Accepted for publication in *Language and Speech* (SAGE)

To what extent is collocation knowledge associated with oral proficiency? A Corpus-based approach to word association

Takumi Uchihara (Waseda University)
Masaki Eguchi (University of Oregon)
Jon Clenton (Hiroshima University)
Kristopher Kyle (University of Oregon)
Kazuya Saito (University College London)

Abstract

This study examined the relationship between second language (L2) learners' collocation knowledge and oral proficiency. A new approach to measuring collocation was adopted by eliciting responses through a word association task and using corpus-based measures (absolute frequency count, t-score, MI score) to analyze the degree to which stimulus words and responses were collocated. Oral proficiency was measured using human judgements and objective measures of fluency (articulation rate, silent pause ratio, filled pause ratio) and lexical richness (diversity, frequency, range). Forty Japanese university students completed a word association task and a spontaneous speaking task (picture narrative). Results indicated that speakers who used more low-frequency collocations in the word association task (i.e., lower collocation frequency scores) spoken faster with fewer silent pauses and were perceived to be more fluent. Speakers who provided more strongly associated collocations (as measured by MI) used more sophisticated lexical items and were perceived to be lexically proficient. Collocation knowledge remained as a unique predictor after the influence of learners' vocabulary size (i.e., knowledge of single-word items) was considered. These findings support the key role that collocation plays in oral proficiency and provide important insights into understanding L2 speech development from the perspective of phraseological competence.

Key words: Collocation, Word Association, Oral Proficiency, Fluency, Lexical Richness

Introduction

Second language (L2) learners with rich mental lexicons are considered proficient in various aspects of communicative abilities (Meara, 1996). Research has provided converging evidence to support the essential role of vocabulary in reading (e.g., McLean, Stewart, & Batty, 2020), listening (e.g., Vafaee & Suzuki, 2020), writing (e.g., Yang, Sun, Chang, & Li, 2019), and speaking (e.g., Uchihara & Clenton, 2020). In exploring the vocabulary-proficiency link, researchers are often faced with a challenge in defining and operationalizing the construct of word knowledge. Earlier studies tend to rely on vocabulary measures of single-word items focusing on knowledge of form-meaning connections (i.e., vocabulary size) using traditional test formats such as multiple-choice (Vafaee & Suzuki, 2020) or yes/no check tasks (Uchihara & Clenton, 2020). This line of research has suggested that learners who can demonstrate knowledge of lower-frequency or more difficult words have richer lexicons (Laufer & Nation, 1995) and more advanced language skills (Miralpeix & Muñoz, 2018). However, the richness of a learner's lexicon is far from simple to delimit using quantity or size of vocabulary alone; quality and/or depth of vocabulary also needs to be considered (Schmitt, 2014). It is possible, for instance, that learners who know a larger number of individual word meanings may not be orally fluent due to lack of knowledge of pronunciation or collocations frequently used in spoken discourses. To address this gap and advance our understanding of the relationship between vocabulary and proficiency, the current study investigates the relationship between collocation knowledge and oral proficiency in spontaneous communication. This study is innovative in two significant ways. First, we measure productive collocation knowledge using a free word association task and adopt a corpus-based approach in analyzing collocational relations between responses and stimulus words. Second, we employ multiple measures to assess two linguistic aspects of oral proficiency—fluency and lexical richness, both of which are considered particularly relevant to learners' vocabulary knowledge (Hilton, 2008; Koizumi & In'nami, 2013; Kormos, 2006; Laufer & Nation, 1995; Miralpeix & Muñoz, 2018; Uchihara & Clenton, 2020; Uchihara & Saito, 2019). With the measures employed in this study, we investigate whether and to what extent L2 speakers with collocationally rich lexicons (or phrasicons) are temporally and lexically proficient in oral production.

Collocation Knowledge and L2 Oral Proficiency

Collocation in its broader sense refers to combinations of two or more words that co-occur very frequently in a target language. It includes various forms of lexical strings, such as idioms (e.g., *kick the bucket*), restricted collocations (e.g., *strong coffee*), phrasal verbs (e.g., *hung up*), binomials (e.g., *bride and groom*), proverbs (e.g., *birds of a feather flock together*), and lexical bundles (e.g., *and so on, I think it*). Defining the construct of collocation is a great challenge due to the many different linguistic characteristics used in research, including the contrast of lexical (e.g., verb + noun) vs. grammatical (e.g., preposition + noun) combinations, fixedness, semantic transparency,

and arbitrariness (Henriksen, 2013). Despite the great variety of collocation types, one feature underlying all such variants relates to the fact that collocations are processed as independently represented or entrenched chunks in the mental lexicon (Siyanova-Chanturia & Martinez, 2015). Research supports the psycholinguistic validity of collocations, meaning that both L1 and L2 speakers can process collocations more rapidly and accurately than novel strings of words in receptive and productive language tasks (Ellis, Simpson-Vlach, & Maynard, 2008; Sonbul, 2015; Tremblay, Derwing, Libben, & Westbury, 2011; see Siyanova-Chanturia & Van Lancker Sidtis, 2018 for the review).

Corpus-based research also acknowledges the importance of collocation knowledge as part of a learners' lexicon in language use. In exploring the relationship between collocation use in written production, Durrant and Schmitt (2009) found that native writers tended to use more sophisticated and low-frequency collocations (indexed by higher mutual information [MI] scores) than non-native writers. Granger and Bestgen (2014) compared L2 learners of intermediate and advanced proficiency levels and supported Durrant and Schmitt's findings. Although limited in number, compared to writing studies, speaking studies examining collocation use (operationalized as lexical bundles) also point to the important role of collocation knowledge in L2 proficiency. Kyle and Crossley (2015) found that orally proficient speakers produce a greater number of target-like combinations frequently used by L1 speakers. Longitudinal studies by Kim, Crossley, and Kyle (2018) and Garner and Crossley (2018) supported Kyle and Crossley's findings that L2 speakers produce a gradually increasing proportion of collocations that are frequent in an L1 spoken reference corpus; L2 speakers also produce more high-frequency two-word sequences over time. In line with writing studies, recent studies have also demonstrated that proficient L2 speakers are more likely to use strongly associated, sophisticated collocations (indicated by higher MI scores) in spontaneous oral production (Eguchi & Kyle, 2020; Saito, 2020). While revealing of the important link between collocation and proficiency, these studies are limited in that they compared L2 proficiency and collocation "use" rather than "knowledge" measured with separate elicitation tasks (e.g., multiple-choice or lexical decision tasks, see Gyllstad & Schmitt, 2018 for a review of collocation elicitation tasks).

The important role of collocation knowledge in oral proficiency does not only have empirical support but also a sound theoretical base. According to the speech production model (Levelt, 1989; Kormos, 2006), speech production involves at least three different stages including conceptualization, formulation and articulation. At the conceptualization stage, speakers plan speech content and its manner of presentation (generation of preverbal message). The preverbal message then proceeds to formulation where lexical selection and grammatical encoding take place. In this stage, appropriate lemmas are activated in the mental lexicon, lemmas are placed into syntactic surface structures, and morphophonological and phonetic encoding are conducted. This product is then submitted to articulation where the phonetic plan is executed before speech is produced. For successful oral performance, the formulation stage is where collocation matters. Phraseologically

proficient speakers are able to retrieve a sequence of words efficiently and rapidly rather than build up individual words to construct that sequence, making their speech fast without too many pauses (Tavakoli & Uchihara, 2020). Such efficiency in lexical retrieval might free up cognitive resources then available to improve other aspects of speech performance such as lexical and grammatical accuracy or complexity (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006).

Corpus-Based Approach to Word Association

Researchers have long used a word association task for the purpose of eliciting L2 responses and assessing learners' collocation knowledge (e.g., Clenton, 2015; Fitzpatrick, 2006; Söderman, 1993; Wolter, 2001; Zareva & Wolter, 2012). In a word association task, learners produce words that come to mind prompted by provided stimuli, and the relation between the response and stimulus is analyzed to determine whether they are collocationally or syntagmatically related (e.g., *hot* > *water*) or represent other types of relations: semantic (e.g., *hot* > *warm*) or phonological (e.g., *hot* > *dot*) (Fitzpatrick, 2013). It is customary that researchers in this area rely on human judgement in determining the types of stimulus-response relations in a categorical manner (e.g., collocational or not).

Adopting this categorical and subjective approach, researchers have investigated the relationship between collocational responses and proficiency. For example, Fitzpatrick (2006) elicited word association responses from native and non-native advanced speakers and compared the proportion of their collocational responses. Fitzpatrick found that native speakers generated a greater proportion of collocational associations compared to non-native speakers. Further analysis of the non-native group demonstrated a significant and positive correlation between the proportion of collocational responses and L2 proficiency measured with a receptive vocabulary size test (the Eurocentres Vocabulary Size Test, EVST; Meara & Jones, 1990). Similarly, Clenton (2015) reported on a study consistent with Fitzpatrick with a finding that successful demonstration of target-like collocational responses to stimuli were significantly correlated with learners' overall receptive vocabulary size (EVST scores).

One alternative method, emerging with recent technological advancements, to identify collocational associations between responses and stimuli is a corpus-based approach. In this approach, the judgement of whether two or more words qualify as collocations is made by frequency of co-occurrence of the words observed in a collection of natural language use or corpus. When a certain combination of words appears more frequently than another in a given corpus, the former is considered to be the more standardized combination (hence, seen as collocation) than the latter. Another technique related to but improved on this raw frequency count is a measure of strength of the association between constituent words. This measure is based on the raw frequency count but controls for the relative frequencies of the words comprising combinations by measuring the conditional probability of word co-occurrence. The most widely used association measures in corpus-linguistic research are the t-score and MI. However, these two measures highlight different

sets of collocations (Durrant & Schmitt, 2009; Granger & Bestgen, 2014). In particular, a t-score tends to emphasize high-frequency combinations (e.g., *good example, hard work*), whereas MI tends to emphasize low-frequency combinations (e.g., *tectonic plates, immortal souls*). Although emerging evidence has suggested that MI is generally an indicator of phraseological sophistication (Paquot, 2019; Saito, 2020), more research is needed to determine the true value of the MI index as a collocation measure in relation to the limitation that it might be largely influenced by test takers' knowledge of low-frequency words (i.e., vocabulary size).

Use of corpus-based frequency and association measures offers advantages over traditional approaches when identifying collocations in word association research. First, a corpus-based approach involves less subjectivity, compared to human judgements, as frequency or association scores are often computed automatically with text analysis software. More importantly, computation of corpus-based scores allows operationalizing the continuous nature of collocational status (Ellis et al., 2008) and avoids the binary identification of collocations. To the best of our knowledge, Zareva and Wolter (2012) is the only study exploring the collocation-proficiency relationship using a word association task and corpus-based measures of collocation. Zareva and Wolter employed t-scores generated from the Bank of English corpus to determine the collocational status between responses and stimulus words. They found that intermediate learners of English produced more collocational responses than either advanced learners (though the difference failed to reach statistical significance) or native speakers (the difference reached statistical significance). Zareva and Wolter's findings countered the prediction that higher proficiency speakers produce more collocational associations than lower proficiency speakers, raising the question of whether the collocational network is linked with L2 proficiency. However, their findings should be interpreted with caution because the study adopted binary identification of collocational status (i.e., a cut-off score rather than a mean score) as well as only using t-scores to measure association strength.

Current Study

In order to advance our understanding of the relationship between phraseological competence and language proficiency, the current study adopts a new and unique approach to the measurement of collocation knowledge and examines the relationship between collocation knowledge and various aspects of oral proficiency. We employ a corpus-based approach to word association to assess how rich L2 speaker collocational networks are. Demonstration of target-like collocational responses to stimuli indexed by frequency and association measures (t-score and MI) is expected to indicate learners' rich lexicons (or phrasicons), wherein a great number of collocations are readily activated and available for L2 use and performance. According to the speech production model (Kormos, 2006; Levelt, 1989), such phraseologically competent speakers are hypothesized to have the ability to access and retrieve words efficiently and rapidly, making them more fluent and able to demonstrate more lexically rich language during spontaneous speech. Based on this

hypothesis, we attempt to examine whether and to what extent collocation knowledge is associated with oral fluency and lexical richness.

Method

Participants

Participants were 40 first-year Japanese undergraduate students (26 females, 14 males) at a prestigious university in eastern Japan. All participants had studied English for six years starting at Grade 7. Their general English proficiency varied according to the Test of English for International Communication (TOEIC) scores, comprising reading and listening sections (*M* = 697.9, *SD* = 125.7, *Range* = 515 to 890) and to the internet-based Test of English as a Foreign Language (TOEFL iBT) scores (*M* = 71.7, *SD* = 13.7, *Range* = 40 to 96). In conjunction with the CEFR benchmarks, these test scores indicated that our participants' proficiency levels ranged from B1 to C1. The current dataset therefore ranged from low to high-proficient learners of L2 English in English-as-a-Foreign-Language settings.

Word Association Task

Lex30 (which was originally designed to elicit single-word items) was used as the word association task (Meara & Fitzpatrick, 2000). It is essentially "an effective and efficient elicitation tool, which can be combined with a range of analytical measures" (Fitzpatrick & Clenton, 2010, p. 551). Following this suggestion, researchers have used Lex30 as a word association task along with other tasks such as proficiency measures and found it an effective tool to elicit free association responses (Clenton, 2015; Fitzpatrick, 2012; Walters, 2012). The task format was considered suitable for eliciting a large number of productive collocations from low- to high-proficiency learners in a time-efficient manner (Barfield, 2009). Participants first worked through a practice set of three different (non Lex30) stimulus words and then completed the Lex30 task individually in a laboratory setting. A total of 30 stimulus words were given and the participants were asked to write the first four words each stimulus made them think of (see Supporting Information-A for the stimulus words and the task format).

The Lex30 stimuli have three elements that make them particularly suitable for the current study. First, all the stimulus words are highly frequent words taken from the first 1,000 most frequent word list (Nation, 1984). Thus, these words were likely to be known to our participants. This feature is particularly important for the objective of the current study, as high-frequency stimulus words are reported to elicit more collocational responses than low-frequency stimulus words (Nissen & Henriksen, 2006). Second, the stimulus words do not tend to elicit a single, dominant primary response (e.g., *black* > *white*, *bread* > *butter*), maximizing the possibility of generating a wide variety of responses. Third, they tend to elicit low-frequency words so that participants can produce a broad range of response words. The last two of these criteria were determined according to data from

the Edinburgh Associative Thesaurus (Kiss et al., 1973) and Nation's (1984) word lists (see Meara & Fitzpatrick, 2000, p. 22-23 for more detailed information).

Following Playfoot et al.'s (2016) and Henriksen's (2008, pp. 30-31) suggestions as to the importance of spontaneity in responding to stimulus words, we imposed a time constraint equally for each stimulus word. Participants were asked to produce a maximum of four responses to each stimulus within 30 seconds. One of the researchers timed the elicitation from the onset of writing a first response to the stimulus to the end of writing a fourth response. When failing to produce a total of four responses or finishing it within the time limit, they were asked to move on to the next stimulus word. The time restriction was necessary in order to (a) avoid strategy use, (b) ensure that each stimulus receives an equal amount of attention from participants, and (c) lead them to complete each response one at a time from the top (1: *attack*) to the bottom (30: *window*). The participants produced on average more than 100 words in response to 30 stimulus words (M = 110.25, SD = 7.78, *Range* = 85 to 120).

Corpus-Based Collocation Measures

Though not mutually exclusive, two approaches—frequency (corpus-based) and semantic (phraseological) traditions—have long been applied to determining collocations (Boers & Webb, 2018; Granger & Paquot, 2008). For this study, we adopted the frequency-based approach, not just because of the objectivity and consistency that this method provides, but because it allows computation of frequency and strength-of-association scores reflective of a continuous nature of the collocational status. The current study evaluates collocations using three statistical measures—raw (absolute) frequency counts and two strength-of-association measures, t-score and MI score—commonly employed in corpus-linguistic research (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Gablasova, Brezina, & McEnery, 2017; Granger & Bestgen, 2014). Collocation frequency is the simplest measure as it counts pure co-occurrence of the node and collocates. Prior to statistical analysis, the frequency score was log-transformed in order to control for Zipfian effects common in word frequency analysis. Although studies show important relationships between frequency (as a proxy of repetition in usage experience) and non-native speakers' knowledge of collocation (Durrant, 2014; González Fernández & Schmitt, 2015; Nguyen & Webb, 2017), frequency count is biased in that it does not distinguish textual co-occurrences at chance level from true collocations. For instance, high frequency words (e.g., *I* and *the*) can co-occur very frequently in close proximities to each other not owing to the association between the unit, but by virtue of the frequencies of the individual components. In short, absolute frequency index may include false positives when component words are highly frequent. The effects of such false positives can be mitigated by using association score such as t-score, which provides "adjusted frequency" (Gablasova et al., 2017, p. 165). Here, such frequent combinations as "*I*" and "*the*" are downgraded because of small differences between observed frequencies (O) and expected frequencies (E). For t-score, mean scores among all the word association responses are calculated as follows:

$$t - score = \frac{O - E}{\sqrt{O}}$$

We also used Mutual Information (MI), "a logarithmic scale to express the ratio between the frequency of the collocation and the frequency of random co-occurrence of the two words in the combination" (Gablasova et al., 2017, p. 163). MI highlights combinations that are either exclusive or rare in its constituents; thus, it should be noted that some collocations (e.g., *okey dokey*) are overemphasized over others due to infrequent nodes and/or collocates (Gablasova et al., 2017, p.160). Mean MI scores are calculated as follows:

$$Mutual\ Information = log_2 \frac{O}{E}$$

Scoring Procedure

In deriving the statistical measures of collocations, the current study used the Corpus of Contemporary American English (COCA; Davies, 2009) with a window span of Left 5 – Right 5 (node word frequency corrected [Evert, 2005]). The window span is determined based on the recent corpus-based lexical network studies in order to increase the number of possible collocational responses to be identified (Baker, 2016; Brezina, McEnery & Wattam, 2015).

To create collocation lists, it is generally considered desirable to search the word forms as they appear in the corpus (e.g., *run*, *runs*, *ran*) separately, given that collocational behaviors are arguably type-specific (Hoey, 2005; Sinclair, 1991). However, this procedure cannot be fully applied to word association data because stimulus words are predetermined and fixed so that researchers have no control over the word forms participants think of in producing associated responses. Therefore, this study followed Zareva and Wolter's (2012) suggestion and computed statistical measures with the lemmatized cue word as the node-word (e.g., *attack*, *attacks*, *attacking* as the same word) and non-lemmatized responses (or types) as collocates.

In word association research, there are two approaches to eliciting and analyzing response data. One is the discrete elicitation approach where participants are required to provide only one word per stimulus (Fitzpatrick, 2006; Jiang & Zhang, 2019), and another is the continuous approach where participants provide multiple responses per stimulus (Clenton, 2015; Fitzpatrick, 2012; Kruse, Pankhurst, & Smith, 1987). We therefore calculated frequency and association scores separately for the first response as well as the total response data.

Oral Task

Spontaneous speech data were elicited using a picture narrative task, a Suitcase Story task (Derwing, Munro, Thomson, & Rossiter, 2009). The rationale behind this choice of speaking task is that communication through an act of describing or narrating accounts for a large proportion of conversation in daily life (Willis & Willis, 2007). In addition, using the oral narrative task can enhance the comparability of the current study, because the task has been most extensively employed in the field of L2 speech research with a view of investigating oral fluency and lexical performance

(Daller, Van Hout, & Treffers-Daller, 2003; Li & Lorenzo-Dus, 2014; Saito, Webb, Trofimovich, & Isaacs, 2016; Trofimovich & Isaacs, 2012; Uchihara & Clenton, 2020; Vermeer, 2000).

In this picture narrative task, the participants describe a sequence of eight-frame pictures without any explicit time restriction after spending approximately one-minute familiarizing themselves with the pictures. The story consists of two strangers carrying suitcases identical in appearance, bumping into each other at the corner of a city street, inadvertently swapping their suitcases, and later finding out their mistake when opening the other's suitcase (accessible at https://www.iris-database.org/iris/app/home/search?new=true).

Speech recordings were carried out individually in a sound-proof laboratory at the university, with each elicited speech sample digitally stored as a WAV file. The total length of each story ranged between 105 and 251 seconds. For lexical analysis, a total of 40 recorded narratives were transcribed with all orthographic markings of pausing (e.g., *uh, um, oh, ehh*) removed and obvious pronunciation errors fixed (e.g., *the story* for *the stoly*). The transcripts in length ranged between 57 and 208 words. All transcripts were then submitted for lexical analysis.

Oral Proficiency Measures

To measure the multifaceted nature of L2 oral proficiency, many scholars have emphasized the importance of adopting both subjective and objective approaches towards analyzing speech samples (Saito & Plonsky, 2019). Following this line of thought, we adopted listener's expert judgements of speech and acoustic analysis measures from the perspectives of perceived fluency and lexical richness. Human judgements of fluency and lexical richness were conducted by trained raters in light of speech rate and lexical variety and sophistication, respectively.

In objective measures of utterance fluency, we adopted one speed fluency measure (articulation rate) and two breakdown fluency measures (silent pause ratio and filled pause ratio). This decision was made on the basis of the operationalization of utterance fluency proposed by Tavakoli and Skehan (2005)—speed fluency (e.g., speech rate, articulation rate), breakdown fluency (e.g., length of run, number of pauses, length of pauses), and repair fluency (e.g., self-corrections, false starts, repetitions, hesitations)—as well as earlier studies suggesting an important relationship between vocabulary knowledge and particularly these two fluency constructs (i.e., speed and breakdown fluency) (De Jong, 2016; Koizumi & In'nami, 2013). Thus, fluency was assessed in terms of listeners' perception of speech rate and three utterance fluency measures (articulation rate, silent pause ratio, and filled pause ratio). Lexical richness was measured in terms of listeners' perception of lexical variety and sophistication and three lexical measures (diversity, frequency, and range).

In conducting objective analysis of richness of vocabulary use, we acknowledge that lexical richness is a composite construct and a great number of measures are available for research purposes (Kyle, 2020; Read, 2000). Among many alternatives, we selected three measures (diversity, frequency, and range) on the basis of robust empirical evidence suggesting that these three measures

are sensitive to and reflective of L2 proficiency and development (Crossley, Salsbury, McNamara, & Jarvis, 2011; Crossley, Subtirelu, & Salsbury, 2013; Daller et al., 2003; Kyle & Crossley, 2015). Diversity scores for each speaker were automatically computed using Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), and frequency and range scores were automatically computed using the Tool for the Automatic Analysis of Lexical Sophistication 2.2 (TAALES; Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2018). In computation of lexical frequency and range scores, we chose spoken corpora as reference (COCA sub-corpus that represents the spoken discourse using the transcriptions of TV programs; Davis, 2009) and transformed indices of the two scores (log transformed data) in order to control for Zipfian effects common in word frequency analysis. The choice of spoken corpora reflected research findings indicating a gap in L2 learners' vocabulary size between spoken and written modes (Milton & Hopkins, 2006; Uchihara & Harada, 2018) and differences in lexical profiles between spoken and written discourses (Dang, Coxhead, & Webb, 2017). Accordingly, we followed the procedures adopted by previous studies by maintaining the consistency between the modality in which L2 words were elicited (oral narrative) and the modality of reference corpora (Eguchi & Kyle, 2020; Saito, 2020; Tavakoli & Uchihara, 2020; Uchihara & Clenton, 2020).

Fluency Measures

Expert ratings. To conduct expert rating of perceived fluency, five native speakers of English (2 males, 3 females) were recruited at an English-speaking university in Montreal, Canada. All raters were graduate students in MA Applied Linguistics with extensive experience on L2 speech analyses of this kind (M_{age} = 28.5 years). They also reported ample English-as-a-Second-Language experience (*M* = 3.5 years; *Range* = 0-7 years). According to Isaacs and Thomson's (2013) definitions, these participants could be considered as linguistically-trained raters.[1]

They listened to and evaluated each sample for one specific construct of L2 speech (i.e., optimal tempo). Building on the notion of perceived fluency in existing L2 fluency literature (e.g., Segalowitz, 2016), and following Saito, Trofimovich and Isaacs's (2017) definition, optimal speed refers to how quickly or slowly someone speaks. It is well known that speaking very quickly can make speech harder to follow but speaking too slowly can have the same effect. A good speech rate should therefore sound natural and be comfortable to listen to (Munro & Derwing, 2001).

As operationalized in Saito et al. (2017), the rating was implemented using the MATLAB-based software. All samples were played in a randomized order. Upon hearing each sample, the listeners were asked to provide fluency ratings (i.e., optimality of speed) using a moving slider. Each end of the continuum had a frowny face (indicating too slow or too slow) and a smiley face

[1] In Saito et al. (2017), it was shown that naïve and trained listeners behaved differently especially as to the assessments of specific phonological features (e.g., segmental and prosodic accuracy, temporal fluency). Thus, it is important to recruit listeners with homogeneous backgrounds in order to attain consistent assessment scores. Indeed, we found relatively high agreement among the five listeners' fluency judgements that we recruited, $\alpha = .90$.

(indicating adequate). Depending on where the slider was located, a 1000-point grading scale was used to indicate L2 proficiency levels (*0 = too fast or slow, 1000 = optimal speed*). In view of statistical analyses, the resulting scores were recorded from 0.001 to 1.000.

Breakdown fluency. For breakdown fluency, measures of both silent pause (i.e., silent pause ratio) and filled pause (i.e., filled pause ratio) were calculated. Building on De Jong et al.'s (2012) study, a silence of longer than 0.35 seconds was counted as a silent pause, and a silent pause ratio was calculated as the percentage of silent pausing time in the total speaking time. In addition to silent pause ratio, we also computed a filled pause ratio. After checking the transcripts as well as speech data and counting the number of filled pauses (e.g., *uh, um*), the total number of filled pauses was divided by the total number of words to then calculate the filled pause ratio.

Speed fluency. For speed fluency, we measured articulation rate (total number of syllables divided by speaking time excluding pauses) as "a pure speed measure" (De Jong et al., 2012, p. 136). Articulation rate is unlikely susceptible to variability in task complexity and appears to reflect "a task-independent articulatory skill" (p. 125).

Lexical Richness Measures

In conducting human rating of lexical richness, we followed an analysis procedure adopted by lexical proficiency studies (Crossley et al., 2011; Saito et al., 2017). To avoid raters' judgements influenced by phonological variables, raters are required to read transcribed samples rather than listen to the samples in lexical analysis. Following this procedure and using a 1000-point sliding scale, the same five trained raters in the audio sessions above evaluated each transcript based on lexical richness. As defined in Saito et al. (2017), lexical richness refers to how sophisticated vocabulary use is. If the speaker uses a few simple, unnuanced words, the speech lacks lexical richness. However, if the speaker's language is characterized by varied and sophisticated uses of English vocabulary, the speech is lexically rich.

Diversity. We computed lexical diversity as the variation of words in a text. Although lexical diversity is normally defined as the number of different words used by a speaker or writer (type-token ratio), the reliability of such measures is questioned due to its dependency on text length (the longer the texts, the lower the values). To circumvent this operational problem, we decided to employ a more sophisticated measure of lexical diversity, or the Measure of Textual Lexical Diversity (MTLD). MTLD involves indices that are mathematically transformed to account for text length so that the computed values can be adequately independent of text-length effect (McCarthy & Jarvis, 2010). A higher score of this measure indicates that speakers produce more lexically diverse language.

Frequency. Word frequency is a traditional construct of lexical sophistication and both oral and written use of infrequent words is considered as an important predictor of L2 lexical proficiency (Daller et al., 2003; Laufer & Nation, 1995). Word frequency scores were calculated as the sum of

log-transformed frequencies divided by the number of word tokens receiving frequency scores. A higher score of this measure indicates that speakers produce more high-frequency words.

Range. Range, referred as contextual diversity alternatively, assesses how widely a word is used across different genres/registers by counting the number of documents (or sub-corpora) in which the word appears in a given reference corpus. Use of L2 words that occur in a narrower range (e.g., specialized vocabulary) has been found to predict L2 proficiency (Crossley et al., 2013; Kyle & Crossley, 2015). Range scores in TAALES were calculated as the sum of log-transformed range counts (i.e., the number of documents in the corpus) divided by the number of word tokens receiving range scores. A higher score of this measure indicates that speakers use words that appear across a wider range of contexts, or more general than specialized vocabulary.

Analysis

First, we examined the validity of our word association task as a tool to elicit collocation by computing the proportion of collocation scores given to response data (i.e., coverage diagnosis: number of responses given collocation scores / number of total responses). The result showed that our corpus-based collocation measure scored most of the responses from the participants (> 80%), *M* = .85, *SD* = .057, *Range* = .71-.96, which supports our approach to measuring collocation with corpus-based metric. Second, interrater reliability for fluency and lexical richness rating was analyzed. Cronbach alpha was high for optimal speed (α = .90) and lexical richness (.85), respectively. For each category (optimal speed, lexical richness), all the rating scores were averaged to yield mean scores per speaker. Second, the descriptive statistics of both collocation measures and oral proficiency scores (fluency and lexical richness) were computed as summarized in Table 1.

In order to investigate the relationships between collocational knowledge measures and oral proficiency measures, we employed a Bayesian approach to the statistical modeling. This approach was taken (a) because it allows estimation of posterior distributions and their uncertainties in an intuitive manner (Kruschke, 2014; McElreath, 2020), and (b) because we prioritized estimations and interpretations of model parameters over significance testing (Norouzian et al., 2018). For a recent application of Bayesian regression in L2 research see Saito et al., (2020).

Bayesian robust correlation. Bayesian robust correlation analyses were conducted using a *rstan* package (Stan Development Team, 2020) in R (R development Core Team, 2014). Stan is a specialized programming language developed for statistical inferences, or Bayesian statistics in particular. We conducted correlation analyses using a freely available stan code robust.mcmc.cor.R (Baez-Ortega, 2018). This code was adopted in the current study because it allowed robust estimates of correlation coefficient rho. In order to conduct the correlation analyses, therefore, the model incorporated multivariate student-t distribution as a prior distribution, heavy tails of which made posterior distributions more robust to the presence of outliers, in comparison to assuming multivariate normal distributions. For other parameters, very weakly informative priors were used (i.e., variances, or sigmas, in the covariance matrix were set to have a half-normal distribution with

the scale of 100). Four Markov chains were implemented with 5,000 iterations (including 500 warm-up iterations). To assess the convergence of the model, trace plots and R-hat values were examined (R-hat values lower than 1.01 was taken to indicate convergence; Vehtari et al., 2020).

An advantage of Bayesian analyses includes its capability for providing full information about the parameter estimates (McElreath, 2020); that is, a posterior distribution represents probability density in which a population parameter is (more or less) likely to fall. This characteristic offers flexible means to test hypotheses in question. For instance, one can set multiple thresholds guided by substantially meaningful effect sizes and obtain probabilities of posterior distribution that exceed (or fall short of) those thresholds. Since we were interested in examining the extent to which collocation knowledge and oral proficiency measures were associated, we interpreted the magnitude of the correlation coefficients through both the point estimates (the mean of the posterior distribution) and 95% Credible Intervals (CIs) of the posterior distributions (as an analogue to conventional confidence interval estimates). To further benefit from the flexibility of Bayesian posterior distribution, we also computed two additional metrics to supplement our interpretations: (a) percentages of the posterior draws that could be considered practically equivalent to the null hypothesis Pr(|*rho*| < .05), or Region Of Practical Equivalence (ROPE; Kruschke, 2014); and, (b) the posterior probabilities of correlation coefficients exceeding the lower bound for the small effect size according to Plonsly & Oswald (2014), or Pr(|*rho*| ≥ .25). The magnitude of correlation coefficients was interpreted according to Plonsky and Oswald's (2014) effect-size benchmark: .25 (small), .40 (medium), and .60 (large).

Bayesian Linear Regression. To further test if the collocation knowledge explains oral proficiency measures above and beyond potential covariates such as vocabulary size (i.e., knowledge of single-word items), we conducted follow-up multiple linear regressions using the *brms* package in R (Bürkner, 2017). Here, vocabulary size was operationalized as the average (logged) frequency of all association responses based on the spoken sub-section of COCA, under the assumption that participants providing more low-frequency single-word items were considered to have larger vocabulary sizes (Fitzpatrick & Clenton, 2010; Meara & Fitzpatrick, 2000). For each of the meaningful bivariate correlations between a pair of an oral proficiency and a collocation knowledge measure, a set of four models were constructed which test the tenability of the relationships. The three models were: (a) unconditional (intercept-only) model with no predictor, (b) collocation-only model, (c) vocabulary-size-only model, and (d) collocation plus vocabulary size model. These four alternative models each corresponded to the following competing hypotheses, or the data generating processes: (a) neither collocation nor vocabulary size predicts oral performance; (b) collocation is a better predictor than vocabulary size; (c) vocabulary size is a better predictor than collocation knowledge; and, (d) both collocation and vocabulary size have unique contributions to oral performance. Subsequent to the model building, analysts may choose the "best-fitting" model semi-automatically using information criteria (i.e., model *selection*; McElreath, 2020), or alternatively, use such information criteria as a tool to learn about the data generating processes each model attempts

to capture (i.e., model *comparison*; McElreath, 2020). The difference between the two approaches will become noticeable when differences among the competing models are small. For instance, one may eventually select a simpler model among comparable alternatives, following the principle of Ockham's razor. However, some statisticians (e.g., McElreath, 2020) advise against selecting the "best" model solely based on its predictive accuracy (i.e., only using information criteria). This advice follows because (a) information criteria inherit uncertainties from the models, and (b) information criteria do not consider the causal structure behind the data. Instead, McElreath (2020) recommends laying out alternative statistical models reflecting possible data generating processes (as already specified above as [a]-[d]) and learning from the relative uncertainties associated with them. The latter approach would allow researchers to avoid making a strong claim about the importance of a particular predictor found in the "best" model while implicitly overlooking alternative interpretaions. In this study, we adapted the model *comparison* approach recommended by McElreath (2020). Specifically, we first computed Leave-One-Out Cross-validation (LOO) Information criterion (LOOIC) for each model to estimate the predictive accuracy of the model. LOOIC is one of the recommended alternative metrics in Bayesian regression analysis, which estimates the degree to which a model captures the population-level regularities (and ignore the sample-specific noises) in the data via cross validation (see McElreath, 2020). Next, we used the LOOIC scores to compute model weights through *loo_model_weight()* function in the *brms* package (Bürkner, 2017) primarily because the values of LOO should be interpreted in relative term and we have specified limited sets of competing models. In the *loo_model_weight()*, the model weights were calculated using Bayesian stacking approach by default (Yao et al., 2018) in order to find an optimal linear combination (i.e., weights) of alternative models so that it minimizes the LOO. The relative weights of the four models (a-d) were then carefully *compared* and *interpreted* to evaluate the plausibilities of the completing hypotheses (i.e., the underlying data generating processes), rather than *selecting* the best model (Bayesian stacking is not developed for model selection; Yao et al., 2018).

In constructing each regression model, both the predictors and outcome variables were standardized so as to interpret the regression parameter in a standardized unit (analogous to standardized beta). For each model, we specified weakly informative prior distributions. Specifically, for any slope parameters, student-t distribution with the degree of freedom of three, the mean of zero, and the standard deviation of one was used, which is recommended by the Stan developer team. The prior distributions for the intercept and the residuals were estimated using the *brms* package (Bürkner, 2017). As in the correlation analyses, a total of four chains were implemented (5,000 iterations but with 1,000 warm-up iterations for each). R-hat values (< 1.01; Vehtari et al., 2020) and trace plots were examined for convergence diagnosis. To further ensure that the local convergence was not an issue, we subsequently doubled the MCMC iterations (9,000 iterations with 1,000 warm-up) and checked for the stability of parameter estimates. Finally, since the choice of a given prior distribution may have a huge impact on the parameter estimation particularly when the dataset is

relatively small (McElreath, 2020), we conducted sensitivity analyses by fitting a series of models using different values for the prior distribution. This allowed us to see the degree to which our prior selection may have impacted the parameter estimation and substantial interpretations of the models. Adapting the rcode used in Saito et al. (2020), we conducted sensitivity analyses on each of the best predictor models using both wider and narrower distributions. Specifically, we fit a total of seven additional priors for each: one Flat prior and six student-t distribution with varying scale values (i.e., 10, 3, 2, 0.9, and 0.8). The readers are referred to the online supplementary material (available at https://osf.io/vfcx2/) for the full information of both correlation and regression analyses, including trace plots of MCMC samples, density plots of the posterior distributions, results of the sensitivity analysis, etc.

Results

Bayesian Robust Correlation

In order to explore the relationship between collocation knowledge and oral proficiency, correlation analyses were conducted between three collocation measures (frequency, t-score, MI score) and eight speech measures (Fluency: fluency rating, articulation rate, silent pause ratio, filled pause ratio; Lexis: richness rating, diversity, frequency, range). In what follows, the results of correlation analyses are presented for total response and first response data separately.

Total response. For collocation measures based on the word association responses (including all responses) and oral proficiency measures (see Table 2), the Bayesian correlation analyses indicated that overall there was a large amount of uncertainty in the parameter estimations, many of which included zero in the 95% Credible Intervals (CIs) (See online supplementary material for plots of the posterior distributions and trace plots). Still, three of the 24 bivariate relationships did not include zero in their CIs, hence considered to have sufficient evidence for the polarity of the correlation coefficients. The first of these was collocation frequency and filled pause (*rho* = -.325). The posterior distribution of the correlation coeffects indicated that 69.9% of the parameter estimates exceeded the cut-off criteria for the small effect size, or Pr(|*rho*| ≥ .25), and the probability of posterior draws overlapping with the Region Of Practical Equivalence (ROPE; Kruschke, 2014), or Pr(|*rho*| < .05), was .032, suggesting that over two thirds of estimated parameter values were considered to have at least small effects (Plonsky & Oswald, 2014) and only 3.2% of the parameter values were practically equivalent to the null hypothesis. This result suggests that speakers who answered more frequent collocations in word association produced fewer filled pauses in spontaneous speech. Additionally, although no meaningful relationships between collocation measures and lexical diversity were found (*rho* < |.25|), MI score seemed to negatively correlate with two lexical richness measures, frequency (*rho* = -.339; Pr(|*rho*| ≥ .25) = .740; Pr(|*rho*| < .05) = 0.024) and range (*rho* = -.347; Pr(|*rho*| ≥ .25) = .746; Pr(|*rho*| < .05) = .024) with small to medium effects. These results indicate that speakers who provided responses strongly associated with cue words were more likely to produce sophisticated vocabulary—words that appear rarely and in a narrow range—

in spontaneous oral narrative. No reliable evidence as to the polarity of the correlation between t-score and oral proficiency was found with all *rho* crossing zero in the 95% CIs, although there was a tendency for negative correlation between t-score and filled pauses (*rho* = -.281).

First response. Bayesian correlation analyses on the first response measures showed weak-to-medium correlations between collocation frequency and three fluency measures: fluency rating (*rho* = -.371; Pr(|*rho*| ≥ .25) = .806; Pr(|*rho*| < .05) = .014), articulation rate (*rho* = -.312; Pr(|*rho*| ≥ .25) = 67.3; Pr(|*rho*| < .05) = .037), and silent pause ratio (*rho* = .329; Pr(|*rho*| ≥ .25) = .72; Pr(|*rho*| < .05) = .028). These results indicate that speakers who provided low-frequency collocations in word association were more likely to produce the L2 faster with fewer silent pauses and be perceived fluent. Comparatively, although no meaningful relationships between collocation measures and lexical diversity were found (*rho* < |.25|), MI score was correlated with lexical richness rating (*rho* = .344; Pr(|*rho*| ≥ .25) = .753; Pr(|*rho*| < .05) = .023) and frequency (*rho* = -.360; Pr(|*rho*| ≥ .25) = .780; Pr(|*rho*| < .05) = .018) with medium effects. These results indicate that speakers who provided responses strongly associated with cue words were more likely to be perceived as lexically proficient and produce low-frequency vocabulary in oral narrative. No reliable evidence as to the polarity of the correlations between t-score and oral proficiency was found with all *rho* crossing zero in the 95% CIs.

Table 1. *Descriptive Statistics of Collocation Measures (Based on All Responses and First Responses) and Oral Proficiency Scores*

	M	*SD*	Range	
			Min	Max
Collocation measures				
All responses				
Collocation frequency (raw)	394.23	129.28	200.70	712.80
Collocation frequency (logged)	5.92	0.34	5.30	6.57
t-score	7.49	1.93	4.10	12.30
MI score	1.82	0.33	1.19	2.60
First responses				
Collocation frequency (raw)	466.11	294.68	170.88	1278.52
Collocation frequency (logged)	5.98	0.57	5.14	7.15
t-score	8.63	3.36	1.93	17.04
MI score	2.02	0.48	1.16	3.25
Oral proficiency measures				
Fluency				
Fluency rating	0.52	0.17	0.24	0.83
Articulation rate	109.12	26.62	50.95	176.20
Silent pause ratio	0.60	0.21	0.28	1.16
Filled pause ratio	0.07	0.06	0.00	0.22
Lexis				
Richness	0.42	0.12	0.19	0.64
Diversity	37.09	10.21	19.20	63.09
Frequency	3.16	0.09	2.92	3.36
Range	-0.38	0.05	-0.52	-0.26

Note. Collocation frequency score was log-transformed.

Table 2. *Correlations between Collocation Measures and Oral Proficiency (All Responses)*

	Collocation frequency			t-score			MI score		
	LCI	*rho*	*UCI*	*LCI*	*rho*	*UCI*	*LCI*	*rho*	*UCI*
Fluency									
Fluency rating	-.436	-.129	.185	-.404	-.094	.222	-.231	.085	.400
Articulation rate	-.366	-.056	.265	-.306	.019	.321	-.173	.142	.458
Silent pause ratio	-.156	.140	.470	-.229	.094	.401	-.313	.015	.333
Filled pause ratio	-.632	-.325	-.028	-.574	-.281	.013	-.442	-.121	.180
Lexis									
Richness rating	-.480	-.177	.135	-.364	-.038	.274	-.169	.140	.452
Diversity	-.521	-.219	.083	-.420	-.110	.206	-.255	.069	.383
Frequency	-.324	.007	.312	-.399	-.087	.235	-.614	-.339	-.047
Range	-.351	-.037	.306	-.421	-.097	.234	-.639	-.347	-.050

Note. LCI = 95% lower credible interval; UCI = 95% upper credible interval.

Table 3. *Correlations between Collocation Measures and Oral Proficiency (First Responses)*

	Collocation frequency			t-score			MI score		
	LCI	*rho*	*UCI*	*LCI*	*rho*	*UCI*	*LCI*	*rho*	*UCI*
Fluency									
Fluency rating	-.642	-.371	-.092	-.549	-.257	.046	-.247	.061	.381
Articulation rate	-.591	-.312	-.012	-.490	-.194	.122	-.310	.019	.328
Silent pause ratio	.041	.329	.612	-.108	.203	.502	-.355	-.045	.275
Filled pause ratio	-.477	-.172	.143	-.570	-.293	.019	-.567	-.255	.048
Lexis									
Richness rating	-.551	-.260	.041	-.225	.090	.403	.060	.344	.621
Diversity	-.507	-.207	.098	-.359	-.039	.268	-.160	.153	.455
Frequency	-.231	.085	.403	-.306	.037	.349	-.636	-.360	-.06
Range	-.212	.122	.426	-.243	.085	.413	-.567	-.279	.040

Note. LCI = 95% lower credible interval; UCI = 95% upper credible interval.

Bayesian linear regression

In order to test whether the relationships found above hold after considering learners' vocabulary size, we constructed a set of four competing regression models for each of the L2 speech outcome variables: (a) unconditional model, (b) collocation-only model, (c) vocabulary-size-only model, and (d) collocation-plus-vocabulary-size model. When more than one collocation measures had meaningful correlations with the outcome variable (oral proficiency measures), the measure with the strongest correlation coefficient was kept, e.g., MI first response was chosen over MI total response in predicting the lexical frequency in oral production. For each model we constructed, convergence criteria were met (R-hat < 1.01; Vehtari et al., 2020). Subsequently, we computed Leave-One-Out cross-validation Information Criterion (LOOIC) to evaluate the relative predictive accuracies of the four models. Table 4 presents the results of the Bayesian stacking through

loo_model_weight() (Yao et al., 2018), where the model weights show how much each model contributes to a linear combination of the models which minimizes LOO.

Three patterns emerged concerning the results of fluency measures (the upper half in Table 4). The first result showed that (d) the collocation-plus-vocabulary model received zero weights for all fluency measures, suggesting that the models did not capture the underlying processes. Second, a great amount of weight is awarded to (b) the collocation-only model, particularly when predicting objective measures of fluency (articulation rate, silent pause ratio, and filled pause ratio). These results indicate that, for predicting oral fluency, the collocation-only model captured the data generating process more consistently than the other models, and thus is considered a more plausible model. A slightly different pattern was observed for fluency rating, where a comparable amount of weight was given to the two single predictor models (b) and (c), but not for (d) the collocation-plus-vocabulary-size model. These results may indicate that neither of the two predictors, when considered together in the same model, added unique contributions to fluency rating compared to when each of the predictor was considered separately. This is possibly because the fluency rating was associated with variances shared between collocation knowledge and vocabulary size ($rho_{collocation\&vocabulary}$ = .512, 95%CI = [.262; .737]).

Regarding lexical richness measures, a contrasting picture emerges; that is, the model weights showed that (d) the collocation-plus-vocabulary-size model may capture the underlying relationships for richness rating (.858) and frequency (.324). The result for richness rating, in particular, contrasts with that of fluency rating, in that the collocation-plus-vocabulary-size model received predominantly more weights than the simpler models (Models b and c). This greater weighting indicates that, in contrast to fluency rating, both collocation knowledge and vocabulary size in combination may capture the underlying data generating process more consistently. Yet another pattern was observed in predicting the frequency measure, where model weights were distributed roughly equally to three models (b)-(d), suggesting none of the three models were superior. This similar weighting illustrates a case where model *selection*, solely according to information criteria, may discard potentially critical information. A closer examination of the relative weights might thus support the following interpretation: there would be a fair amount of contribution by vocabulary size as in (c) and (d), but we cannot entirely dismiss the importance of collocation based on the results of (b) and (d). Finally, a distinct pattern was found for the range measure, where (b) the collocation-only model received 1.617 (= .618/.382) times more weights than (c) the vocabulary-size-only model. Although including two predictors may overfit to the data (probably due to the shared variance in predicting range), collocation knowledge may well capture the data generating process slightly better than vocabulary size.

Table 4. *The summary of model weights through Bayesian stacking.*

Outcome variables	Model weights			
	(a) Unconditional model	(b) Collocation-only model	(c) Vocabulary-size-only model	(d) Collocation-plus-vocabulary-size model
Fluency measures				
Fluency rating	0	.575	.425	0
Articulation rate	.185	.711	.104	0
Silent pauses	.079	.921	0	0
Filled pauses	.371	.629	0	0
Lexical richness measures				
Richness rating	.139	.003	0	0.858
Frequency	0	.293	.383	0.324
Range	0	.618	.382	0

Note. Model weights were estimated using loo_model_weight(), which uses Bayesian stacking (Yao et al., 2018) method. Importantly, Bayesian stacking estimates a weighted linear combination of models so that such combined posteriors minimize Leave-One-Out cross-validation Information Criterion (LOOIC), a measure of predictive accuracy in Bayesian statistics (McElreath, 2020). Filled pause and range measures were based on total response data, and the remaining measures were based on first response data.

The results of the model weights indicated that (d) the collocation-and-vocabulary-size model can be important in two lexical measures (richness rating and frequency). To further explore the relative contribution of the two predictors (vocabulary size and collocation) in such combined models, summaries of the two models are shown in Tables 5 and 6. R-hat values were smaller than 1.01, meaning that the models converged. In the richness rating model, the positive effects of MI indicated that the more exclusively associated collocations a learner provided in word association task, the higher lexical richness rating they tended to score. The negative coefficient for vocabulary size indicated that speakers providing lower frequency of lexical items (i.e., larger vocabulary size) tended to produce lexically richer language in spontaneous speech. The credible intervals for these coefficients were fairly wide, showing the uncertainties of the model overall as well as the specific parameters (particular that of vocabulary size). Notably, MI (B = 0.319, CI [0.023, 0.618]) appeared to be more closely related to lexical richness than vocabulary size (B = -0.290, CI [-0.588, 0.007]). In the lexical frequency model (Table 6), the results indicated that (a) a higher MI was associated with lower frequency words in spoken production, and (b) a lower mean frequency for word association responses was associated with a lower frequency for lexis in oral narratives. Unlike the result for lexical richness rating, vocabulary size (B = 0.358, CI [0.067, 0.644]) appeared to be more closely related to the frequency measure than MI (B = -0.289, CI [-0.578, -0.001]). For a complete analysis result, readers are referred to the online supplementary material.

Table 5. *The final regression model on richness rating.*

		95% Credible Interval for *B*		
	B	Lower	Upper	*Rhat*
Intercept	0.001	-0.289	0.295	1.0005
MI (first response)	0.319	0.023	0.618	1.00025
Vocabulary size	-0.290	-0.588	0.007	1.00025
R^2	0.222	0.047	0.397	
ΔLOOIC (vs. unconditional model)	-5.797			

Note. LOOIC stands for Leave-One-Out cross-validation Information Criterion, where smaller LOOIC indicates better predictive accuracy of the model.

Table 6. *The final regression model on frequency.*

		95% Credible Interval for *B*		
	B	Lower	Upper	*Rhat*
Intercept	-0.0005	-0.293	0.286	1.0005
MI (first response)	-0.289	-0.578	-0.001	1.0002
Vocabulary size	0.358	0.067	0.644	1.0002
R^2	0.248	0.0608	0.425	
ΔLOOIC (vs. unconditional model)	-6.281			

Note. LOOIC stands for Leave-One-Out cross-validation Information Criterion, where smaller LOOIC indicates better predictive accuracy of the model.

Sensitivity analysis

The result of sensitivity analyses indicated that the choice of the prior distribution had negligible effects on the slopes of the collocation knowledge measures. To illustrate this process, Figure 1 presents the result on fluency rating (for the entire result of the sensitivity analysis see Online supplementary material). Whether or not we chose other prior distributions, the substantial interpretation was not affected, suggesting the robustness of the current findings.

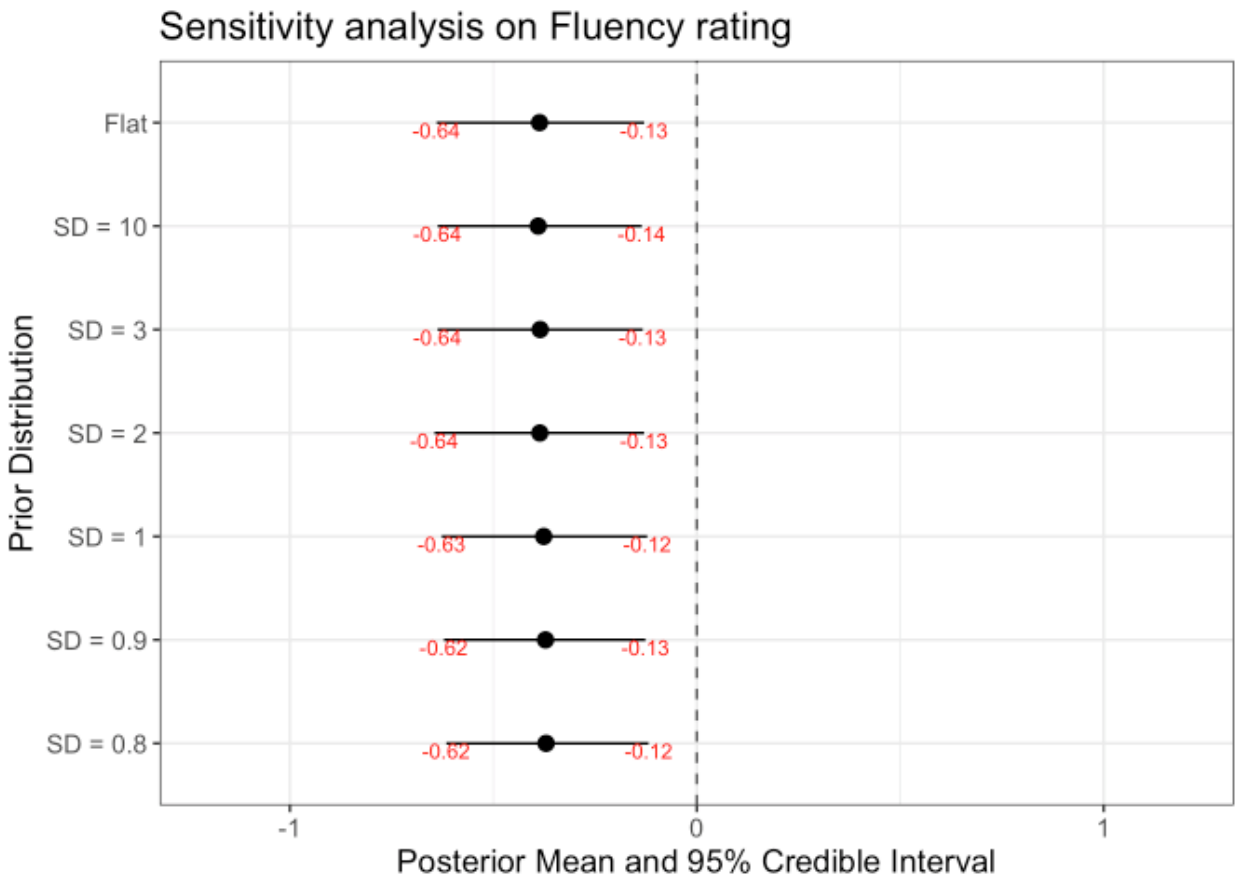

Figure 1. Sensitivity analysis result from the fluency rating model.

Discussion

The current study took an exploratory approach to examining the extent to which collocation knowledge (operationalized as collocational associations elicited via a word association task) is linked with L2 oral proficiency measured as oral fluency (perceived fluency rating, articulation rate, silent pause ratio, filled pause ratio) and lexical richness (perceived richness rating, diversity, frequency, range). Expanding upon an earlier study (Zareva & Wolter, 2012) and drawing from corpus-based techniques (Durrant & Schmitt, 2009), the collocational status of the relation between responses and cue words was determined by means of three corpus-based measures: frequency, t-score, and MI score. To calculate these scores, we conducted two separate analyses, one with all responses and the other with the first responses. Our prediction based on a speech production theory (Levelt, 1989; Kormos, 2006) was that speakers with rich lexicons where collocation is readily accessible for use would be able to orally and spontaneously produce language faster and in a lexically richer manner.

With respect to the relationship between collocation and oral fluency, correlation analysis based on the total and first response data showed that collocation frequency was correlated with fluency: fluency rating (*rho* = -.371), articulation rate (*rho* = -.312), silent pause ratio (*rho* = .329),

and filled pause ratio (*rho* = -.325). Follow-up regression analyses showed that although collocation-plus-vocabulary-size model was not statistically supported, model weight indices suggested a clear pattern supporting the predictive role of collocation frequency for all fluency measures.[2] These results indicated that speakers who provided more low-frequency collocational responses in word association were able to speak more rapidly with fewer silent pauses, and listeners perceived them as optimally fluent. The current findings add to the L2 vocabulary literature demonstrating that knowledge of lower-frequency words is linked with higher oral proficiency (Hilton, 2008; Uchihara & Saito, 2019) as well as support the L2 assessment literature that frequency plays a key role in measuring collocation knowledge (Durrant, 2014; González Fernández & Schmitt, 2015; Nguyen & Webb, 2017). However, the current study was different from earlier studies because we focused on the role of collocation, whereas existing research tends to rely on measures of single-word items alone. The findings therefore provide further insights into our understanding of the speech production system, suggesting that L2 speech production is not only lexically driven (Hilton, 2008; Koizumi & In'nami, 2013; Uchihara & Saito, 2019) but also, to an even greater degree, collocationally driven (Saito, 2020; Tavakoli & Uchihara, 2020). Perhaps learners with a larger repertoire of collocations in the lexicons could (a) benefit from processing advantages (Siyanova-Chanturia & Van Lancker Sidtis, 2018) in accessing and retrieving appropriate lemmas more efficiently at the formulation stage (Kormos, 2006), (b) spare more attentional resources for other aspects of language processing (e.g., building syntactic structure; Skehan, 2009), (c) produce the L2 faster with fewer pauses (De Jong, 2016), and (d) give listeners an impression that speakers are fluent (Uchihara & Saito, 2019).

Regarding the relationship between collocation and lexical richness, the results based on the first and all response data showed small to medium correlations between MI score and three lexical richness measures (but not with lexical diversity): for all responses, frequency (*rho* = -.339) and range (*rho* = -.347); for first responses, richness rating (*rho* = .344) and frequency (*rho* = -.360). Subsequent regression analyses supported collocation-plus-vocabulary-size model, further indicating that MI score was more predictive of lexical richness rating than vocabulary size score (as the 95% CIs for vocabulary size included zero; [-0.588, 0.007]). This result indicated that collocation knowledge (strength of association) had more of an impact on perceived lexical richness than knowledge of single-word items. For another oral proficiency measure indexed by lexical frequency, both MI and vocabulary size measures remained as important predictors in the regression model with the latter ($B_{vocabulary\ size}$ = 0.358) considered a slightly better predictor than the former (B_{MI} = -0.289). The results of the slightly greater weights for vocabulary-size-only model as well as the slight

[2] One of the reviewers highlighted how the issue with the raw frequency measure (vs. t-score, MI) influenced the validity of this result. The issue surfaces particularly when researchers interpret high frequency scores to define collocation. In the present study, lower frequency scores, not higher frequency scores, were associated with fluency measures. Also, a potential confounding factor (i.e., frequency of single words) was considered in the analysis. Thus, the inherent issue with the collocation frequency measure was less problematic in the current study.

superiority of vocabulary size over MI index in the regression model are probably due to the congruency in the calculation of the scores between the predictor and outcome variables (i.e., vocabulary size and lexical frequency) as both of the measures were based on corpus-based frequency of single-word items. To summarize the results, speakers who demonstrated more target-like and sophisticated collocations in word association task (indexed by higher MI scores) produced more low-frequency vocabulary in speech, and their resulting production was perceived to be lexically richer and advanced by trained raters. Notably, collocation knowledge remained as a key predictor of oral proficiency after the effect of vocabulary size was controlled for. This finding is in line with our prediction that richer lexicons enable speakers to produce linguistically more sophisticated language (Skehan, 2009) and supports findings in corpus linguistics and psycholinguistics that advanced language users demonstrate sensitivity to and use of collocations with higher MI scores (Durrant & Schmitt, 2009; Eguchi & Kyle, 2020; Ellis et al., 2008; Granger & Bestgen, 2014; Saito, 2020; Sonbul, 2015). Although our interpretation of these findings seems reasonable and as generally expected, a question emerges, important for further exploration and discussion: Why was MI score exclusively related to lexical richness measures, not fluency measures, despite the presence of relationships between collocation frequency and fluency measures, but not with lexical richness? The answer to this question would advance our understandings of the nature of phraseological competence (Granger & Bestgen, 2014) and its interface with spontaneous speech production (Kormos, 2006).

A possible reason for this pattern may be because the two kinds of the collocation measures tap into differential aspects of phrasal competence and reflect different stages of L2 oral development. More specifically, compared with raw frequency score, MI score indicating a higher level of phraseological sophistication may be of little relevance to an initial stage of L2 oral development including temporal features (e.g., fluency); instead, it may be more closely aligned with a linguistic aspect which develops at a relatively later stage including lexical richness (Saito, Trofimovich, et al., 2016; Tavakoli, 2018). From a theoretical perspective, the link between raw frequency score and fluency could also be explained in terms of power law of L2 practice (DeKeyser, 2001; Ellis, 2002). Learners who can demonstrate knowledge of larger low-frequency collocations are likely to have received a greater amount of exposure to high-frequency collocations, as a result of which such high-frequency collocations become even more entrenched in the lexicon, and thus, highly activated and immediately retrievable for L2 use. This view of lexical development reflects the theory of power law of learning (see Ellis, 2002), which also aligns well with the development of oral fluency. L2 fluency research has suggested that learners can improve oral fluency relatively quickly after completing an intensive repetition practice (e.g., 3-day practice sessions 30 minutes each; Suzuki, 2020). Such a rapid change in fluency system is often explained by power law of practice under the skill acquisition theory (DeKeyser, 2001). Perhaps the observed association between knowledge of infrequent phrases and fluency could therefore be attributed to the proceduralization of more frequent phraseological expressions, which in turn served as the basis for

fluent oral production (Tavakoli & Uchihara, 2020). In contrast, lexical richness improvement involves more than proceduralization of available lexical resources through a short-term fluency practice (Suzuki, 2020) or immersion (Tavakoli, 2018). It necessitates not only vocabulary expansion by learning lower-frequency words (breadth of knowledge) but also attending to various aspects of lexical knowledge other than L2 word forms and meanings including contextual constraints in using the word and its grammatical functions (depth of word knowledge).

Contrary to the results of collocation frequency and MI measures, no meaningful correlations between t-score and oral proficiency measures were observed with all mean coefficients smaller than |.30| and all 95% CIs including zero. These findings were not consistent with findings from Zareva and Wolter (2012) that the number of collocational responses identified using t-score was descriptively greater for intermediate learners than for advanced learners. Several methodological differences such as outcome measures (oral proficiency vs. TOEFL scores) and scoring procedures of t-score (mean vs. cut-off scores) makes simple comparison between the current study and Zareva and Wolter's difficult. This study instead sheds light on different aspects of collocation knowledge that t-score and MI highlight (Durrant & Schmitt, 2009; Granger & Bestgen , 2014), although both indices are subsumed under the set of association measures (Gablasove et al., 2017). Before drawing any conclusion regarding the value of t-score as a collocation measure, future investigations in the field of word association research are needed to further explore whether it relates to L2 proficiency.

Finally, our findings showed that the first response elicited in a word association task made more salient the relationship between collocation measures and oral proficiency compared to the total response data. This clear pattern suggests that the very first responses that came to learners' minds might reflect their quality or richness of mental lexicons more accurately (Playfoot et al., 2016; see also Eguchi et al., in press); hence first response data might be more sensitive to and reflective of L2 proficiency than total response data. A possible reason for such insensitivity of total response data might be attributed to chaining effects—i.e. the first response acting as a prompt for the next response (Fitzpatrick, 2006). To avoid this issue, future studies should include more stimulus words to elicit primary responses and use a method that repeats the presentation of the stimulus word after every blank (Wolter, 2002).

Conclusion

The current study set out to examine the relationship between collocation knowledge elicited via a word association task and oral proficiency (in terms of temporal and lexical features), and found that two corpus-based collocation measures (frequency and MI) were related to oral fluency and lexical richness. Our findings support the view that speakers with rich collocational networks can produce language in a fluent and lexically rich manner during spontaneous speech.

This study provides an important implication for word association research and oral proficiency development. Previous research relies on intuitive judgements to determine whether the

relation between association responses and stimuli is collocational (Fitzpatrick, 2013). However, the current study demonstrated that corpus measures such as frequency and association measures (t-score and MI) might be useful alternative tools for the purpose of assessing and identifying collocations elicited through a word association task. This field will benefit from inclusion of such objective measures with the view to advancing the understanding of the relationship between collocation knowledge and L2 proficiency.

Finally, future research should replicate the findings of the current study in order to determine the role of collocation in oral proficiency. The small sample size (N = 40) prevents this study from making a stronger claim regarding some of our findings. Especially for total response data, Bayesian analyses indicated a great amount of uncertainty in the estimated parameters (see 95% CIs; Table 2 and 3). It should also be noted that there was variation in the amount of evidence (or plausibility) to support the results even among the cases where 95% CIs did not include zero. For example, a greater amount of evidence was available to interpret some results meaningfully, such as 80.6% probability to find the correlation coefficient between (first response) collocation frequency and fluency rating above a small effect-size value determined in SLA research (Plonsky & Oswald, 2014), in comparison to the results of the relationship between (first response) collocation frequency and articulation rate (67.3% probability to find the correlation above the set benchmark value). Despite the small sample size, however, the results of Bayesian analytic procedure allowed us to draw a tentative conclusion regarding the meaningful roles of differential collocation knowledge on oral proficiency constructs. Since the Bayesian analysis provided the full information about the posterior distribution, a reader can examine the validity of our argument by referring to the online supplementary material, which provides full information about the analytic process. In the meantime, a simple recommendation for future studies is to increase sample size so that the amount of uncertainty would be reduced. An alternative suggestion—given that speech measurement procedures are highly cost-intensive (e.g., eliciting speech samples individually, training raters, implementing rating sessions, identifying temporal properties such as silent and filled pauses)—is to use this study results as prior to inform the subsequent Bayesian analysis. The use of informative priors can advance our understanding of the current issue by updating the effect size using the new data. This approach reflects a recommended practice in scientific research, highlighting cumulative, collaborative aspect of scientific endeavor (Kruschke, 2010).

References

Baez-Ortega, A. (2018). Bayesian robust correlation with Stan in R (and why you should use Bayesian methods). In *Silico Naïve Thoughts on Data*. https://baezortega.github.io/2018/05/28/robust-correlation/

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, *21*(2), 139–164.

Barfield, A. (2009). Exploring productive L2 collocation knowledge. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners: Papers and perspectives in honour of Paul Meara* (pp. 95-110). Bristol, UK: Multilingual Matters.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.

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

Boers, F., & Webb, S. (2018). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, *51*(1), 77-89.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *202*(2015), 139–173.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01

Clenton, J. (2015). Testing the revised hierarchical model: evidence from word associations. *Bilingualism: Language and Cognition*, *18*(1), 118-125.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*(1), 182-193.

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production? *Studies in Second Language Acquisition*, *35*(4), 727-755.

Daller, M. H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197-222.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 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*(2), 159-190.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 121-142). Amsterdam: John Benjamins.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (1st ed., pp. 125–151). Cambridge University Press.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533-557.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157-177.

Eguchi, M., & Kyle, K., (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381-400.

Eguchi, M., Suzuki, S., & Suzuki, Y. (In press). Lexical competence underlying second language word association tasks: Examining the construct validity of response type and response time measures. *Studies in Second Language Acquisition*.

Ellis, N. C. (2002). Frequency effects in language processing: A Review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188.

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.

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations.* Ph.D. thesis. Stuttgart: University of Stuttgart.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook*, *6*(1), 121-145.

Fitzpatrick, T. (2012). Tracking the changes: vocabulary acquisition in the study abroad context. *The Language Learning Journal*, *40*(1), 81-98.

Fitzpatrick, T. (2013). Word associations. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-6). Oxford, UK: Wiley Blackwell.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*, 537-554.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(1), 155-179.

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

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? *International Journal of Applied Linguistics*, *166*(1), 94-126.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229-252.

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

Gyllstad, H., & Schmitt, N. (2018). Testing formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.) *Understanding formulaic language: A second language acquisition perspective* (pp. 174-191). London, New York: Routledge.

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language: Processes and development* (pp. 22-66). London, UK: Palgrave Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development—a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29-56). Amsterdam: John Benjamins.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal*, *36*(2), 153-166.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London, UK: Routledge.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159.

Jiang, N., & Zhang, J. (2019). Form prominence in the L2 lexicon: Further evidence from word association. *Second Language Research*.

Kim, M., Crossley, S., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120-141.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153-165). Edinburgh: Edinburgh University Press.

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research, 4*(5), 900-913.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293–300.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, JAGS, and stan*. Academic Press.

Kruse, H., Pankhurst, J., & Sharwood Smith, M. (1987). A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9*(2), 141-154.

Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454–476). London: Routledge.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 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*(3), 1030-1046.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Li, H., & Lorenzo-Dus, N. (2014). Investigating how vocabulary is assessed in a narrative task through raters' verbal protocols. *System, 46*, 1-13.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor and Francis, CRC Press.

McNamara, D. S., Grasesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Competence and performance in language learning* (pp. 35-53). Cambridge, UK: Cambridge University Press.

Meara, P., & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19-30.

Meara, P., & Jones, G. (1990). *The Eurocentres Vocabulary Size Tests 10Ka*. Zurich: Eurocentres Learning Service.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographical vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *The Canadian Modern Language Review*, *63*(1), 127–147.

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1-24.

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*, 451–468.

Nation, I. S. P. (1984). *Vocabulary lists*. Wellington, New Zealand: Victoria University of Wellington, English Language Institute.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298-320.

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383-410.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach: Bayesian revolution in L2 research. *Language Learning*, *68*(4), 1032–1075.

Nissen, H. B., & Henriksen, B. (2006). Word class influence on word association test results. *International Journal of Applied Linguistics*, *16*(3), 389-408.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121-145.

Playfoot, D., Balint, T., Pandya, V., Parkes, A., Peters, M., & Richards, S. (2016). Are word association responses really the first words that come to mind? *Applied Linguistics*, *39*(5), 607-624.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548-588.

Saito, K., Macmillan, K., Mai, T., Suzukida, Y., Sun, H., Magne, V., Ilkan, M., & Murakami, A. (2020). Developing, Analyzing and Sharing Multivariate Datasets: Individual Differences in L2 Learning Revisited. *Annual Review of Applied Linguistics*, *40*, 9–25.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652-708.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*, 439-462.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, *38*(4), 677-701.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913-951.

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.

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

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549-569.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Söderman, T. (1993). Word associations of foreign language learners and native speakers: The phenomenon of a shift in response type and its relevance for lexical development. In H. Ringbom (Ed.), *Near-native proficiency in English* (pp. 91-182). Abo, Finland: Abo Akademi.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition*, *18*(3), 419-437.

Suzuki, Y. (2020). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*.

Tavakoli, P. (2018). L2 development in an intensive study abroad EAP context. *System*, *72*, 62-74.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: John Benjamins.

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

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569-613.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916.

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*(3), 564–587.

Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*(1), 64-75.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65-83.

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly*, *9*(2), 172-185.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford, UK: Oxford University Press.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition*, *23*, 41-69.

Wolter, B. (2002). Assessing proficiency through word associations: Is there still hope? *System*, *30*(3), 315-329.

Yang, Y., Sun, Y., Chang, P., & Li, Y. (2019). Exploring the relationship between language aptitude, vocabulary size, and EFL graduate students' L2 writing performance. *TESOL Quarterly*.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). Bayesian Analysis, 13(3), 917–1007. https://doi.org/10.1214/17-BA1091

Zareva, A., & Wolter, B. (2012). The 'promise' of three methods of word association analysis to L2 lexical research. *Second Language Research*, *28*(1), 41-67.

Supporting Information-A: 30 stimulus words and the Lex30 task format

1. attack				
2. board				
3. close				
4. cloth				
5. dig				
6. dirty				
7. disease				
8. experience				
9. fruit				
10. furniture				
11. habit				
12. hold				
13. hope				
14. kick				
15. map				
16. obey				
17. pot				
18. potato				
19. real				
20. rest				
21. rice				
22. science				
23. seat				
24. spell				
25. substance				
26. stupid				
27. television				
28. tooth				
29. trade				
30. window				