These insights underscore the pedagogical potential of longer LBs for enhancing oral fluency, while emphasizing the necessity for a comprehensive focus on different lengths and types of multiword sequences in EFL pedagogy. Our findings could inform more effective, data-driven language teaching strategies and materials. We discuss the findings in relation to L2 speech production models and provide important suggestions for future LB-fluency research.

Keywords: oral fluency; multi-word sequences; lexical bundles; learner-corpus research; dominance analysis; random forests analysis; multiple regression analysis

1. Introduction

The past 40 years have seen influential research in second language (L2) speech fluency and multiword sequences (MWSs) (e.g., Pawley & Syder, 1983; Schmidt, 1992; Wray, 2002; Myles & Cordier, 2017; Tavakoli & Wright, 2020). MWSs are essential for L2 learners because of the acknowledged influences on general speaking proficiency (Garner & Crossley, 2018) and overall language proficiency (Nation, 2013). One specific type of MWS that has gained popularity over the past two decades is lexical bundles (LBs: contiguous MWSs identified primarily by means of frequency and range). A growing number of studies investigating the influence of LB use on oral fluency have, however, focused on the use of relatively short (two- and three-word) bundles (e.g., McGuire & Larson-Hall, 2021; Tavakoli & Uchihara, 2020; Zhang et al., 2021). Emphasis on shorter two- to three-word sequences in prior studies may overlook the depth of learners' phraseological knowledge and its impact on oral fluency. Investigating longer LBs, along with shorter LBs, is crucial, as longer LBs may not only bolster oral fluency but also reveal processing efficiencies essential for fluent speech production. For instance, longer sequences such as “in light of recent events” or “with respect to the” frequently appear in fluent speech, suggesting that there are quite a few longer useful phrases that the investigation of longer LBs can help to promote. What remains unexplored and is the focus of the current article is the extent to which longer LB use relates to oral fluency. We respond to calls (e.g., Tavakoli & Uchihara, 2020) to explore the link between the use of longer LBs and aspects of utterance fluency. The current study aims to identify the relative contributions and importance of LBs of various lengths to aspects of utterance fluency.

2. Literature review

2.1 L2 Oral fluency

We can define L2 fluency in a broad sense (i.e., general L2 proficiency) and many use the terms fluency and speaking ability interchangeably (Tavakoli & Hunter, 2018). We can also consider L2 fluency in a narrower sense, as a component of oral proficiency (Lennon, 1990). Segalowitz (2016) outlines three types of L2 fluency: cognitive, utterance, and perceived fluency. Cognitive fluency means the smooth operation of the cognitive processes (i.e., speed and efficiency of word-meaning retrieval) underlying L2 speech acts. Utterance fluency refers to the temporal aspects of speech (i.e., speech rate, hesitation, and pausing phenomena). Perceived fluency refers to listener-based subjective judgements of fluency. The present study is concerned with fluency in the narrow sense focusing on utterance fluency.

Utterance fluency can be conceptualized in terms of three dimensions in which fine-grained analysis of audible speech is measured according to: (a) speed (i.e., the flow and continuity of speech), (b) breakdown (i.e., pauses that disrupt the speech flow), and (c) repair (i.e., self-monitoring mechanisms such as self-corrections and reformulations) (e.g., Skehan, 2003; Suzuki et al., 2021; Tavakoli & Skehan, 2005). Such ‘objective’ research contrasts with ‘subjective’ research (e.g., Boers et al., 2006) that employs human judgements to measure perceived fluency according to various assessment criteria (e.g., fluency, range of expression, and accuracy). One of the key findings revealed by studies taking the objective approach is that fluent and dysfluent speech differ in terms of pause location. Fluent speech is characterized by less frequent pausing in mid-clause positions, whereas dysfluent speech often contains more mid-clause pauses (Tavakoli, 2011; Kahng, 2014). Further, mid-clause pauses have been found to be related to proficiency (more proficient learners produced fewer silent pauses within clauses; de Jong, 2016a) and perceived fluency (Kahng, 2017; Suzuki et al., 2021). These studies show that pause location is an important indicator of how successful a speaker might be in producing speech fluently. Taking these various threads together, the current study examines fluency in the narrow sense, reflecting the interest and importance of investigating the use of objective L2 fluency measurements (speed, repair, and breakdown, including pause location in terms of mid- versus end-clause position) in the fields of language teaching and testing (Tavakoli et al., 2017; Thomson et al., 2019).

2.2 Multiword sequences and lexical bundles

MWS is an umbrella term encompassing combinations of words that occur together in language including various types such as idioms (e.g., *break a leg*), collocations (e.g., *fast food*), phrasal verbs (e.g., *cheer up*), proverbs (e.g., *the squeaky wheel gets the grease*) and LBs (e.g., *one of the*) (Garner & Crossley, 2018). We can broadly identify MWSs according to one of two approaches: a phraseological or a frequency-based approach (Granger & Paquot, 2008; Nesselhauf, 2004). The phraseological approach defines MWSs using linguistic criteria on a continuum ranging from free combinations to pure idioms (Cowie, 1981). This approach often relies on L1-speaker intuition in defining and measuring MWSs. Previous studies taking a phraseological approach (e.g., Boers et al., 2006) often involve the use of subjective (human) raters who measure MWSs by counting the number of multi-word chunks they consider as formulaic sequences. Such subjectivity might cause a relatively low agreement between experienced judges because the level of inter-rater reliability ($r < .60$), for instance, reported by Boers et al. (2006) is less than the median inter-rater reliability ($r = .92$) in L2 acquisition research reported by Plonsky and Derrick (2016).

In a more recent approach known as the frequency-based approach, corpus-based automated extraction techniques are used to identify a wider range of recurrent word combinations, according to quantitative criteria such as frequency threshold (i.e., a minimum number of occurrences of the unit in a corpus) and range (i.e., a minimum number of texts in which the units occur or a minimum number of learners using them). Researchers often refer to these recurrent word combinations as lexical bundles or n-grams where n can be two, three, four, or more consecutive words, including structurally complete word combinations (e.g., you know what I mean) and incomplete ones (e.g., so I think the), "regardless of their idiomaticity, and regardless of their structural status" (Biber et al., 1999, p. 990). LBs are a type of MWS that have become the focus of an increasing number of learner corpus studies because of objective, frequency-based measurement techniques (Paquot & Granger, 2012; Granger, 2019). By using the lexical bundle approach, we can acquire a large quantity of data for quantitative investigation, which serves as an ideal starting point for phraseological explorations (Ebeling & Hasselgård, 2015). In the present article, the term LB is used interchangeably with n-gram, referring to sequences identified using automatic extraction software with specified criteria such as minimum frequency, minimum range, and mutual information (*MI*) scores. *MI* scores evaluate the strength of association between pairs of words. In the review of research and the discussion that follows, we strive to distinguish between LBs and other types of MWSs wherever possible. However, it is important to recognize that the conceptual boundaries between LBs and other MWSs are sometimes ambiguous. For instance, the automatic extraction of bigrams with high *MI* scores could be seen as imbuing LBs with a distinct collocational quality.

When considering LB production in learner corpus research, two approaches of analysis can be taken: text-internal or text-external. Studies taking a text-*internal* approach (e.g., Biber & Gray; 2013) analyze LB production using only data within learner corpora (i.e., electronic collections of learner-produced text), tallying the number of adjacent sequences of a specified number of words. Alternatively, studies taking a text-*external* approach (e.g., Tavakoli & Uchihara, 2020) determine the formulaic status of LBs based on their frequency of co-occurrence in an external reference corpus, thus examining the extent to which L2 speakers use word combinations that occur in L1 speaker language. Both approaches, however, have limitations. Myles and Cordier (2017) argued that the text-internal approach is limited because the frequency cut-offs adopted are arbitrary and likely to be low in learner-produced texts in a single data set, and raw frequency alone is an inadequate measure of formulaic language. The text-external approach is limited because certain software programs (e.g., the Tool for the Automatic Analysis of Lexical Sophistication: TAALES; Kyle & Crossley, 2015; Kyle et al., 2018) used to determine the frequency of co-occurrence can only analyze shorter (two- and three-word) bundles. A further limitation of text-external techniques is that frequency of co-occurrence in external corpora is no guarantee that word bundles have psycholinguistic reality for the specific learners investigated (Ellis et al., 2009; Myles & Cordier, 2017). The current study aims to overcome these limitations by adopting a frequency-based approach using text-internal techniques to measure the contribution of longer LBs and using text-external techniques to measure the contribution of shorter LBs.

2.3 Previous theories explaining the LB-fluency relationship

To justify the LB-fluency link, Levelt's (1989) speech production model provides theoretical support. Levelt proposes three main stages of speech production: conceptualization, formulation, and articulation. First, speakers plan speech content at the conceptualization stage. Second, speakers encode lexical and grammatical items in the mental lexicon at the formulation stage, at which point appropriate lemmas are activated and put into syntactic surface structures, followed by morphological and phonetic encoding. The third and final articulation stage requires the phonetic plan to be implemented and speech is produced (Levelt, 1989). Initially proposed for L1 speakers, Levelt's model was later refined by Kormos (2006) to include L2 speakers. L2 processing and production face challenges not commonly found in L1 because of the constraints of the L2 mental lexicon. The L2 mental lexicon is considered smaller, less structured, less easily accessed, with fewer formulaic expressions. Theoretically, speakers with a larger MWS repertoire in their mental lexicon can retrieve longer MWSs at a similar processing rate to those required for single-word lexical item retrieval at the formulation stage. Retrieval of longer MWSs could allow them to free up cognitive resources and processing time at the formulation stage, to prepare for other processing needs such as phonological encoding, further message generation and articulation, thus enabling them to enjoy a processing advantage. A larger MWS repertoire may enable L2 learners to decrease the demands on cognitive resources that can be employed in simultaneous processing during other speech production aspects (e.g., lexical and grammatical accuracy or complexity) (Kormos, 2006; Skehan, 2014). Understanding and utilizing MWSs can enhance oral fluency, especially during lexical selection at the formulation stage (Kormos, 2006; Levelt, 1992). Conversely, speakers with a smaller MWS repertoire might not have this processing advantage, as they might use up cognitive resources trying to retrieve single lexical items one by one during the formulation stage. Speakers with a limited range of MWSs are more likely to show hesitations or pauses within MWSs. By comparing the impact of longer versus shorter language blocks on speech production, the current study provides an innovative perspective on the relationship between processing and fluency.

2.4 Previous studies examining the LB-fluency relationship

To date, most learner-corpus-based studies have focused on LBs used in writing rather than speaking (e.g., Appel & Wood, 2016; Siyanova-Chanturia & Spina, 2020; Staples et al. 2013). Such studies highlight the complexity of the MWS-L2-proficiency link, but general trends have emerged, including (a) the quantity of LBs produced by learners decreases as proficiency increases (Appel & 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, a trend resembling the use of "phraseological teddy bears" (i.e., overusing high-frequency phrases with which one feels comfortable; Hasselgård, 2019). Research has demonstrated that proficient learners have a firmer and more creative command of lower-frequency bundles whose constituent words are non-associated (Siyanova-Chanturia

& 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 & 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 fluency, 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, providing richer insights into speech fluency 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 & Crossley, 2018; Kyle & Crossley, 2015; Zhang et al., 2021). The findings from these studies differ from the typical trend from 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, suggesting that more skillful L2 speakers use a larger number of highly frequent trigrams. Garner and Crossley (2018) found that beginning-level L2 learners indicated 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 & 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 & Gray's (2013) study of spoken and written responses to the TOEFL iBT revealed a slightly more complex pattern than other n-gram studies. They reported that intermediate-level test-takers produced a greater number of bundles (four-word units) than the lower and higher proficiency groups, suggesting a general developmental progression in which lower-level test-takers use a smaller number of bundles, middle-level test-takers overuse a larger number of bundles, and high-scoring test-takers show greater control and creativity in using the bundles they have acquired (p. 37). In summary, these studies indicate a multifaceted view and a necessity for more study into how language utilization differs among individuals from divergent backgrounds.

While the aforementioned 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. This is evident 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; McGuire & Larson-Hall, 2017, 2021; Stengers et al., 2011;

Suzuki et al., 2022; Tavakoli, 2011; Tavakoli & 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 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, 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 recycling of more complex MWSs through task repetitions seemed to facilitate proceduralization (i.e., more efficient retrieval of MWSs), but that such reuse had both positive and negative influences on mid-clauses pauses, specifically, fewer but longer pauses within clauses, which may show that learner encoding systems were in the process of restructuring.

While the studies reviewed here support the hypothesis of a positive relationship between MWS use and speech fluency, they might be limited in at least six important ways. First and 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 & Crossley, 2018; Kyle & Crossley, 2015; Kyle et al., 2018; McGuire & Larson-Hall, 2021; Suzuki et al., 2022; Tavakoli & Uchihara, 2020; Zhang et al., 2021). For example, using longer LBs might be more beneficial for improving aspects of oral fluency and increase high-stakes assessment 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. Despite this insight, the contributions of these longer LBs to fluent speech production remain underexplored. Given what we know about Levelt's speech production model and previous empirical findings, 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 relying on subjective human judgments of speaking (e.g., Boers et al., 2006; Stengers et al., 2011; Zhang et al., 2021), focusing on only one aspect of fluency (e.g., speed fluency in Thomson, 2017; McGuire & Larson-Hall, 2017, 2021; Wood, 2009), or measuring MWSs subjectively using a criteria checklist and L1 speaker intuition (e.g., McGuire & Larson-Hall, 2017; Wood, 2009). With some exceptions (e.g., Suzuki et al., 2022; Tavakoli & Uchihara, 2020), there is a paucity of studies designed to examine the link between LB use and objectively measurable aspects of utterance fluency (speed, breakdown, and repair). Suzuki et al. (2022)

and Tavakoli and Uchihara (2020) aside, no previous study has examined the link between LB use and all aspects of utterance fluency, including, importantly, pause location (mid- versus end-clause position). Third, most previous research is restricted to investigating the MWS-fluency 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 & Crossley, 2018; Tavakoli & Uchihara, 2020; Zhang et al., 2021). There has been no systematic attempt to employ both text-external and text-internal methods to analyze learner-produced LBs in relation to oral fluency. Fourth, most previous MWS-fluency studies have been conducted with upper-intermediate to advanced students in ESL contexts (e.g., Garner & Crossley, 2018; McGuire & Larson-Hall, 2017; Tavakoli & Uchihara, 2020; Wood, 2009, 2010). Few studies have investigated the MWS-fluency link with lower or intermediate proficiency learners in EFL contexts where students have had little L2 exposure or have had minimal opportunities for L2 use (see Thomson, 2017, for an exception). Fifth, some previous studies have suffered from small sample sizes (e.g., *N* = 19 in McGuire & Larson-Hall, 2017; *N* = 1 in Wood, 2009; *N* = 11 in Wood, 2010). Sixth, because the LBs investigated in learner-corpus-based studies are often of different lengths (i.e., ranging from two to six words) and the extraction criteria used (frequency and dispersion) vary considerably from study to study, the findings and general trends previously discussed should be seen as mere hypotheses that need to be tested against other corpus data in different contexts.

3. The present study

With these identified gaps in mind and driven by the shortage of studies examining the degree to which extended LB use is associated with oral fluency aspects, the current study assesses the contribution of longer LBs to the LB-fluency linkage, expanding on and elaborating previous research in several critical respects. This is the first study to focus on the use of longer LBs in relation to shorter LBs and objectively measured aspects of utterance fluency (speed, breakdown, and repair), including, importantly, pause location (mid-ASU versus end-ASU position). The current study is also among the few studies focusing on the LB-fluency link with lower- to intermediate-level learners in an EFL context. It is also the first study of its kind that systematically employs both text-external and text-internal approaches, which not only enhances our understanding of the LB-fluency relationship but also paves the way for future research in this area. Additionally, we employ a larger sample size and systematically compare different lengths of LBs across varied extraction criteria, aiming for robust and generalizable findings.

Research questions

The current study builds on and extends Tavakoli and Uchihara's (2020) study by investigating the relationship between the use of various LBs, both shorter (bi- and trigrams) and longer (four- to five-word), and aspects of oral fluency. The following research questions guide this study:

RQ1: To what extent is the use of longer (four- to five-word) LBs associated with three aspects of fluency (speed, breakdown, and repair fluency?)
RQ2: How are shorter (bi- and trigram) LBs related to the three aspects of fluency?

Based on theoretical models of speech production (Kormos, 2006; Levelt, 1989), we hypothesize:

H1: The use of longer LBs will positively correlate with speed fluency, implying that learners who utilize longer LBs will produce speech at a faster processing rate.
H2: Speakers who use longer LBs are less likely to hesitate or pause within LBs and are likely to repair their speech less often.
H3: Similarly, the use of shorter LBs (bi- and trigrams) will also show a positive correlation with speed fluency, but to what degree this correlation differs from the longer LBs remains an exploratory aspect of this study.
H4: Speakers employing shorter LBs are less likely to hesitate or pause within LBs and are likely to repair their speech less often.
Overall, we anticipate that the use of both shorter and longer LBs will be positively associated with enhanced aspects of oral fluency.

4. Method

4.1 Participants

Participants were 50 L1 Japanese undergraduate learners of English (M_{age} = 19.5, range = 18–21) recruited from two Japanese universities, of whom 36 were female and 14 were male. They had studied English as a foreign language in classrooms in Japan for at least six years but did not use English regularly outside classrooms. Participants' English vocabulary size ranged from 2000 to 4800 words (M = 3805, SD = 658), indicating lower-intermediate to intermediate proficiency levels as measured by the X_Lex vocabulary size test (Meara & Miralpeix, 2016). Following previous research in the area of L2 oral speech production (e.g., de Jong & Mora, 2019), we used X_Lex because lexical size is an important dimension of lexical competence that is generally acknowledged as an indicator of general L2 English proficiency. The X_Lex test is a frequency–based yes/no instrument assessing learners' familiarity with the first five 1000-word frequency bands. Learners are presented with 120 words sequentially and must indicate whether they know each word. The test includes 20 imaginary words that serve as a simple method of correction for guessing: The learners' score is reduced if there is a large number of false alarms (i.e., "yes" responses to imaginary words). Their score reflects the number of recognized real words and is an estimate of learners' overall vocabulary knowledge (Meara & Miralpeix, 2016).

4.2 Speaking tasks

All participants completed three fluency tasks adapted from those used in Clenton et al. (2021) and de Jong et al. (2013). We chose the tasks because they varied in terms of their complexity, formality, and discourse type: (a) a formal descriptive task (describing a crime scene to a police officer), (b) a formal persuasive task (responding in a town hall meeting to whether a new casino should be built next to an elementary school), and (c) an informal persuasive task (expressing an opinion on solutions to climate change). Although the original tasks used in de Jong et al. (2013) were for higher-level learners, Clenton et al. (2021) used adapted versions of these tasks with lower-level Japanese EFL learners and found that one aspect of fluency (frequency of silent pauses) related to productive vocabulary knowledge (as measured by Lex30). On this basis, although the three tasks were challenging for the participants' level, they were adapted appropriately enough to elicit speech samples. Each task began by presenting participants with a detailed bilingual (Japanese-English) explanation and photos of the situation. We asked participants to imagine they were speaking in the situation presented. We instructed the participants to complete the tasks themselves and to follow the directions presented on a computer screen. Participants had a 30-second period within which to prepare their response following the directions presented on the computer screen and then spoke the response aloud within a 2-minute response time. All tasks were completed and recorded on a personal computer.

4.3 Measuring fluency

Following Clenton et. al. (2021) and de Jong et al. (2012), we defined a silent pause as a 350-millisecond (or longer) occurrence of silence. We chose (a) mean syllable duration as a measure of speed fluency, (b) frequency of total repairs (reformulations, self-corrections, repetitions and hesitations) per minute of speaking time (excluding silent pauses) as a measure of repair fluency, and (c) frequency of mid- and end-clause pauses per minute of speaking time (excluding silent pauses) as a measure of breakdown fluency. The use of frequency measures corrected for speaking time, rather than total phonation time, provides benefits by removing the impact of silent pausing time and enhancing the measures' accuracy (de Jong, 2016b, p. 213). All measures were collated over the three tasks.

As for the pause analysis procedure, we employed an expert who has experience of using PRAAT scripts. First, the "mark_pauses.praat" script (from the Speech Corpus Toolkit for Praat, Lennes, 2021) was used to compute the silent pause measures automatically. The automatic pause measurements were then checked and edited manually while listening to the audio recordings to ensure a high degree of precision using spectrograms created in PRAAT. We then used the "calculate_segment_durations.praat" script (Lennes, 2021) to automatically calculate the pause segment durations. To investigate pause location in terms of occurrences in mid- and end-clause location, the first author transcribed the audio recordings using AI-powered transcription software (www.otter.ai). A research assistant then checked the transcriptions, and the first author double-checked them for accuracy. The first author then divided the unpruned transcriptions into analysis of speech units (ASUs, Foster et al., 2000). The pause location analysis then entailed several iterative steps, including listening to the audio recordings, examining the spectrograms created in PRAAT, inspecting the transcriptions, and marking pauses in either mid- or end-ASU position on the transcripts.

4.4 Measuring lexical bundles: A two-pronged approach

In this 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.

4.4.1 Text-external n-gram analysis and measures.

Following previous studies (Garner & Crossley, 2018; Tavakoli & 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 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 sub-section 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 & Harada, 2018) and differences in lexical profiles between spoken and written modes (Dang et al., 2017). We opted to maintain the consistency between the modality in which L2 words were elicited and the modality of the reference corpus based on practice used in previous studies (e.g., Uchihara & Clenton, 2020; Uchihara et al., 2021).

In the present study, proportion score indices measure the proportion of bi- and trigrams in our learner speech sample data, which are also found among the 30,000 most frequent bi- and trigrams in the external reference corpus (COCA). Higher proportion scores show that participants in our sample produced a higher percentage of high-frequency, target-like bi- and trigrams. Frequency score indices measure the number of high-frequency, target-like n-grams produced by our participants. Logarithmic bi- and trigram scores, instead of raw frequency scores, were used to control for Zipfian effects common in word frequency lists (Kyle & Crossley, 2015; Tavakoli & Uchihara, 2020). Higher frequency scores show that participants in our sample produced a larger number of high-frequency target-like bi- and trigrams. 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 (*MI*) score.[1] MI score measures the strength of association between two words. While higher *MI* scores indicate words are more strongly associated, *MI* highlights word pairs which are relatively infrequent (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 28 and 415 words (*M* = 175.8, *SD* = 80.2).

4.4.2 Text-internal lexical bundle analysis and measures

We adopted a text-internal approach to isolate and measure the unique contribution of longer LBs to fluency aspects. As a first step, we conducted frequency analyses using AntConc (Anthony, 2022) to generate lists of the most frequently used four-word LBs in our 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 ours, a raw cut-off frequency is often used, ranging from two to ten occurrences (e.g., Altenberg, 1998; Biber & Barbieri, 2007; De Cock, 1998). Given the small size of the spoken corpus in the current study (8792 words), for four-word combinations to qualify as lexical bundles, we used a cut-off point of three or more occurrences in at least three texts, following Biber & Barbieri (2007). These minimum figures help to ensure that the identified bundles are not idiosyncrasies confined to occurrences produced by an individual speaker. The number of bundles generated at the minimum frequency and minimum range of three was 104. We deemed this number a suitable size for the identification and refinement of longer bundles, as well as manually scoring the usage of text-internal bundles to keep the scope of our study manageable.

4.5 Identification and refinement of longer text-internal bundles

We began our exploration of longer LBs by analyzing the occurrences of overlapping four- and five-word sequences to gain a clearer understanding of what longer LBs were frequently used among these participants. To better understand overlapping longer bundles in our data set, we examined the profiles of the most frequent four-word bundles used by all participants, using AntConc's KWIK (key word in context) concordance feature. To address overlapping sequences, we applied the procedure outlined in Appel and Wood (2016, p. 61). We undertook manual analysis to identify which partially recurrent four-word sequences were actually longer stretches (five to seven words) of repeated language. For example, two of the most frequent four-word sequences in our data are "solar *panel is the"* and "*panel is the* best" with frequencies of 16 and 14, respectively. Comparing the concordance lines of these similar sequences revealed the fact that *all* 14 occurrences of "panel is the best" are overlapping with "solar panel is the". These two sequences arguably should thus be combined to form the longer 5-word sequence "*solar panel is the best*". Any two or more four-word sequences that had three words in common (e.g., 'solar *panel is the'* and '*panel is the* best') and had a similar frequency of occurrence and range (within four of each other) were combined to form a longer five-word entry. We subsequently checked these longer word sequences for frequency of occurrence/range within the corpus. If the figures for both frequency of occurrence and range were within four of the frequency of occurrence and range of either of the two individual four-word sequences, the four-word sequences were removed from analysis and replaced with the new, longer sequence that had been identified. Using the previously noted procedure of identifying longer stretches of repeated language and eliminating their partially recurrent four-word structures, the original list of 104 recurrent four-word sequences was reduced to 84 sequences, 71 of which were four-word sequences (e.g., *I think it is)* and 13 of which were five-word sequences (e.g., *solar panel is the best)* (see Table A1 in Appendix for the full list). By extending the investigation to include not only four-word sequences but also longer five-word sequences, a higher percentage of the longer LBs within the corpus could be identified and overlapping bundles could be largely reduced. The resulting list of LBs includes both grammatically incomplete sequences such as *I think the best* and grammatically complete sequences such as *the problem of global warming*. Because of the small size of our corpus and limited figures for both frequency of occurrence and range, we found it was not applicable to include six- and seven-word structures in the analyses.

4.6 Scoring the use of text-internal lexical bundles

We created a scoring system that quantifies high-frequency text-internal LB usage. One point was awarded for each identified LB used from our combined list of four- and five-word combinations. For example, we gave a participant one point for use of the four-word unit *I think it is*. This scoring system assumes that learners who string together four-word or longer combinations are more phraseologically proficient and likely more fluent and rewards them accordingly (see a similar system as in Thomson, 2017). To calculate our text-internal measure labeled "four & five-word *MI*", we extracted *MI* scores for each identified LB in our combined list of four- and five-word combinations using the *Collocate 2.0* software program (Barlow, 2015) (see Appendix for the full list ranked according to MI scores). MI scores for three- to five-word sequences have been used in several studies (e.g., Ellis et al., 2008; Simpson-Vlach & Ellis, 2010) as they appear to offer reasonably reliable indication of phrasal coherence. For each bundle used by each speaker, we awarded the corresponding *MI* score. We then tallied all *MI* scores, gave each speaker a total score and divided each speaker's total *MI* score by the total number of 4-grams produced by that speaker.

4.7 Data analysis

We analyzed four fluency measures, six text-external n-gram (two- and three-word) measures, and one text-internal n-gram (four- and five-word) measure. To investigate the relationships between MWS measures and aspects of fluency, we conducted a robust regression analysis using MM estimation with the 'rlm' function in the MASS package in R (R Development Core Team, 2019). As recommended by Larson-Hall (2016, p. 264), we chose the 'rlm' function because it is suited for non-homoscedastic datasets containing outliers.

Initially, we examined correlations among the predictor (MWS) variables.[2] Subsequently, we ran multiple regressions with the predictor variables (PVs) and each of the criterion variables (aspects of fluency) separately. To investigate the relative importance of each PV in explaining the variance in the model, we ran a dominance analysis (DA) using the "calc.relimp" function in the "relaimpo" package in R (Grömping, 2006). DA can be used to effectively address correlations among PVs and can help in better understanding how each PV uniquely contributes to the criterion variable in multiple regression analysis, as opposed to solely relying on standardized beta coefficients which can be misleading (Mizumoto, 2022). DA facilitates comprehension by computing dominance weights for each predictor, which indicate the mean impact of a variable on the predictability of all potential subset of predictors, consequently presenting a thorough comprehension of the influence of each predictor on the outcome.

To achieve a more precise estimation of variable importance in the multiple regression model, it is important to conduct DA 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). The use of random forests allows researchers to gain 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 & 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. This will be depicted using boxplots in the figures presented in the following section.

In what follows, we present the descriptive statistics first, then we report the multiple regression, DA, and random forests results.

5. Results

Table 1 shows the descriptive statistics for the different n-gram and fluency measures in our data set. There are six score indices for text-external n-grams (two proportion, two frequency, and two association), three of which are bigram measures and three are trigram measures. There is one text-internal n-gram measure (four-five-word *MI* scores). And there are four fluency measures.

Table 1

Descriptive Statistics for All N-Gram and Fluency Measures (N = 50)

	Median	Mean	*SD*	Minimum	Maximum
Bigram frequency	1.395	1.382	0.169	0.617	1.744
Bigram proportion	0.505	0.506	0.080	0.259	0.676
Bigram *MI*	1.522	1.516	0.271	0.901	2.638
Trigram frequency	0.765	0.738	0.194	-0.147	1.050
Trigram proportion	0.134	0.137	0.058	0.000	0.299
Trigram *MI*	2.154	2.092	0.332	1.135	2.987
Four & five-word *MI*	0.700	0.737	0.460	0.000	2.496
Frequency of end-ASU pauses	4.375	6.400	4.207	1.900	15.100
Frequency of mid-ASU pauses	19.320	24.595	14.543	6.620	71.910
Frequency of total repair	4.500	5.403	3.878	0.900	24.310
Mean syllable duration	365.000	385.940	79.864	246.000	588.000

Note. MI = Mutual Information. ASU = analysis of speech unit. *SD* = standard deviation. Frequency of mid- and end-clause pauses, and frequency of total repairs are reported per minute of speaking time (excluding silent pauses).

Table 2 shows the results from the robust regression analysis with the criterion variable being mean syllable duration, and the corresponding dominance weights (see online supplementary materials for R code and detailed results of the dominance analysis). Our robust regression analysis found several key predictors for mean syllable duration in the dataset. Bigram frequency showed a positive association, suggesting that participants who produced more high-frequency bigrams spoke at a faster rate (*b* = 313.41, 95% *CI* [40.22, 586.60], *t* = 2.25). This variable accounted for a significant 18.12% of the variance in mean syllable duration as per dominance analysis. Similarly, bigram *MI* significantly contributed to our model, although in the opposite direction, suggesting that participants who produced more bigrams of collocational quality, spoke at a slower rate (*b* = -120.23, 95% *CI* [-229.36, -11.10], *t* = -2.16). This variable, despite its negative effect, still accounted for a notable 22.65% of the variance in mean syllable duration.

Trigram proportion, although showing a negative association with mean syllable duration (*b* = -553.90, 95% *CI* [-1401.68, 293.87], *t* = -1.28), was the most dominant predictor, accounting for the highest proportion of variance at 23.31%. This shows that even though trigram proportion decreases mean syllable duration, its role in explaining the variability in our response variable is highly significant. Lastly, four-five-word *MI* demonstrated a smaller positive effect (*b* = 13.07, 95% *CI* [-46.53, 72.67], *t* = .43) but accounted for a substantial 10.53% of the variance in mean syllable duration. It is notable that the confidence intervals for bigram frequency and bigram *MI* do not span zero, suggesting a degree of certainty in the precise effect sizes of these two predictors. Confidence intervals that do not include zero indicate a statistically significant effect. In the context of regression coefficients, this means that we can be reasonably confident the predictor has a true effect on the outcome variable, and it's not due to random chance. Their contribution to the overall model as indicated by relatively high dominance weights confirms their importance in predicting mean syllable duration. The confidence intervals for all other measures all span zero, suggesting a degree of uncertainty about the precise effect sizes of the other predictors.

Table 2

Robust Regression and Dominance Analysis (Criterion: Mean Syllable Duration)

	b	95% CI	SE	t	Dominance weight (%)
Intercept	456.77	[65.52, 848.02]	199.62	2.29	
Bigram frequency	313.41	[40.22, 586.60]	139.38	2.25	.044 (18.12%)
Bigram proportion	-359.85	[-914.44, 194.75]	282.95	-1.27	.024 (9.59%)
Bigram *MI*	-120.23	[-229.36, -11.10]	55.68	-2.16	.056 (22.65%)
Trigram frequency	-167.99	[-344.10, 9.02]	90.31	-1.86	.016 (6.69%)
Trigram proportion	-553.90	[-1401.68, 293.87]	432.54	-1.28	.057 (23.31%)
Trigram *MI*	21.58	[-63.84, 106.99]	43.58	.50	.022 (9.10%)
Four-Five-Word *MI*	13.07	[-46.53, 72.67]	30.41	.43	.026 (10.53%)
Total					.245 (100%)

Note. *N* = 50. *MI* = Mutual Information. *CI* = confidence interval. *SE* = standard error.

Figure 1 shows the dominance weights in descending order from the PV with the largest dominance weight (i.e., trigram proportion) to that with the smallest dominance weight (trigram frequency) from Table 2. For ease of interpretation, the colors in the figure transition from spring green (indicating the largest dominance weight) through light golden yellow (moderate dominance weights) to misty rose (the smallest dominance weight). Figure 2 gives a different perspective showing the variable importance plot of random forests based on the data from Table 2. By comparing Figure 2 to Table 2 (and Figure 1), the result of random forests (Boruta) partially corroborates that of dominance analysis. In Figure 2, only one of the MWS variables was confirmed important: bigram MI, shown on the right side of the “shadowMax” variable. All other MWS variables were confirmed unimportant (see online supplementary materials for R code and detailed results of the random forests and Boruta analysis).

For clarity in interpreting the boxplots generated by the Boruta algorithm, the color scheme is: green represents "confirmed" variables, red denotes "rejected" variables, yellow signifies "tentative" variables, and blue corresponds to randomized shadow variables. Note that the three shadow variables generated by the Boruta algorithm are randomized copies of the original variables, serving as a reference to test whether the importance of the original variables is higher than random chance.

Figure 1

Dominance Weights in Descending Order (Criterion: Mean Syllable Duration)

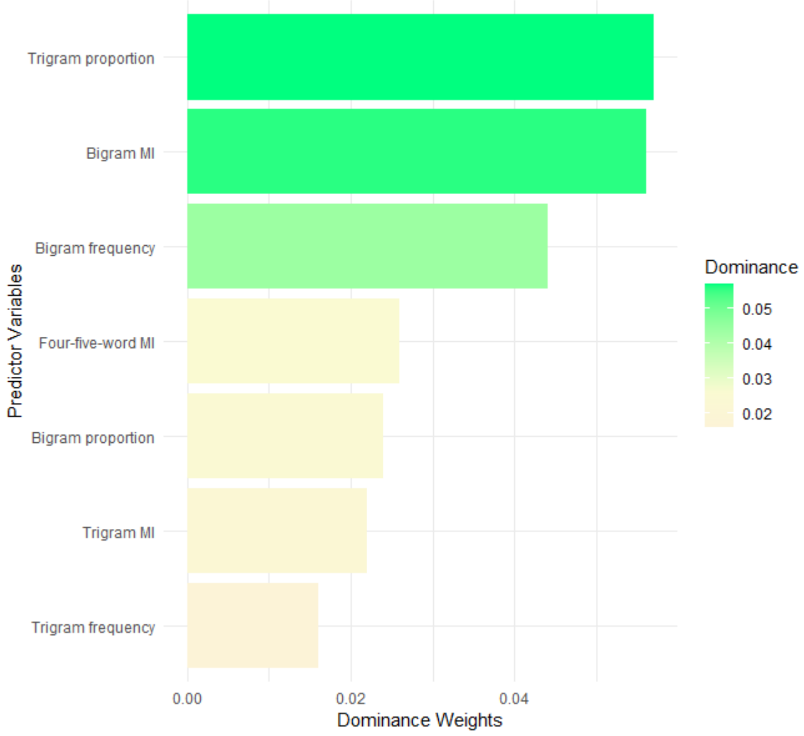


Figure 2

Variable Importance Plot Obtained from Random Forests Using the Boruta Algorithm (Criterion: Mean Syllable Duration)

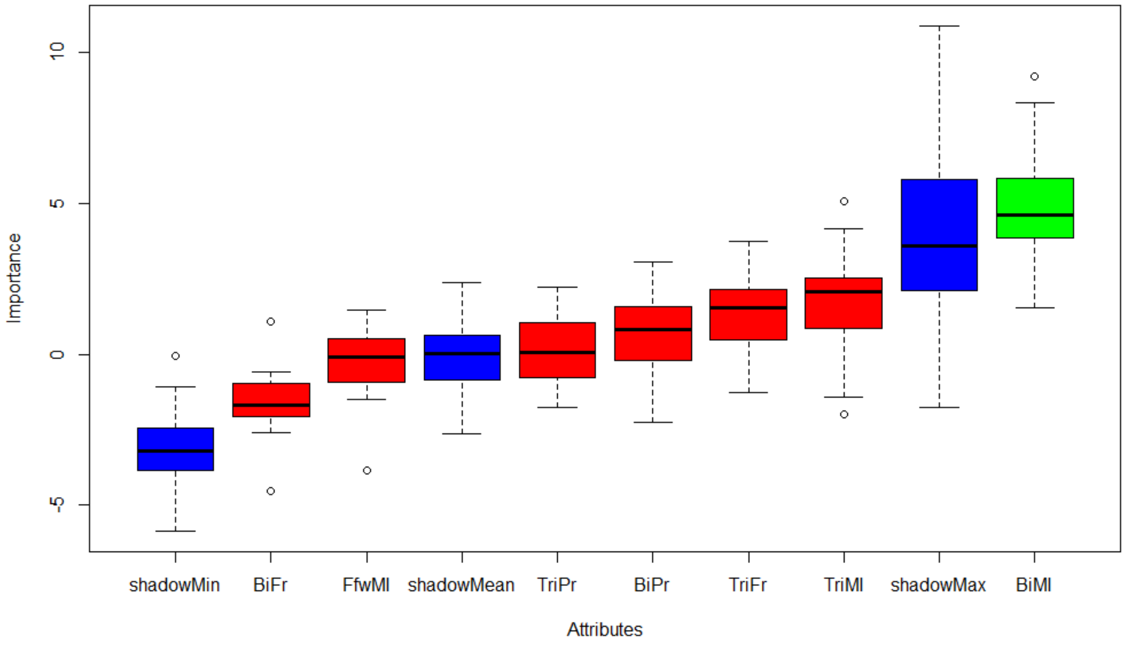


Table 3 shows that none of the variables appeared to have a significant effect on the frequency of mid-ASU pauses at conventional statistical thresholds (alpha level of .05) based on their associated t-values. However, four-five-word *MI* emerged as the predictor that was closest to significance with a *t*-value of -1.72 and an effect size of -9.99 (95% *CI*: -21.34, 1.37). This suggests a negative relationship, where an increase in four-five-word sequences of collocational quality corresponds to a decrease in the frequency of mid-ASU pauses. Furthermore, despite the non-significance in the regression analysis, dominance analysis confirmed the relative importance of four-five-word *MI* from a different angle. Four-five-word *MI* contributed the most to the model, with a dominance weight of 39.59%, suggesting that it has the highest relative influence on the frequency of mid-ASU pauses among the considered predictors. Trigram frequency also showed a notable contribution with a dominance weight of 27.19%. In contrast, the dominance weights of other measures like bigram frequency, bigram *MI*, and trigram *MI* were lower (7.08%, 11.73%, and 8.53% respectively).

Table 3

Robust Regression and Dominance Analysis (Criterion: Frequency of Mid-ASU Pauses)

	B	95% CI	SE	t	Dominance weight (%)
Intercept	46.20	[-28.34, 120.74]	38.03	1.21	
Bigram frequency	-13.98	[-66.03, 38.07]	26.55	-0.53	.019 (7.08%)
Bigram proportion	2.37	[-103.28, 108.03]	53.91	0.04	.008 (2.98%)
Bigram *MI*	1.52	[-19.27, 22.31]	10.61	0.14	.032 (11.73%)
Trigram frequency	7.79	[-25.93, 41.51]	17.21	0.45	.073 (27.19%)
Trigram proportion	44.86	[-116.66, 206.37]	82.40	0.54	.008 (2.91%)
Trigram *MI*	-6.12	[-22.39, 10.16]	8.30	-0.74	.023 (8.53%)
Four-five-word *MI*	-9.99	[-21.34, 1.37]	5.79	-1.72	.106 (39.59%)
Total					.269 (100%)

Note. N = 50. *MI* = Mutual Information. *CI* = confidence interval. *SE* = standard error.

Figure 3 displays the dominance weights in a decreasing sequence, from the PV that carries the heaviest dominance weight (four-five-word MI) to the one with the lightest dominance weight (bigram proportion), as outlined in Table 3. Figure 4 shows the variable importance plot of random forests, derived from the data in Table 3. By cross-referencing Figure 3 with Table 3 (and Figure 4), the outcomes from the random forests (Boruta) confirm those of dominance analysis. In Figure 4, two of the variables were confirmed important (trigram frequency and four-five-word *MI*), which are the same two with the heaviest dominance weights shown in Figure 3. Only one variable (bigram *MI*) was labelled “tentative”, meaning Boruta could not make a clear and definitive decision about its importance, perhaps due to the variable having a borderline (neither strong nor weak) predictive power.

Figure 3

Dominance Weights in Descending Order (Criterion: Frequency of Mid-ASU pauses)

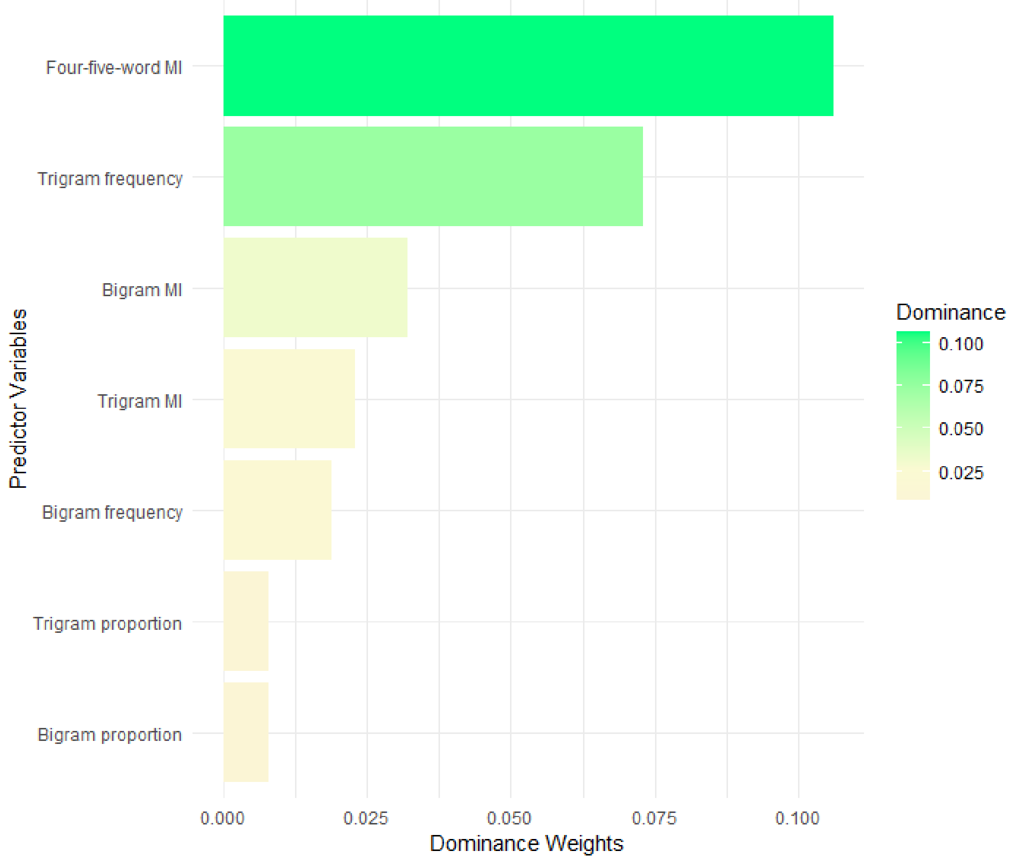


Figure 4

Variable Importance Plot Obtained from Random Forests Using the Boruta Algorithm (Criterion: Frequency of Mid-ASU Pauses)

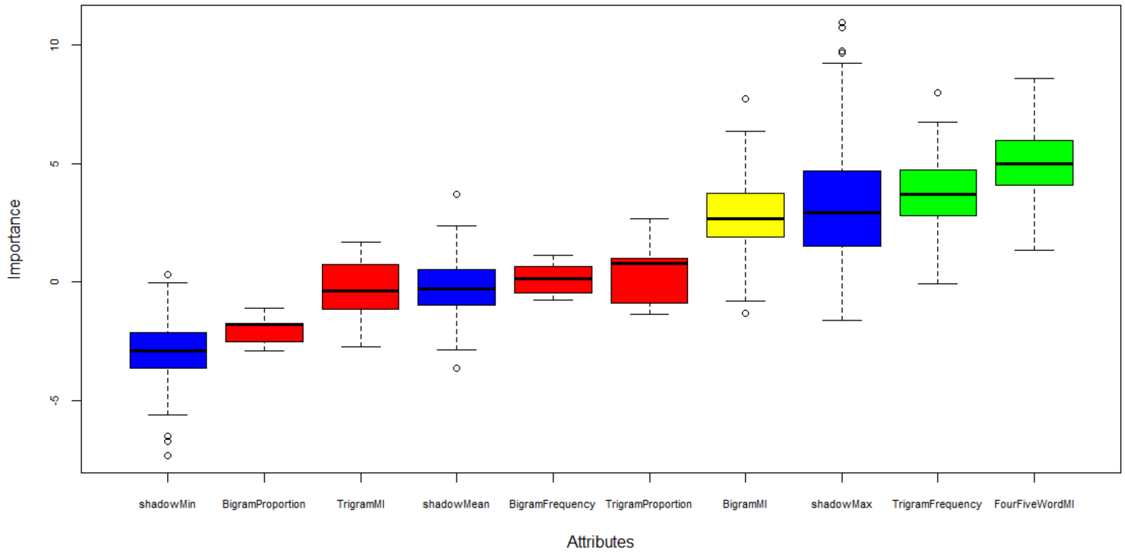


Table 4 shows that none of the predictors showed a significant impact on the frequency of end-ASU pauses at conventional statistical thresholds (alpha level of .05), based on their associated *t*-values. Nevertheless, the dominance analysis shows that four-five-word *MI* appears to have the highest relative importance, with a dominance weight of 34.17%. This suggests that it may have the strongest influence on the frequency of end-ASU pauses among the factors analyzed, despite its non-significant impact as per the regression analysis.

Table 4

Robust Regression and Dominance Analysis (Criterion: Frequency of End-ASU Pauses)

	b	95% CI	SE	t	Dominance weight (%)
Intercept	30.09	[.99, 59.19]	14.85	2.03	
Bigram frequency	-19.26	[-39.58, 1.06]	10.37	-1.86	.036 (18.55%)
Bigram proportion	14.34	[-26.91, 55.59]	21.05	0.68	.015 (7.89%)
Bigram *MI*	-1.67	[-9.79, 6.45]	4.14	-0.40	.045 (22.80%)
Trigram frequency	5.28	[-7.88, 18.45]	6.72	0.79	.015 (7.79%)
Trigram proportion	25.51	[-37.55, 88.56]	32.17	0.79	.007 (3.74%)
Trigram *MI*	-3.09	[-9.44, 3.27]	3.24	-0.95	.010 (5.07%)
Four-five-word *MI*	-4.10	[-8.53, .33]	2.26	-1.81	.067 (34.17%)
Total					.195 (100%)

Note. *N* = 50. *MI* = Mutual Information. *CI* = confidence interval. *SE* = standard error.

Figure 5 shows the dominance weights in descending order from the PV with the largest dominance weight (i.e., four-five-word *MI*) to that with the smallest dominance weight (trigram proportion) from Table 4. Figure 6 shows the variable importance plot of random forests based on the data from Table 4. By comparing Figure 6 to Table 4 (and Figure 5), the result of random forests (Boruta) corroborates that of the regression analysis. In Figure 6, none of the variables were confirmed important. Only one variable (four-five-word *MI*) was labelled ‘tentative’, indicating that four-five-word *MI* is of borderline importance in predicting end-ASU pauses.

Figure 5

Dominance Weights in Descending Order (Criterion: Frequency of End-ASU pauses)

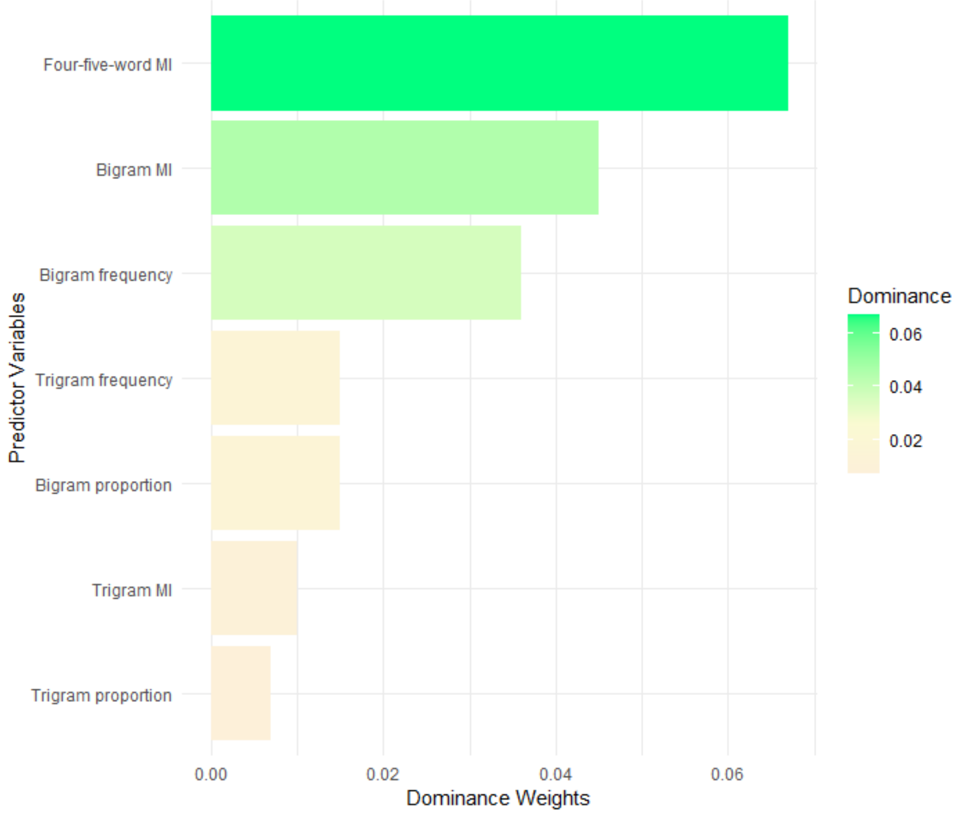


Figure 6

Variable Importance Plot Obtained from Random Forests Using the Boruta Algorithm (Criterion: Frequency of End-ASU Pauses)

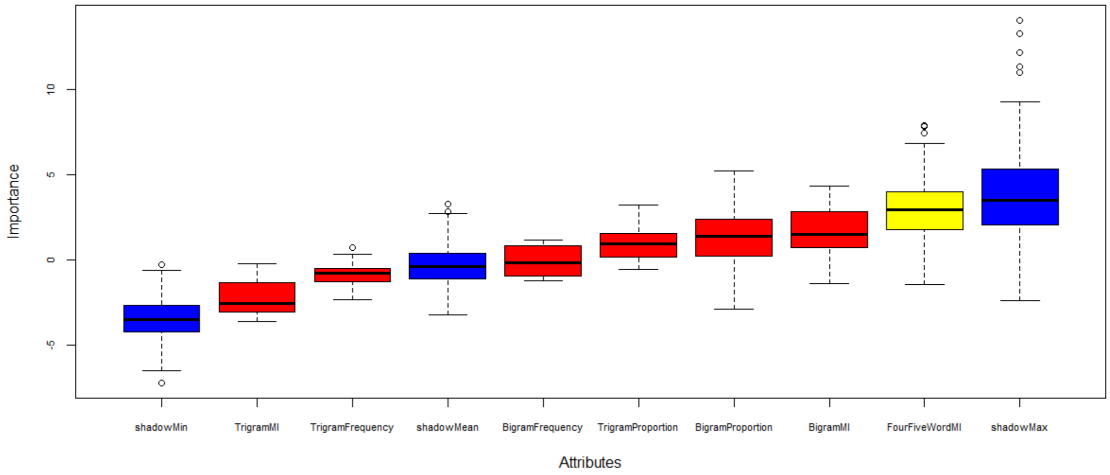


Table 5 shows the results from the robust regression and dominance analysis with the criterion variable being frequency of total repair. The analysis shows that four-five-word *MI* has the strongest effect, with an effect size of -2.41 (95% *CI*: -4.24, -0.59), reflecting a negative association with frequency of total repair. This indicates that as participants produce more four-five-word *MI* sequences, they repair their speech less frequently. Moreover, this variable accounted for the largest proportion of the model's predictive power, with a dominance weight of 42.75%. Bigram *MI* also showed a substantial dominance weight of 24.04%, suggesting a significant contribution to the model's overall predictive power. However, its effect size of -1.36 was smaller and its confidence interval crossed zero (95% *CI*: -4.70, 1.98), indicating a weak negative association with frequency of total repair. The effect sizes for other predictors, such as bigram frequency, bigram proportion, trigram frequency, and trigram proportion, were smaller, and their 95% confidence intervals included zero, suggesting that these associations may not be consistent across different samples. Their respective dominance weights ranged from 1.70% to 6.90%, indicating that they contributed less to the model's predictive power.

Table 5

Robust Regression and Dominance Analysis (Criterion: Frequency of Total Repair)

	b	95% CI	SE	t	Dominance weight (%)
Intercept	13.49	[1.51, 25.46]	6.11	2.21	
Bigram frequency	-6.19	[-14.55, 2.17]	4.27	-1.45	.022 (6.90%)
Bigram proportion	8.85	[-8.12, 25.82]	8.66	1.02	.005 (1.70%)
Bigram *MI*	-1.36	[-4.70, 1.98]	1.70	-0.80	.075 (24.04%)
Trigram frequency	0.93	[-4.49, 6.35]	2.76	0.34	.007 (2.09%)
Trigram proportion	14.63	[-11.31, 40.57]	13.24	1.11	.006 (2.05%)
Trigram *MI*	-1.85	[-4.46, 0.77]	1.33	-1.38	.064 (20.47%)
Four-Five-Word *MI*	-2.41	[-4.24, -0.59]	0.93	-2.60	.133 (42.75%)
Total					.312 (100%)

Note. *N* = 50. *MI* = Mutual Information. *CI* = confidence interval. *SE* = standard error.

Figure 7 shows the dominance weights in descending sequence, starting from the PV with the greatest dominance weight (four-five-word *MI*) and ending with the one with the least dominance weight (bigram proportion), as displayed in Table 5. Figure 8 provides a different perspective by presenting the variable importance plot of random forests, utilizing the data from Table 5. By comparing Figure 8 to Table 5 (and Figure 7), the findings from the random forests (Boruta) once more partially affirm those from the dominance analysis. In Figure 8, only two of the variables (those appearing on the right side of shadowMax) were confirmed important (four-five-word *MI* and bigram *MI*), of which bigram *MI* appears to be the single most important predictor of frequency of total repair by a large margin. One variable (bigram frequency) was labelled ‘tentative’, likely due to the variable having a borderline predictive power.

Figure 7

Dominance Weights in Descending Order (Criterion: Frequency of Total Repair)

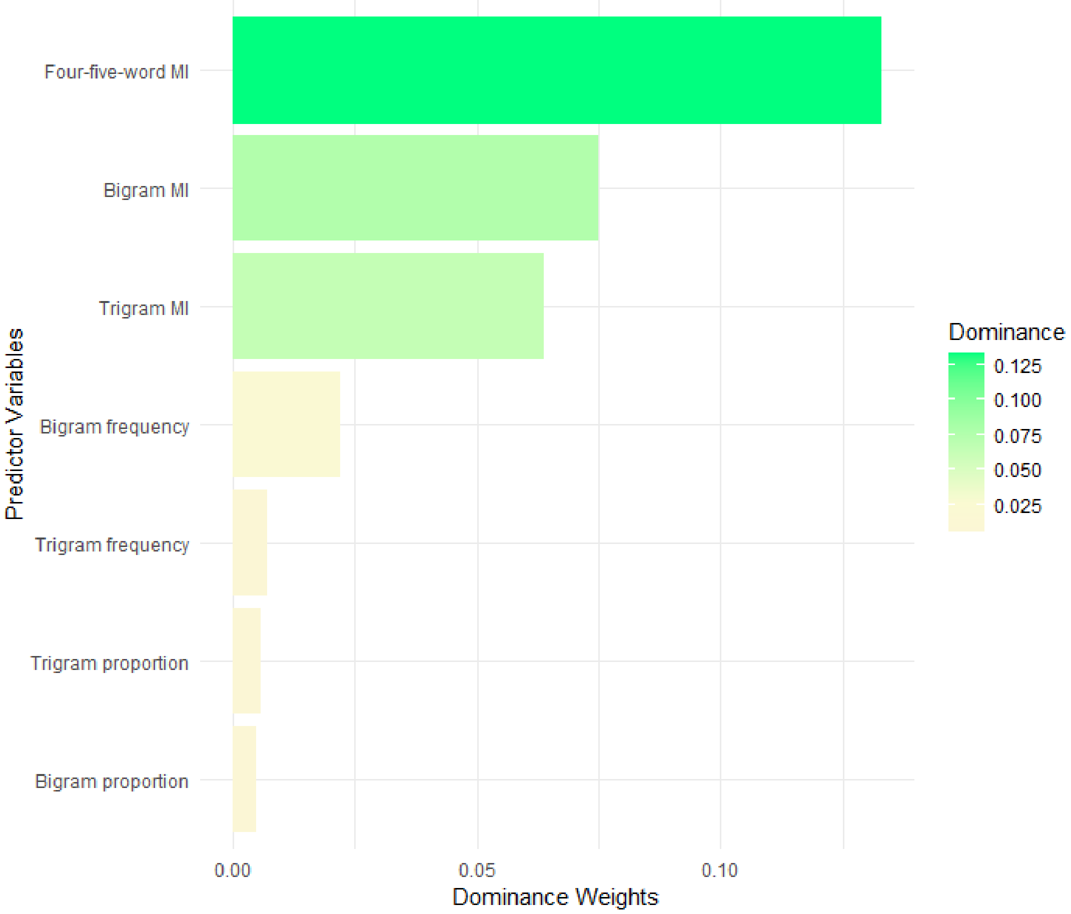
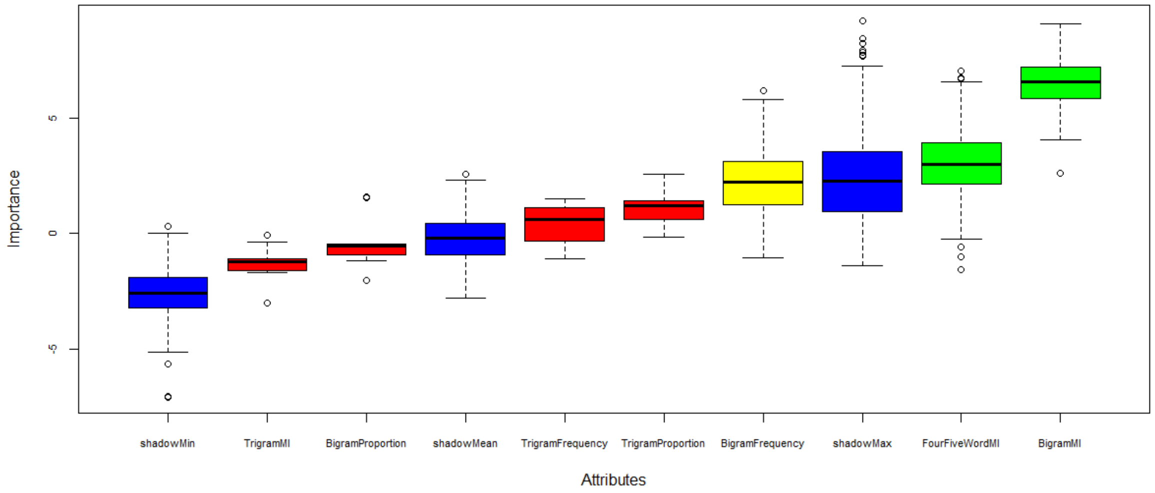


Figure 8

Variable Importance Plot Obtained from Random Forests Using the Boruta Algorithm (Criterion: Frequency of Total Repair)



To sum up, considering the detailed statistical analysis presented, it becomes clear that the structure and frequency of word sequences in speech play a crucial role in oral fluency. Specifically, the results show that using longer strings of words in speech helps speakers talk more smoothly with less frequent pauses for self-correction. Furthermore, the way certain word combinations are used can influence speed of talk, with more common pairs of words speeding up speech, while less common, more specialized word combinations can lead to a slower rate of speech.

6. Discussion

The overarching aim of this study was to unpack the nuanced relationship between the use of lexical bundles of different lengths and three distinct dimensions of oral fluency. To achieve this, we guided our investigation with two research questions—one focusing on the association of longer LBs with fluency aspects, and the other on the relationship of shorter LBs with these fluency dimensions. The following discussion has been structured to contrast the results for each fluency dimension across both LB lengths, thereby facilitating direct comparisons.

6.1 Speed fluency and lexical bundles

6.1.1 Longer (four- & five-word) lexical bundles

The findings from our robust regression analysis point to a complex relationship between the use of longer lexical bundles and speed fluency. Our results show that the usage of four- and five-word MI lexical bundles has a marginal positive effect on speech rate, with the confidence intervals crossing zero, suggesting a degree of uncertainty about the precise effect sizes of this predictor. Despite its minor impact, this variable accounted for a significant 10.53% of the variance in mean syllable duration, indicating its meaningful contribution to the speed of talk and confirming (albeit tentatively) our earlier prediction that speakers who use longer LBs would produce speech at a faster rate. While longer LBs may be associated with faster speech, the magnitude of this effect is likely to be minor.

6.1.2 Shorter (bi- & trigram) lexical bundles

Our analysis also showed a negative association between mean syllable duration and bigram *MI*, indicating that the more frequently participants used bigrams of collocational quality, the slower they spoke. In other words, participants who often used high-quality bigrams tended to speak more slowly. Notably, the random forests (Boruta) method confirmed the importance of bigram *MI*, underscoring its significant role in influencing speech speed. This suggests that while longer lexical bundles may have a minor effect on enhancing speech speed, certain shorter bundles may slow it down. The findings also showed that the Trigram Proportion, despite its negative association, accounted for the highest proportion of variance at 23.31%, indicating a stronger relationship with speech speed compared to the four- and five-word *MI* lexical bundles. These findings do not support our earlier prediction that the use of shorter LBs (bi- and trigrams) would show a positive correlation with speed fluency.

These findings contrast with Tavakoli and Uchihara (2020) who found a slight positive link between their n-gram Factor 1 (high-frequency trigram) and speed fluency (measured by pruned and unpruned articulation rate; r = .325 and r = .396, respectively). Their study also found no link between their n-gram factor 2 (association measure of *MI*), whereas our study found a negative association between bigram *MI* and speed fluency (mean syllable duration). These differences could be due to several reasons. First, the difference in proficiency levels and learning context of the speakers studied could play a role. Our study examined lower- to intermediate-level learners in an EFL context, while Tavakoli and Uchihara's study focused on speakers in the UK ESL context with higher proficiency levels. These higher proficiency speakers might have developed a more nuanced usage of n-grams, reflected in the different correlations observed. As our results suggest, it could be the case that text-external measures (such as those from COCA) may not work as well for lower-level learners in EFL contexts. It is plausible to think that learners in EFL contexts such as ours do not get as much exposure to L2 input as those in Tavakoli and Uchihara’s ESL context. Using COCA as a reference corpus may not reflect EFL learners’ L2 experience. Secondly, the difference in n-gram measures used might contribute to the discrepancies. Tavakoli and Uchihara used a factor (derived from principal component analysis) that encompasses high-frequency trigrams, whereas our study analyzed bigrams and trigrams separately, in addition to four- and five-word *MI* bundles. This difference in granularity might have led to different insights into the role of n-grams in speech rate.

To summarize, considering our findings, longer lexical bundles (four- & five-word) might play a modest role in enhancing speed fluency. However, the prominent impact comes from shorter lexical bundles, especially bigrams with high MI scores, albeit in an opposite direction than expected. The intricacies of how these bundles influence speed fluency differ across studies, highlighting the significance of context, such as proficiency levels, learning environments, and choice of reference corpus in different contexts.

6.2 Breakdown fluency and lexical bundles

6.2.1 Longer (four- & five-word) lexical bundles

Although none of the predictors, including four- and five-word MI, showed a significant impact on the frequency of mid- and end-ASU pauses according to our robust regression analysis, the dominance analysis and random forests provided a nuanced perspective. Four- and five-word *MI* exhibited the highest dominance weight in both models, implying it was the most influential predictor despite its non-significance in regression analysis. This suggests a potential negative relationship, with an increase in these sequences of collocational quality linked to a decrease in the frequency of mid-ASU pauses. This finding illuminates the complex role of longer lexical bundles in shaping fluency, hinting at their potential to reduce disruption in speech flow, and underscores the value of utilizing methods like dominance analysis to unveil key influences masked by conventional analysis. The random forests (Boruta) results corroborated the dominance analysis findings, further indicating the importance of four- and five-word *MI*, but also confirming the importance of shorter LBs (trigram frequency and bigram *MI*) in influencing mid-ASU pauses. This confirms our earlier prediction that speakers who use longer LBs would show fewer pauses in mid-ASU position, supporting the notion that the use of longer LBs is linked with smoother speech production.

These findings support those of previous research which found that L2 learners seldom pause in the middle of multiword units (Tavakoli, 2011) and that lower-level L2 learners produce more silent pauses within ASUs (de Jong, 2016a), confirming that pause location is an important indicator of how successful a speaker is in producing speech fluently. These findings are important because pauses within clauses are thought to be associated with the formulation stage of speech production (Skehan, 2014). Some argue that pauses within clauses reflect an L2 speaker's speech processing (e.g., lexical and morphosyntactic) difficulties and that MWS use helps to ease such demands (Felker et al., 2019; Kahng, 2014). Our findings support this argument because they suggest that less phraseologically knowledgeable speakers are more likely to pause within clauses while attempting to retrieve individual words, whereas more phraseologically knowledgeable speakers are more likely to retrieve LBs as whole chunks without mid-ASU pauses.

6.2.2 Shorter (bi- & trigram) lexical bundles

In contrast, our findings show that shorter LBs (bigrams and trigrams) demonstrated less influence than longer LBs. One notable exception is trigram frequency which showed a slightly positive relationship, implying that as the frequency of three-word sequences increases, the number of mid-ASU pauses also increases. This could indicate that more frequent use of target-like trigram sequences may actually contribute to speech disruption and increased pausing. While our finding regarding trigram frequency was not statistically significant according to the regression analysis, trigram frequency accounted for a significant 27.19% of the variance in mid-ASU pauses according to dominance analysis. This indicates its potentially meaningful contribution to breakdown fluency, but this potential contribution needs to be interpreted with caution. The apparent contradiction between increased trigram frequency and mid-ASU pauses can be explained by the cognitive processing demands on EFL learners at lower to intermediate levels. Producing trigrams might represent a transitional phase for these learners; while they've moved beyond relying on single words, they are still developing the automaticity required to fluidly produce longer sequences. Therefore, while the usage of these target-like trigrams may be higher, the cognitive effort needed to retrieve and express them could result in a rise in pausing. It is plausible that learners, in their efforts to produce more complex and fluent speech, may be overloading their working memory with these trigrams, leading to a momentary disruption in speech. This finding could be related to the non-linear nature of language acquisition, where certain developmental stages, although indicative of progress, might momentarily introduce challenges in fluency. However, these findings have contradicted our initial prediction that the usage of shorter LBs would result in fewer hesitations or pauses within those bundles. In fact, the evidence points to the opposite, especially in the case of trigrams: as their usage increases, so does the occurrence of mid-ASU pauses.

While none of the variables were confirmed important for end-ASU pauses, the "tentative" label for four- and five-word *MI* suggests borderline predictive importance, reaffirming its potential influence.

In summary, while the regression analysis did not point to significant predictors for mid- and end-ASU pauses, the dominance analysis and random forests provided more depth. It seems longer lexical bundles may play a more pivotal role in influencing the frequency of mid-ASU pauses, suggesting smoother speech production. The frequent use of trigram sequences could lead to an increase in pauses, as the generation of these trigrams places a considerable cognitive burden on intermediate EFL learners. This underscores the complexity of fluency development and the nuances involved in the acquisition and production of MWSs.

Further, the incongruity between regression, DA and random forests findings underscores the value of multi-method analyses in disentangling these complex relationships.

6.3 Repair fluency and lexical bundles

6.3.1 Longer (four- & five-word) lexical bundles

The results from the robust regression and dominance analysis indicate a compelling negative relationship between the frequency of total repair and the utilization of four-five-word *MI* lexical bundles. The effect size of -2.41 underlines this inverse relationship, suggesting that as the use of such long lexical sequences increases, instances of speech repair decrease. This relationship is further underscored by (1) the substantial dominance weight of 42.75% held by the four-five-word *MI* variable, and (2) the confirmed importance as per random forests (see Figure 8) indicating its substantial contribution to the model's predictive power. This high dominance weight and confirmed importance suggests that the utilization of four and five-word lexical bundles is a key determinant of repair fluency. These results provide strong evidence to assert that greater use of longer lexical bundles is associated with enhanced repair fluency, confirming our earlier prediction that participants who use longer LBs would likely repair their speech less often. This might imply that learners who can readily deploy longer sequences have a firmer grasp on language structures, leading to fewer instances of self-repair.

6.3.2 Shorter (bi- & trigram) lexical bundles

Our findings indicate a complex relationship between the usage of shorter LBs and speech repair. The bigram *MI* variable also emerged as a noteworthy contributor to the model's predictive power, with a dominance weight of 24.04%, despite its relatively weaker negative association with the frequency of total repair. However, the random forests (Boruta) analysis positioned bigram *MI* as the single most important predictor of frequency of total repair. This suggests that as the quality of bigrams increases, the frequency of speech repairs tends to decrease. This finding is consistent with that of Tavakoli and Uchihara (2020) who also found a significant but small negative correlation between their n-gram Factor 2 (association measure of *MI*) and frequency of total repair (r = .308, p = .021). This supports the notion that employing certain high-quality bigrams might indeed result in fewer speech repairs. However, it is important to note that other studies (e.g., Saito et al., 2018; Tavakoli et al., 2020) have reported inconsistent findings regarding repair fluency and oral development. This implies that our finding is not conclusive, highlighting the still-evolving nature of our understanding regarding repair fluency.

The other PVs such as bigram frequency, bigram Proportion, trigram frequency, and trigram proportion exhibited weaker effect sizes and contributed less significantly to the model's predictive power. The Boruta algorithm also indicated ambiguity regarding the role of bigram frequency. This suggests that certain lexical bundles could play a nuanced role in speech repair, warranting more granular analyses in future studies. Accordingly, even though there is some data that backs our first hypothesis about shorter LBs causing fewer speech repairs, the findings imply that the correlation is intricate and more elaborate than we originally thought.

In summary, our findings underscore the potential role of longer lexical bundles in enhancing repair fluency. The findings also hint at the intertwined relationship between lexical bundle length, cognitive processing, and repair fluency, underscoring the need for nuanced pedagogical approaches that account for these interactions. However, given the complexity of linguistic processes, more comprehensive studies encompassing wider aspects of language production and their interrelationships may be essential to fully understand this relationship.

7. Limitations and future research

While our study yields valuable insights, indicating that the use of longer LBs often corresponds with improved aspects of oral fluency, and that shorter LBs may serve distinct roles in communicative competence, the relationship is intricate and multidimensional, and there are several limitations to consider. First, the study's participants come from a narrow range of linguistic backgrounds, which could limit the generalizability of the findings. Second, there is uncertainty regarding the fairly wide confidence intervals reported in the current study's regression analyses. Third, additional uncertainty arises from the frequency-based automatic extraction techniques used to identify contiguous bundles. Our frequency-based LB approach remains limited in that it certainly did not capture all MWSs, such as the ones used infrequently or idiosyncratically, or the ones that blend into surrounding language, in our small dataset. Most of the LBs captured in our approach are in fact structurally and semantically incomplete units. Fourth, some of the LBs captured in our study (e.g., "the problem of global warming") were borrowed from the speaking task prompts, instead of being LBs generated by the learners themselves. Some of the LBs in our list may therefore not be considered conventional multi-word expressions. Fifth, our findings may be limited because our longer text-internal LB measure uses *MI* score. Since *MI* scores were originally developed to measure the collocational strength of two-word collocations and since they do not consider the order of the words (Biber, 2009; Hyland, 2012), they may not be a highly reliable measure for strings of three or more words.

We need future studies to address these limitations. Future research should examine the LB-fluency relationship among learners with diverse linguistic backgrounds. To decrease the width of the confidence intervals and increase the reliability of the regression analyses, employing a bigger sample size may help (see Larson-Hall & Plonsky, 2015). While it might be argued that text-external measures are sufficient for investigating the LB-fluency relationship, our findings suggest that employing text-internal measures or a combined approach can provide a more accurate representation of learner-produced LBs and their relationship with aspects of fluency. Further work is needed to establish the viability of text-internal MWS measures that can extrapolate lower proficiency learners' aspects of fluency. Further research is also needed to examine the relationship between aspects of oral fluency and refined LBs, using LB rating or refinement techniques, such as those employed in Wood and Appel (2014) and Coxhead, Dang, and Mukai (2017), that can help identify more structurally and semantically complete units. Future studies should also employ more sophisticated statistical techniques, such as mixed-effects models which are well-suited to the analysis of learner corpus data (Siyanova-Chanturia & Spina, 2020). Other promising areas for future research include examining how LB usage varies in terms of function and the link to different aspects of fluency across proficiency levels and disciplines.

8. Pedagogical implications and suggestions

Our findings can inform fluency teaching and learning practices in various ways. Familiarity with and use of MWSs, especially longer bundles with high *MI* scores like "positive for our community but", "the problem of global warming", and "I agree with the idea", plays a pivotal role in the development of oral fluency. Such specific sequences often reflect more sophisticated language use and are indicative of advanced proficiency. This underscores the importance of their explicit instruction. To maximize fluency outcomes, educators should prioritize teaching longer (four- to five-word) lexical bundles, preferably with high *MI* scores. High *MI* scores, referring to the statistical strength of association between words in a sequence, indicate the specificity or sophistication of longer bundles. Hence, when we emphasize the teaching of bundles with high *MI* scores, we highlight the importance of sequences that carry more specific and perhaps nuanced meanings. This approach can enhance both speed and repair fluency. However, caution is necessary for lower- and intermediate-level EFL learners, as increased use of certain trigram sequences might lead to temporary speech disruptions. Considering the varying impacts of different lexical bundle lengths on fluency, a tailored teaching approach is advised. This approach should consider both learners' proficiency and the cognitive demands of the sequences. The goal is to optimize the fluency outcomes across different contexts, leveraging the crucial role diverse MWSs play in shaping learner fluency.

For educators seeking a practical method to identify worthwhile LBs for instruction, the following steps offer a guideline:

1. Select a text that aligns with your learners' needs.
2. Use Tom Cobb's web-based "Phrase Extractor" tool (https://www.lextutor.ca/multiwords/phrase/) to extract core two-word collocations with high *MI* scores. Note that this tool focuses on two-word sequences, because *MI* scores were originally designed to measure the strength of association between pairs of words. These two-word collocations often serve as the foundation for longer lexical bundles. For example, using the "Barack Obama" text sample on the Phrase Extractor webpage yields word pairs like "health care" and "carbon pollution".
3. Build upon these foundational collocations using advanced AI-powered tools, such as Microsoft's new Bing search engine (https://www.bing.com/chat). This can help identify how these two-word sequences integrate into longer MWSs like "access to quality health care" or "the biggest source of carbon pollution".[3] These longer sequences often have a nuanced meaning, essential for fluent speech. Consider creating tailored teaching materials, such as bilingual phrase lists, based on these structures to enhance learners' awareness and contextual understanding of extended lexical sequences.

9. Conclusion

The present study is the first attempt to investigate how the use of longer (four- to five-word) LBs and shorter LBs relate to aspects of oral fluency (speed, breakdown, and repair). By systematically employing both text-internal and text-external approaches, our pioneering study sheds light on the role of longer LBs in oral fluency, demonstrating that their use has a marginal positive effect on articulation rate, a potential negative association with the frequency of mid- and end-ASU pauses, and a strong negative relationship with the frequency of total repair. While the longer lexical bundles had positive influences, shorter two- and three-word sequences had unique effects on fluency, potentially causing temporary disruptions and highlighting the nuanced interplay between LB length and fluency outcomes.

Taken together, the combination of findings discussed underscores the nuanced relationship between lexical bundle length, both short and long, and fluent speech production. Our research contributes to the field by demonstrating that while longer LBs marginally enhance speed fluency, they play a more substantial role in reducing speech disruptions and repairs. This suggests a more complex interaction between LB length and fluency than previously assumed, challenging the conventional notion that longer MWSs straightforwardly facilitate fluency. In contrast, LBs, particularly bigrams and trigrams, show a distinct, sometimes counterintuitive, impact on fluency. The use of these LBs, particularly by intermediate-level EFL learners, may lead to slower speech and increased pausing, highlighting a critical developmental stage in fluency acquisition where cognitive processing demands are high. These insights have important implications for MWS research, suggesting a need for a more differentiated approach (e.g., combining text-internal and text-external techniques) in studying the impact of LB length on various aspects of fluency.

In conclusion, our study not only adds a new dimension to the understanding of the relationship between LB length and aspects of fluency but also lays the groundwork for future investigations into the intricacies of this relationship. It encourages a more nuanced view of how MWSs of various lengths contribute to the development of L2 oral fluency.

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).
2. We also tried running a principal component analysis (PCA) to see if we could identify meaningful underlying n-gram factors in a similar way to Tavakoli and Uchihara (2020). However, various attempts at running PCAs found mixed clusters of factor loadings, so we decided to run multiple regression analyses instead of PCA.
3. The prompt we used to generate these example phrases in the new Bing is: *The following collocations have high mutual information (MI) scores computed by a Phrase extractor tool (from https://www.lextutor.ca/multiwords/phrase/): "health care" and "carbon pollution". Please scan the entire Internet and create a separate short list of 5 frequently occurring phrases containing each of these collocations. The phrases should be at least 4 or 5 words each and could be sentence stems or clauses. Provide the list in a two-column table with the phrases in the left column and corresponding Japanese translations in the right column. We intend to use the lists for EFL teaching purposes at a Japanese university, so kindly make them easy to understand.*

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT Plus (GPT-4) to improve the text's readability and ensure it conforms with APA 7 style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: theory, analysis, and applications* (pp. 101–122). Oxford University Press.

Anthony, L. (2022). AntConc (Version 4.0.5) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Appel, R. & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high- and low-proficiency levels, *Language Assessment Quarterly*, *13*(1), 55–71.

Barlow, M. (2015). Collocate (Version 2.0) [Computer Software]. Athelstan. https://athel.com/

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275– 311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286.

Biber, D. & Gray, B. (2013), Discourse characteristics of writing and speaking task types on the *Toefl iBT*® test: A lexico-grammatical analysis. *ETS Research Report Series*: i–128.

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

Biber, D., Conrad, S., & Cortes, V. (2004). 'If you look at. . .: Lexical bundles in university teaching and textbooks,' *Applied Linguistics 25*(3), 371–405.

Boers, F., Eyckmans, J., Kappel, K., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Language Teaching Research, 10*, 245–261.

Boersma, P., & Weenink, D. (2016) PRAAT: Doing phonetics by computer (Version 6.1.39) [Computer program]. https://www.fon.hum.uva.nl/praat/

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics*, *37*, 575–596.

Chen, Y., & Baker, P., (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1, *Applied Linguistics*, *37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Clenton, J., de Jong, N., Clingwall, D. & Fraser, S. (2021). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency for a small group of pre-intermediate Japanese L1 users of English (L2). In Clenton, J. and Booth, P. (Eds.), *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 126–145). Routledge.

Cobb, T. (2012). Phrase extractor (Version 1.2) [Web application]. Retrieved September 18, 2023, from https://lextutor.ca/multiwords/phrase/

Coxhead, A., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks, *Journal of English for Academic Purposes*, *30*, 66–78.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics 2*(3), 223–235.

Dang, T. N. Y., Coxhead, A. & Webb, S. (2017), The academic spoken word list. *Language Learning*, *67*, 959–997.

Davies, M. (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.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, *3*(1), 59–80.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, New Series*, *2*, 225–246.

de Jong, N. (2016a). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*, 113–132.

de Jong, N. (2016b). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 203–218). De Gruyter.

de Jong, N., Steinel, M., Florijn, A., Shoonen, R., & Hulstijin, J. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In Housen, A., Kuiken, F. and Vedder, I. (Eds.) (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 121–142). John Benjamins.

de Jong, N., Steinel, M., Florijn, A., Shoonen, R., & Hulstijin, J. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893–916.

de Jong, N., & Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition*, *41*(1), 227–239.

Ebeling, S., & Hasselgård, H. Learner corpora and phraseology. In Granger, S., Gilquin, G. & Meunier, F. (Eds.), 2015. *The Cambridge Handbook of Learner Corpus Research* (pp. 207–229). Cambridge University Press.

Ellis, N., Frey, E. & Jalkanen, I. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In U. Römer & R. Schulze (Eds.), *Exploring the lexis-grammar interface* (89–114). John Benjamins.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396.

Felker, E., Klockmann, H., & de Jong, N. (2019). How conceptualizing influences fluency in first and second language speech production. *Applied Psycholinguistics, 40*(1), 111–136.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375.

Garner, J., & Crossley, S. (2018). A latent curve model approach to studying L2 N-Gram development. *The Modern Language Journal*, *102*(3), 494–511.

Granger, S. (2019). Formulaic sequences in learner corpora: Collocations and lexical bundles. In Siyanova-Chanturia, A. & Pellicer-Sanchez, A. (Eds.) *Understanding formulaic language: A second language acquisition perspective* (pp. 228–247). Routledge.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–49). John Benjamins.

Gries, S. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies 5*(1), 1–33.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1). https://doi.org/10.18637/jss.v017.i01

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 21–33). Palgrave Macmillan.

Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In M. Mahlberg and V. Wiegand (Eds.), *Corpus linguistics, context and culture (pp. 339*–362). De Gruyter.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21.

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

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809–854.

Kahng, J. (2017). The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39*(3), 569–591.

Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030–1046.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly, 18*, 154–170. https://doi.org/10.1080/15434303.2020.1844205

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lennes, M. (2021). *SpeCT* - The speech corpus toolkit for Praat (Version 1.0.0) [Computer software script]. https://lennes.github.io/spect/

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387–417.

Levelt, W. J. (1989). *Speaking: From intention to articulation.* MIT Press.

Levelt, W. J. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition, 42*, 1–22.

Liakhovitski, D., Bryukhov, Y., & Conklin, M. (2010). Relative importance of predictors: Comparison of random forests with Johnson's relative weights. *Model Assisted Statistics and Applications*, *5*, 235–249. https://doi.org/10.3233/MAS-2010-0172

McGuire, M., & Larson-Hall, J. (2017). Teaching formulaic sequences in the classroom: Effects on spoken fluency. *TESL Canada Journal*, *34*(3), 1–25.

McGuire, M. & Larson-Hall, J. (2021). The contribution of high-frequency multi-word sequences to speech rate and listening perception among EFL learners. *Vocabulary Learning and Instruction*, *10*(2), 18–29.

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters.

Mizumoto, A. (2015). Langtest (Version 1.0) [Web application]. Retrieved July 21, 2023, from http://langtest.jp

Mizumoto, A. (2022). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning, 1–36.*

Myles, F., & Cordier, C. (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.

Nation, P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Nesselhauf, N. (2004). What are collocations? In Allerton, D., N. Nesselhauf & P. Skandera (Eds.). *Phraseological units: Basic concepts and their application* (pp. 1–21). Schwabe.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & W. Schmidt (Eds.), *Language and communication* (pp. 29–59). Longman.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*, 538–553.

R Development Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/

Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, *39*, 593–617.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(4), 357–385.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching, 54*(2), 79–95.

Simpson-Vlach, R. & Ellis, N. C. (2010). An Academic Formulas List (AFL). *Applied*

Linguistics, 31(4), 487–512.

Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning, 70*, 420–463.

Siyanova-Chanturia, A., & Van Lancker Sidtis, D. (2018). What on-line processing tells us about formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 38–61). Routledge.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*, 1–14.

Skehan, P. (Ed.) 2014. *Processing perspectives on task performance*. John Benjamins.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP development: Lexical bundles in the TOEFL iBT writing section. *English for Specific Purposes 12*(3), 214–25. https://doi.org/10.1016/j.jeap.2013.05.002

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics, 49*, 321–343.

Suzuki, S., & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition. 42*(1), 251–251.

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal, 105*(2), 435–463.

Suzuki, Y., Eguchi, M. & de Jong, N. (2022), Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective. *TESOL Quarterly*, *56*, 1290–1319. https://doi.org/10.1002/tesq.3103

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*, 71–79.

Tavakoli, P., Nakatsuhara, F. & Hunter, A. M. (2017). Scoring validity of the Aptis Speaking Test: Investigating fluency across tasks and levels of proficiency. ARAGs Research Reports Online. British Council. https://www.britishcouncil.org/exam/aptis/research/publications/tavakoli

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*, *105*(1), 169–191.

Tavakoli P. & Hunter A. M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, *22*, 330–349.

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and Task Performance* (pp. 239–273). John Benjamins.

Tavakoli, P. & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547.

Tavakoli, P. & Wright, C. (2020). *Second language speech fluency*. Cambridge University Press.

Thomson, H. (2017). Building speaking fluency with multiword expressions. *TESL Canada Journal, 34*(3), 26–53.

Thomson, H., Boers, F. & Coxhead, A. (2019). Replication research in pedagogical approaches to spoken fluency and formulaic sequences: A call for replication of Wood (2009) and Boers, Eyckmans, Kappel, Stengers & Demecheleer (2006). *Language Teaching, 52*(3), 406–414.

Tremblay, A., Derwing, B., Libben, G. & Westbury, C. 2012. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569–613.

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, *24*(4), 540–556.

Uchihara, T., & Harada, T. (2018), Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly, 52*, 564–587.

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (2022). To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech, 65*(2), 311–336.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics, 12*(1), 39–57.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. Continuum.

Wood, D. and Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks, *Journal of English for Academic Purposes, 15*, 1–13.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge University Press.

Zhang, X., Zhao, B., & Li, W. (2023). N-gram use in EFL learners' retelling and monologic tasks. *International Review of Applied Linguistics in Language Teaching, 61*(3), 939–965. https://doi.org/10.1515/iral-2021-0080

Appendix

Table A1

The Most Frequent Four- And Five-Word Recurrent Word Combinations Used by all Speakers (N = 50): Ranked According to *MI* Score

Rank (FREQ)	Rank (*MI*)	N-Gram	FREQ	RANGE	FREQ as 5-Gram	*MI* Score
29	1	positive for our community but	5/4	5/4	4	28.01
4	2	the problem of global warming	9/8	9/7	9	26.03
18	3	I agree with the idea	6/5	6/5	5	23.69
56	4	be positive for your	3	3		22.16
63	5	it is bad for children	3/3	3/3	3	21.02
24	6	be positive for our	5	5		20.46
3	7	will be positive for	11	11		20.42
23	8	next to the elementary school	5/5	5/5	5	20.40
40	9	woman who wear blue	4	4		19.76
2	10	solar panel is the best	16/14*	13/12	14	19.62
12	11	it is not good for	11/8	10/8	6	19.47
82	12	wearing a blue jacket	3	3		19.44
59	13	is good to build casino	3/3	3/3	3	19.07
31	14	elementary school because casino is	4/4	4/4	4	18.94
19	15	so I think we should	8/4	5/4	5	18.68
67	16	put her bag on	3	3		18.65

14	17	bag and ran away	6	6		18.51
13	18	our community but I	6	6		18.22
57	19	black bag and ran	3	3		17.80
11	20	I think solar panel is	7/6	7/6	6	17.79
68	21	she put her bag	3	3		17.72
51	22	elementary school so it is	4/3	4/3	3	17.39
22	23	the idea of the casino	5/5	5/5	5	17.39
55	24	away from elementary school	3	3		17.15
7	25	next to elementary school	7	7		16.98
36	26	the income will be	4	4		16.93
83	27	who is wearing blue	3	3		16.91
73	28	solutions for the problem	3	3		16.84
53	29	a woman who wear	3	3		16.81
15	30	so we should build	6	6		16.27
34	31	not good for children	4	4		16.24
49	32	think we should build	4	3		15.82
46	33	the woman who wear	4	3		15.82
60	34	her bag and run	3	3		15.75
81	35	we should not build	3	3		15.71
72	36	solution for the problem	3	3		15.69
62	37	if we build casino	3	3		15.45
61	38	I disagree with the	3	3		15.44
79	39	we should build casino	3	3		15.26
54	40	after that the woman	3	3		15.16
28	41	far from the school.	4	4		15.09
47	42	think that we should	4	3		15.04
6	43	near the elementary school	8	7		14.96
84	44	with the idea of	3	3		14.94
17	45	to use solar panel	6	5		14.85
21	46	I do not think	5	5		14.61
8	47	from the elementary school	7	6		14.58
66	48	on the ground and	3	3		14.51
78	49	using solar panel is	3	3		14.36
48	50	think the best solution	4	3		14.34
41	51	the elementary school because	4	4		13.81
76	52	to elementary school because	3	3		13.81
27	53	use solar panel is	5	3		13.66
80	54	we should build the	3	3		13.52
26	55	is the best solution	5	5		13.34
32	56	elementary school so it	4	4		13.33
45	57	the best solution is	4	3		13.02
77	58	to elementary school so	3	3		12.97
20	59	think it is good	6	4		12.89
64	60	is the best idea	3	3		12.87

30	61	casino near the school	4	4		12.83
58	62	casino near the elementary	3	3		12.76
9	63	because it is not	7	6		12.74
70	64	solar panel on the	3	3		12.68
43	65	I think that we	4	3		12.61
1	66	I think it is	22	12		12.48
25	67	but I think it	5	5		12.47
71	68	solar panels is the (best)	3	3		12.24
5	69	I think the best	8	7		12.18
39	70	school because casino is	4	4		12.16
50	71	so I think that	4	3		12.12
65	72	it is good to	3	3		11.50
52	73	a casino near the	3	3		11.29
69	74	solar panel is a	3	3		11.22
37	75	think it is not	4	4		11.04
35	76	so I think it	4	4		10.99
16	77	the solar panel is	6	5		10.81
44	78	if the casino is	4	3		10.64
75	79	the elementary school and	3	3		10.60
33	80	it is the best	4	4		10.53
42	81	but I think the	4	3		10.29
10	82	so I think the	7	6		10.14
38	83	think it is the	4	4		8.84
74	84	the casino because the	3	3		8.65

Note. MI = Mutual Information.
*Multiple-frequency figures listed in column 3 represent the individual frequencies of the four-word sequences that make up the longer five-word structure.