To be published in *Language Learning* (Wiley Blackwell)

LANGUAGE LEARNING
A Journal of Research in Language Studies

Multi- or Single-Word Units? The Role of Collocation Use in Comprehensible and Contextually Appropriate Second Language Speech

Kazuya Saito[1]

Abstract

The current study examined the degree to which collocation use (i.e., meaningful co-occurrences of multiple words) was related to L1 raters' intuitive judgements of L2 speech. Speech samples from a picture description task performed by eighty-five Japanese learners of English with varied L2 proficiency profiles were transcribed and assessed by 10 L1 raters for global comprehensibility (how easily speech can be understood) and lexical appropriateness (the extent to which words are used adequately and naturally in context). The samples were then submitted to a range of lexical measures tapping into the collocation (frequency, association), depth (abstractness) and breadth (frequency, range) aspects of L2 vocabulary use. Results of the statistical analyses showed that the raters' comprehensibility and lexical appropriateness scores were strongly determined by the L2 speakers' use of low-frequency combinations containing infrequent, abstract and complex words (i.e., mutual information).

Key words: Second language speech, collocation, oral proficiency, comprehensibility, vocabulary use

[1] I am grateful to Masaki Eguchi, Takumi Uchihara, George Smith, Parvaneh Tavakoli, and four anonymous *Language Learning* reviewers for their helpful input, feedback and advice at every stage of the manuscript writing and revising processes. I also thank Kokoro Muramoto for his professional assistance for data collection and analysis.

Foreign accent is a normal characteristic of L2 learning in adulthood. This realization has led to a wide-ranging scholarly consensus that the linguistic quality of L2 speech should be evaluated based on intelligibility, comprehensibility and communicative adequacy rather than L1-like accuracy. Given that not all linguistic errors equally impact successful comprehension and communication, an increasing number of studies have begun to advocate for combining subjective judgements of L2 oral proficiency with the notion of error gravity (Derwing & Munro, 2015 for comprehensibility; Foster & Wigglesworth, 2016 for weighted accuracy; Saito, Trofimovich, & Isaacs, 2017 for lexical appropriateness). Much of the attention especially in the area of L2 comprehensibility has been given to examining the phonological aspects most related to raters' behaviours during global L2 speech evaluation (e.g., Kang, Rubin, & Pickering, 2010). By comparison, the lexical profiles of comprehensible and contextually-appropriate L2 speech are largely under-researched and as a consequence poorly understood. In the field of L2 vocabulary learning, recent evidence has indicated that collocational information (i.e., meaningful co-occurrence of multiple words) may be a relatively strong determiner of L2 speaking proficiency (e.g., Kyle & Crossley, 2015).

Interfacing perspectives in L2 speech and vocabulary, the current study explored the extent to which collocation association factors, operationalized via two different n-gram association measures (t-scores, Mutual Information), could predict the comprehensibility and lexical appropriateness judgements of Japanese learners' L2 speech. Following methodological discussion and innovation in the precursor research (e.g., Saito, Webb, Trofimovich, & Isaacs, 2016a), raters assessed for the lexical quality of L2 speech by reading transcripts rather than listening to actual audio samples. The unique methodology here allowed us to further refine our understanding of the lexical correlates of L2 comprehensibility and appropriateness controlling for the influence of phonological accentedness.

Background

Collocations in L2 Vocabulary Research

According to usage-based accounts of SLA, language is formulaic in nature, and most linguistic information is stored in the form of multi-word units, or "chunks". Through sufficient exposure in meaningful contexts, L2 leaners are thought to develop the ability to accurately, rapidly, and subconsciously access these chunks in response to specific contextual and linguistic cues (Ellis, 2012). Furthermore, it is believed that such formulaic bundles can spur morphosyntactic development through abstract, schematic analysis of the chunks' constituent parts (Tomasello, 2003).

Corpus-based investigations have shown that a large proportion of oral discourse between and among L1 speakers (30-40%) is formulaic in nature (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Erman & Warren, 2000; Leech, 2000). Second language users, on the other hand, have been observed to produce spontaneous speech that is comparatively lacking in the use of formulaic language, and often contains recurrent dysfluencies such as filled pauses (Foster, 2001). From this point of view, the attainment of collocational knowledge can be seen as instrumental to becoming a functional, communicatively successful L2 user (Siyanova-Chanturia & Martinez, 2015; Wray, 2000).

A precise definition of a collocation has been elusive in the L2 literature (Boers & Webb, 2018 for a concise review on the phraseology vs. corpus-based approaches to defining collocations). Following Baker, Hardie, and McEnery's (2006) definition in corpus linguistics, this study defines collocation as "the phenomenon surrounding the fact that certain words are more likely to occur in combination with other words in certain contexts" (p. 36). This is inclusive in that it covers a range of multiword units with different degrees of transparency (e.g., "kick a ball" vs. "kick him out" vs. "kick the bucket").

One corpus-based operationalization of collocations is n-gram frequency, i.e., how often continuous strings of *n* words ($n \geq 2$) are used in a specific corpus (Biber, Conrad, & Cortes, 2004). However, the interpretation of raw n-gram frequency is complex, because counting the number of occurrence does not necessarily capture the actual use of multiple words in context. Raw n-gram frequency scores may include not only formulaic sequences (e.g., *think* and *of*), but also random co-occurrences of lexical items (e.g., *think* and *chair*).[2] It is crucial to take into account not only how often certain word combinations have appeared, but also the extent to which the constituent words associate with each other.

Two measures have been devised to index the strength of meaningful (rather than by-chance) n-gram associations (for a comprehensive overview, see Gablasova, Brezina, & McEnery, 2017). First, t-scores mathematically represent the above-chance co-occurrence of *n* words without taking into account the frequency of individual words (for the details of the calculation procedure, see the Method section). Notably, t-scores favor the combination of high-frequency words, because they occur together in a corpus more often than low-frequency words. In essence, multiword units with higher t-scores comprise more frequent words, which are relatively concrete, meaningful and transparent in nature.

Importantly, the concept of t-scores does not necessarily reflect the strength of word partnerships. While high frequency words form a frequently-occurring collocation, they can also collocate with many other different words. In this sense, the strength of the partnership could be considered relatively weak. Mutual information (MI) scores, on the other hand, represent the relative exclusivity of word combinations, i.e., the statistical likelihood of *n* words occurring together with each other but not with any other word. Consequently, the calculation of MI favors

[2] Random co-occurrences are calculated by dividing the number of any possible combinations (multiplying the frequency of node by collocate) by the total number of tokens in a reference corpus. On a related note, the inherent problem in the adjusted n-gram, t- and MI-scores, is that it does not well capture the extent to which multiword units can be structurally and semantically complete (formulaic sequences) or incomplete (lexical bundles) (Jeong & Jiang, 2019). For example, t-scores can identify and eliminate "think chair" as an instance of random co-occurrence; but cannot distinguish the difference between "think of" as a formulaic sequence (verb phrase) and "consider the" (verb phrase + unfinished noun phrase) as a lexical bundle.

lower-frequency words, because the number of partner words that they can collocate with is exclusively limited, resulting in relatively strong partnerships. In essence, n-grams with higher MI scores likely comprise low-frequent, abstract and complex words.[3]

Previous research in L2 writing has revealed how different measures of association strength operate differently with beginner and advanced writing samples. It has been shown that lower-level production tends to feature high-frequency collocations (and consequently higher t-scores), while higher-level writers employ low-frequency but strongly associated collocations (with higher MI scores) (Bestgen, 2017; Durrant & Schmitt, 2009; Kim, Crossley, & Kyle, 2018). Several studies have also demonstrated that certain collocation measures (MI-scores) are weakly but significantly correlated with global L2 writing quality (e.g., Kyle & Crossley, 2016 for r = .10-.20 in TOEFL Writing; Garner, Crossley, & Kyle, 2018 for r = .20-.30 in CEFR Writing).

When it comes to the lexical correlates of L2 oral proficiency, however, most of the existing studies have focused on the analyses of single-word units. The results of this body of literature indicate that most index categories (e.g., frequency) can explain only a small portion of the variance (<10%) in raters' L2 speaking judgements, e.g., based on ACTFL (Crossley, Salsbury, McNamara, & Jarvis, 2011 for r = -.29) and TOEFL proficiency guidelines (e.g., Crossley & McNamara, 2013 for r = -.23). Investigations into the role of collocation use in speaking proficiency is comparatively limited. Kyle and Crossley (2015) examined the predictive power of both single-word and collocation measures for L2 oral proficiency. Among the many indices predicting variance in speaking scores, multiword units (trigram frequency) emerged as the strongest predictor in the model (r = .59), accounting for 35% of the shared variance in holistic TOEFL speaking proficiency scores. Similarly, Eguchi and Kyle (forthcoming) found relatively strong roles of collocation (r = .49) in the context of ACTFL Oral Proficiency Interview responses (for the results using rater judgements of formulaic sequences, see also Boers, Eyckmans, Kappel, Stengers, & Demecheleer. 2006; Stengers, Boers, Housen & Eyckmans, 2011).

[3] The parallel relationship between frequency and abstractness and its impact on acquisition is an ongoing empirical question that needs a careful examination with a thorough, longitudinal research design. There is some evidence that both frequency and a range of abstractness measures (e.g., hypernymy, meaningfulness, concreteness, imageability) can be predictive of the L2 oral proficiency development (e.g., Crossley et al., 2018). The findings suggest that frequency and abstractness are somewhat inter-connected as both of them are relevant to the way how L2 learners expand their spoken vocabulary repertoire. Interestingly, however, our dataset (based on N = 85 Japanese learners with diverse L2 English proficiency levels) demonstrated a very unique relationship between MI scores, frequency and abstractness. According to the results of factor analyses (see Table 6), MI scores and abstractness tapped into the same construct of L2 oral proficiency (higher MI scores entail lower abstract scores); but MI scores and frequency could be grouped into two different constructs of L2 oral proficiency. It would be intriguing to examine the same topic but by using a large-scale L2 speech corpus in particular (e.g., Ishikawa, 2014 for the International Corpus Network of Asian Learners of English: Spoken Monologue and Dialogue).

From the brief review of the literature presented here, it seems that L2 learners may be judged as more proficient if they employ more common combinations of words in their writing and speech (Durrant & Schmitt, 2009; Kyle & Crossley, 2015). Whereas L2 learners' formulaic sequences initially comprise more high-frequency collocations, they are likely to increasingly employ more low-frequency collocations in the later stages of L2 development (Garner & Crossley, 2018; Kim et al., 2018). Notably, it appears that the size of the collocation-proficiency link is relatively large in speaking (r = .59 in Kyle & Crossley) compared to writing (r = .10-.30), a difference which can be ascribed to the fact that speech production includes more formulaic and idiomatic expressions than written production (Biber et al., 1999).

Though revealing, the findings in support of the relationship between collocation and L2 speaking (and writing) proficiency need to be interpreted with caution. In such studies, raters first engage in extensive training so that they can provide holistic proficiency scores in accordance with certain descriptors and guidelines (e.g., 0-4 points in TOEFL). As aptly pointed out by Koizumi (2012), such trained raters may pay attention to the lexical features of a text (e.g., collocation), simply because they are asked to do so. It is thus valid to wonder which lexical factors L1 speakers rely on when assessing the global quality of L2 speech in the absence of detailed descriptors.

Examining the generalizability of Crossley and Kyle et al.'s findings on TOEFL Speaking Tasks to more intuitive and subjective judgements of L2 oral proficiency is crucial for expanding our knowledge of the collocation-proficiency link. Many scholars have stressed that it is such human perception that ultimately matters in a real-life communication between L1 and L2 users (e.g., Derwing & Munro, 2015). Given the theoretical, pedagogical and practical importance of the topic (the role of collocation in perceived L2 oral proficiency), the current study was designed to explore the extent to which collocation information relates to L1 judges' *intuitive* and *subjective* judgements of comprehensible and lexically appropriate speech.

Intuitive Judgements of L2 Comprehensibility

Many students and teachers in foreign language contexts tend to perceive the attainment of L1-like speech as an ideal goal of language learning (e.g., Tokumoto & Shibata, 2011). The attainability of this goal is somewhat questionable, however, as there is ample evidence that post-pubertal L2 learners' speech is generally marked by a foreign accent—a consequence of these learners' strongly-developed L1 phonological systems (Flege, 2016). In addition, it has been observed that the majority of the world's English-speaking population is made up of L2 speakers, and that a large proportion of interaction in English occurs in a lingua franca context (Pennycook, 2017). All of these factors combined have led to a major paradigm shift in applied linguistics which specifies that L2 proficiency should be examined against L2 users themselves rather L1-speaker norms (e.g., Levis, 2005 for Intelligibility Principle; Ortega, 2013 for Multilingual Turn). Accordingly, researchers have developed the consensus in L2 pronunciation research that it is more important to obtain comprehensible speech rather than L1-like accuracy (Crowther, Trofimovich, Saito, & Issacs, 2017; Derwing & Munro, 2015).

Comprehensibility has traditionally been operationalized as human raters' *intuitive* judgements of how easily a talker can be understood. Procedurally, studies on L2 comprehensibility tend to give raters a brief overview of the definition of the construct (i.e., ease of understanding), and then expose them to L2 speech samples presented in a randomized order. These raters are then generally asked to rate their overall impression of each sample based on a 9-point scale (*1 = difficult to understand, 9 = easy to understand*). This "intuition-based" approach is essentially different from judgements conducted by trained raters in high-stakes testing settings, wherein specific constructs of L2 speech (e.g., pronunciation, fluency, lexicogrammar) are evaluated in accordance with prescribed, task-specific descriptors or rubrics (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008). It is intriguing to note that raters in intuition-based studies demonstrate relatively high interrater agreement despite receiving minimal explanation of a construct's definition, and not being able to reference a rubric when rating. This in turn suggests that naïve raters have a shared, intuitive notion of what constitutes comprehensible L2 speech.

To date, a number of empirical studies have extensively investigated how phonological information affects L2 comprehensibility judgements. For example, raters have been shown to pay selective attention to features such as segmental contrasts with high functional load (e.g., /ɹ/ vs. /l/ but not /s/ vs. /θ/) (Munro & Derwing, 2006; Suzukida & Saito, 2019), prosodic accuracy (e.g., Isaacs & Trofimovich, 2012; Kang et al., 2010), and temporal fluency (Suzuki & Kormos, 2019). The amount of phonological influence on these judgements also varies in accordance with non-linguistic factors, such as task demands (e.g., Crowther et al., 2017) and listeners' familiarity with foreign-accented speech (e.g., Ludwig & Mora, 2017). During their understanding of L2 speech, however, raters do take into account a wide range of linguistic information beyond pronunciation.

More recently, a growing number of scholars have examined how L2 comprehensibility could be influenced by other linguistic features such as lexicogrammatical appropriateness and sophistication (e.g., Crowther et al., 2017; Isaacs, Trofimovich, & Foote, 2018; Saito et al., 2017). Because raters make comprehensibility judgements by listening to speech samples, phonological factors not surprisingly explain a great deal of variance in L2 comprehensibility judgements (50-60%). This means that even when L2 learners can handle vocabulary adequately, they could be perceived as difficult to understand when they make phonological errors. To provide a more detailed and refined picture of the relationship between vocabulary use and L2 speech comprehension, it is thus crucial to develop, adopt and elaborate a methodology by which to separate raters' processing of vocabulary from that of phonological information during L2 comprehensibility judgements.

To correspond to this concern, Saito, Webb, Trofimovich, and Isaacs (2016a) explored the lexical profiles of comprehensible L2 speech with 40 L1 French learners of English. Diverging from the traditional intuitive approach, wherein raters listen to and assess *audio* recordings, all speech samples were presented to raters as *transcribed* texts to control for the influence of phonological factors on their judgements. The raters' lexical processing during L2 comprehensibility judgements was analyzed by comparing their comprehensibility ratings and the vocabulary use of the speech samples. Among the wide range of single-word measures, the extent to which L2 speakers used different types of more abstract words (diversity, abstractness) explained small amounts of variance (5-10%) in L2 comprehensibility; however, the use of more infrequent words (frequency) did not have any significant associations with raters' L2 comprehensibility judgements ($p > .05$). While this study was an important step in examining the lexical correlates of comprehensible L2 speech, the study was limited to single-word measures.

In this particular study, Saito et al. (2016a) pursued the same *concept* of L2 comprehensibility as in the previous literature—ease of understanding (e.g., Derwing & Munro, 2015). To isolate the vocabulary influence on the phenomenon of comprehensibility, however, Saito et al. adopted the different *methodology* of measuring comprehensibility (reading transcripts rather than listening to audio files). It is noteworthy that the methodology proposed here (the transcript assessment approach) has been widely used in various L2 acquisition literature (but outside L2 comprehensibility studies), whereby researchers aim to look at how raters assess the vocabulary aspects of L2 speech regardless of the degree of phonological accentedness (e.g., Crossley et al., 2011). Below, I also provide a literature review on another line of L2 speech assessment research which has adopted the analyses of transcribed speech as a main methodological option.

Subjective Judgements of Global Appropriateness

In the field of applied linguistics, few would disagree with the fundamental idea that lexical and morphosyntactic appropriateness are key components of L2 oral proficiency. Furthermore, it well-known that L2 learners' language becomes more accurate as a function of increased practice, experience and exposure to the target language (Tavakoli, 2018). Historically, the global accuracy of L2 speech has been analyzed by tallying the number of times speakers make specific lexicogrammar errors in obligatory contexts (e.g., every clause or 100 words; for a review, see Housen, Kuiken, & Vedder, 2012). However, the binary coding of local accuracy (correct vs. incorrect) has been questioned because it may not capture the impact each error has on communicative adequacy (Révész, Ekiert, & Torgersen, 2016). A growing number of scholars have thus far explored how the global quality of L2 speech can be assessed intuitively by humans while taking into account error gravity, using indices such as weighted clause ratio (Foster & Wigglesworth, 2016), semantic and lexical appropriateness (Saito et al., 2017), and morphosyntactic accuracy (Ruivivar & Collins, 2017). In these studies, raters are typically asked to pay attention to specific aspects of language (e.g., accuracy rather than fluency and complexity) in their ratings, but their judgements can still be considered to be largely *subjective*, since there is no reference to descriptors or rubrics of any kind. In other words, it is the raters that decide which errors should be considered as more or less important during their evaluations of global L2 accuracy.

Similar to L2 comprehensibility judgements, subjective analyses of L2 accuracy have been found to lead to high levels of inter-rater agreement, indicating that raters (especially those with linguistics and L2 teaching experience) can reliably assess the linguistic appropriateness of L2 speech (Saito et al., 2017). There is some empirical evidence that L2 learners' improvement patterns can be clearly observed when the global accuracy of their speech is evaluated by human raters, but not when the same dimension is analyzed simply by counting the number of linguistic errors made per counting unit (Foster & Wigglesworth, 2016). These findings provide indirect evidence that this approach to accuracy analysis may serve as a better index of L2 development than the dichotomous coding of linguistic errors. To our knowledge, however, little is known about what lexical characteristics raters actually use to determine the different levels of perceived L2 oral proficiency.

Motivation for Current Study

Collocation has been recognized as a key construct of L2 acquisition (Ellis, 2012) and has been extensively researched in L2 writing research (e.g., Garner et al., 2018). In the context of the TOEFL, which employs judgement scores based on detailed descriptors, Kyle and Crossley's (2015) work has recently indicated that collocation may be a primary lexical determinant of L2 learners' speaking scores. Contrary to the trained-descriptor approach, many scholars have begun to acknowledge the value of human raters' intuitive judgements as an ecologically valid way to assess the global comprehensibility and accuracy of L2 speech production. To date, however, the relationship between collocational information and perceived L2 oral proficiency is relatively unknown and ripe for further investigation. The main objective of the current study was thus to examine the extent to which two different types of collocation association measures—t-scores, MI-scores—could predict variance in raters' subjective judgements of comprehensible and lexically appropriate speech. In order to further examine the relative weights of the collocation factors, the study focuses on how two groups of raters' judgements relate to the use of collocation vs. other major dimensions of vocabulary, i.e., breadth and depth (for details, see below).

A set of predictions are formulated in regard to the relationship between different types of collocation measures (t-scores, mutual information) and perceived L2 oral proficiency (comprehensibility, lexical appropriateness). Given that L2 learning is characterized by learners' increasing control over not only frequent, but also infrequent multi-word sequences (Ellis, 2012), we predict that the collocation indices would be relatively strong predictors of L2 comprehensibility and accuracy (more so than the single-word breadth and depth indices) (Kyle & Crossley, 2015). Since L1 speakers tend to be more sensitive to lower frequency combinations of words when judging the targetlikeness of multiword strings (Ellis, Simpson-Vlach, & Maynard, 2008), it is likely that mutual information, rather than t-scores, would be a better predictor of L2 comprehensibility and lexical appropriateness judgements (cf. Garner et al., 2018 for L2 writing proficiency). Finally, the predictive power of the collocation indices may be slightly stronger in lexical appropriateness than in comprehensibility, because the former judgements concern lexical qualities of L2 speech more directly than the latter judgements do.

Method

Participants

Speakers (*n* = 85). Eighty-five Japanese learners of English provided the speech samples used for rating in the current study. Thirty-eight were university students in Japan with relatively homogeneous L2 English learning experiences—i.e., six to seven years of English-as-a-Foreign-Language education in Japan without any experience studying abroad. The participants varied greatly in their proficiency levels (TOEIC *M* = 520 out of 950, *Range* = 460-910), ranging from A2 "Basic Users" to B2 "Independent Users" according to the CEFR.

The remaining 47 participants were recruited from various cities in the USA. Due to the difficulty in accessing the target population (long-term Japanese residents) in a single city, the decision was made to distribute digital flyers on a number of community websites throughout the country. All interested participants were interviewed by the researcher using a video-based conversation programme, Google Hangouts. While all of these Japanese residents had arrived at the USA after the age of 16, the length of their stay in English-speaking countries varied considerably, ranging from 1 month to 34 years (summarized in Table 1). Efforts were made to recruit such a wide range of length of residence profiles, as the experience variable was used as a means of having low- to high-level proficient performers.

At the same time, however, length of residence can only be considered as a rough parameter of L2 experience. This is because some L2 learners choose to operate in their L1 despite their extended stays in an L2 speaking environment (Jia & Aaron, 2001). In terms of the current study, the participants reported that their main language of communication at work, school or home was English and rated the frequency of L2 use beyond 4 out of 6 (*1 = very infrequent, 6 = very frequent*). Thus, the assumption here was that the current dataset covering the wide range of length of residence profiles could represent a wide range of L2 proficiency levels comprising low- to high-level proficient performers.

Table 1 *Length of Residence Profiles of 85 Japanese Learners of English*

Length of Residence	No. of participants
0 months	38
0.1-5 years	10
6-10 years	10
11-20 years	17
21-30 years	10

Raters (*n* = 10). A total of 10 L1 speakers of English were recruited at a university in the USA, and then assigned to two different groups: (a) *n* = five raters to assess global comprehensibility; and (b) *n* = five raters to assess lexical appropriateness. In terms of the number of raters, a great deal of methodological variation has been observed in the previous literature. For example, a few raters were recruited and trained in certain studies (e.g., Crossley et al., 2015 for *n* = 3 raters), more raters participated in other studies (e.g., Saito et al., 2016a for *n* = 10 raters). Importantly, it has been shown that it is not the number of raters, but their backgrounds that can significantly influence the process and product of L2 judgements of this kind (Kennedy & Trofimovich, 2008 for the role of professional L2 assessment experience). Instead of increasing the number of raters by recruiting as many individuals as possible despite their potentially varied linguistics and teaching experience and familiarity with foreign accepted speech, efforts were made to recruit an adequately sufficient number of raters with relatively homogeneous backgrounds (*n* = 5 for novice raters for L2 comprehensibility judgements and *n* = 5 for linguistically trained raters for L2 lexical appropriateness judgements). As a result, a strong agreement among each rater group was observed, suggesting that their comprehensibility and lexical appropriateness judgments could be considered sufficiently reliable (see the Results section below). Here, it was important to avoid asking the same raters to engage in both comprehensibility and lexical appropriateness judgements at the same time, because they are assumed to tap into two different constructs of L2 assessment and require two different kinds of rater backgrounds—novice raters' intuitive judgements (comprehensibility) vs. expert raters' professional evaluations (lexical appropriateness).

Speaking Task

The L2 learners' spontaneous speech was elicited via a picture description task widely adopted in L2 pronunciation (e.g., Derwing & Munro, 2015). A decision was made to use the same task, since it has been increasingly applied to L2 vocabulary research with a view of the generalizability of the topic (the linguistic correlates of L2 comprehensibility and lexical appropriateness) (e.g., Saito et al., 2016a). This task requires participant to describe an eight-frame cartoon, wherein two strangers bump into each other on the street and unintentionally swap their suitcases, which have a similar appearance. In the current study, participants were given one minute to familiarize themselves with the content of the story before describing the cartoon at their own pace. All the speech samples were recorded via a high-quality digital recorder set to a 44.1 kHz sampling rate with 16-bit quantization (for the foreign language learners in Japan); or the recording function available in Google Hangouts (for the Japanese residents in the USA). In most L2 comprehensibility research, a certain duration of the speech samples (e.g., 30 seconds) is excised and used for listeners' L2 comprehensibility judgements. To provide samples of sufficient duration for robust vocabulary analyses (e.g., Koizumi & In'nami, 2012), however, in the current study we used the entirety of each speech sample (*M* = 162.3 words, *Range* = 95-424 words). Given that the main focus of the study lay in vocabulary rather than fluency and pronunciation, we cleaned up each transcript by eliminating all filled pauses (ah, eh, oh, hmm) and fixing obvious pronunciation problems based on contexts (e.g., "I think" rather than "I sink"). Efforts were made to ensure that the administration of the tasks was comparable across the face-to-face meetings (*n* = 38) and video-based sessions (*n* = 47) by creating and following a strict protocol.

Comprehensibility vs. Lexical Appropriateness Judgements

Comprehensibility is defined as naïve raters' intuitive judgements about ease of understanding while assessing vocabulary aspects of language through reading transcribed L2 speech (Saito et al., 2016a). Thus, a decision was made to recruit undergraduate students who had never taken any courses in linguistics or language teaching, and reported little familiarity with Japanese-accented English on a 6-point scale (*1 = not all, 6 = highly familiar*) ($M_{familiarity}$ = 1.6). In contrast, lexical appropriateness refers to linguistically trained raters' holistic assessment of vocabulary use in transcribed L2 speech. The lexical appropriateness judgments require raters to receive some form of training so as to pay specific attention to the appropriate (rather than complex and fluent) use of vocabulary (rather than grammar). The precursor literature has shown that to execute such linguistic assessment without confusion, raters need a sufficient amount of experience related to L2 speech analyses (Saito et al., 2017). Thus, the appropriateness raters were graduate students in applied linguistics programs who had extensive experience with linguistics, speech analyses and EFL teaching (1-5 years). These raters' familiarity with Japanese accented English was reported to be relatively high ($M_{familiarity}$ = 5.4).

Comprehensibility and lexical appropriateness are two different constructs of L2 speech assessment, as the former is concerned with novice raters' intuitive judgements; and lexical appropriateness with expert raters' accuracy (but not fluency nor complexity) judgements. Importantly, lexical appropriateness judgements are also distinguishable from traditional accuracy analyses (i.e., counting errors as per obligatory contexts) and from general speaking proficiency ratings (i.e., evaluating multiple aspects of L2 speaking proficiency in line with certain descriptors as in TOEFL and IELTS). The notion and method of lexical appropriateness echoes Foster and Wigglesworth's (2016) weighted accuracy measures (i.e., classifying lexicogrammar errors according to their impacts on overall communicative success and adequacy (see also Révész et al., 2018).

Comprehensibility Judgements. Following the same rating procedure as in Saito et al. (2016a), all the transcripts were displayed on a screen in a randomized order via a MATLAB-based software program, Z-LAB. A cursor was available to scroll across the texts if a transcript could not completely fit on the screen. For each transcript, raters made an intuitive judgement in terms of comprehensibility by using a moving slider. Depending on where the cursor was located, their comprehensibility scores were automatically recorded on a 1000-point scale (*0 = difficult to understand, 1000 = easy to understand*).

Each rating session took place individually at a university in the USA. First, raters received an explanation from a trained research assistant on (a) the definition of comprehensibility (how easy to understand); (b) the purpose of the project; (c) the cartoon picture that L2 speakers needed to describe; and (d) the rating procedure (using a moving slider for judgements). Subsequently, the raters practiced the procedure by evaluating three transcripts (not included in the main dataset), and proceeded to the assessment of the main dataset. The entire session lasted for approximately 60-75 minutes per session (including both training and main ratings). As summarized in Table 2, the onscreen labels and training scripts for comprehensibility judgements fully focus on raters' processing of meaning without any mention of vocabulary use in L2 speech.

Table 2 *Training Scripts and Onscreen Labels for Comprehensibility*

Comprehensibility	This dimension refers to how much effort it takes to understand what someone is trying to convey. If you can understand (what the picture story is all about) with ease, then the speaker is highly comprehensible. However, if you struggle and must read very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.
Difficult to understand	↔ Easy to understand

Lexical Appropriateness Judgements. In previous research (Saito et al., 2017), raters have had some relevant backgrounds in linguistic analyses so that they can selectively (but subjectively) attend to and evaluate the appropriate use of vocabulary (for a similar subjective approach to L2 accuracy, see Foster & Wigglesworth, 2016 for weighted accuracy; Ruivivar & Collins, 2017 for morphosyntactic accuracy). The same methodology, training and materials were used in the current study. The final transcripts were displayed on screen in a randomized order via the MATLAB software. Using the same moving slider, recorded on a 1000-point scale: *0 = many inappropriate words, 1000 = consistently appropriate*), the raters read and assessed the transcripts in terms of the appropriate use of words in context. The five raters first received explanations of the definition of lexical appropriateness (adequate and natural choice of words in contexts), and then practiced the rating procedure with five speech samples not included in the main dataset. For each of these practice files, the raters were asked to explain their decisions, and received feedback from the researcher to ensure that they had correctly understood the nuanced aspect of L2 lexical proficiency targeted in the current study (appropriateness, but not fluency or complexity). Afterwards, the raters moved on to the main dataset (i.e., 85 transcripts). All the sessions took place individually. As summarized in Table 3, the training scripts and onscreen labels for lexical appropriateness judgements highlight the contextually accurate use of words but without providing any detailed guidance on breadth and depth aspects of L2 vocabulary use. The entire session lasted for 90-100 minutes (including training and main ratings).

Table 3 *Training Scripts and Onscreen Labels for Lexical Appropriateness*

Lexical appropriateness	This dimension refers to the semantic appropriateness of the vocabulary words used by the speaker. If the speaker uses incorrect or inappropriate words in context, including words from the speaker's native language, lexical accuracy is low. On the other hand, lexical accuracy is high if the speaker has all the lexical items required to accomplish the speaking task and does so using semantically precise lexical expressions.	
Many inappropriate words	↔	Consistently appropriate

Collocation Frequency and Association Measures

As conceptualized in Gablasova et al. (2017) and operationalized in L2 vocabulary research (e.g., Kyle & Crossley, 2015), the collocational qualities of L2 speech were objectively analyzed via the bigram and trigram measures available in the Tool for the Automatic Analysis of Lexical Sophistication 2.0 (TAALES) (Kyle & Crossley, 2015). Since the main objective of the study concerned the lexical profiles of L2 speech (not writing) among Japanese learners of English in EFL classrooms (where GA [General American] has generally been taught as a domain model) and in the USA (where GA is used as a main language of communication), the decision was made to use the Corpus of Contemporary American English (COCA) (Davies, 2009) as the reference corpus. Among the five subsections of the corpus, we chose the "spoken" dimension, which is comprised of conversations from a wide variety of TV and radio programs in the USA over the past 25 years. Three different types of collocation frequency and association scores were examined.

Collocation Frequency. To examine the extent to which using more frequent collocations relate to L2 comprehensibility and lexical appropriateness ratings, average frequency scores were calculated for each transcript. Since raw frequency scores are likely subject to a Zipfian distribution (intensive use of frequent combinations), these scores were adjusted via logarithmic transformation to approximate a normal distribution for use in the statistical analyses.

High-Frequency Associations (t-scores). To examine the strength of the associations of word combinations, we calculated t-score by dividing the difference between raw frequency and random co-occurrence frequency by the square root of the raw frequency. This index allows us to look at the extent to which the speech samples featured high-frequency collocations (e.g., *the* and *man*) while downgrading relatively random combinations of frequent words (e.g., *this* and *you*). In this sense, t-scores could be labelled as high-frequency associations.

Low-Frequency Associations (MI-scores). MI-scores were calculated by dividing the frequency of collocations by the frequency of random co-occurrence of the words. The scores were then logarithmized so that the figures could reveal how frequently speech samples constituted low-frequent combinations (e.g., *bump* and *into*) over high-frequent combinations (e.g., *look* and *into*). Since MI scores represent the degree of exclusivity in word partnerships, collocations with higher MI scores typically include lower-frequency words which do not have many partner words. In this sense, MI scores could be labeled as low-frequency associations (Gablasova et al. 2017).

To provide a more concrete picture of the n-gram analyses, several bigram and trigram examples with higher t- and MI-scores from the current dataset were listed in Table 4. By definition, t-scores weigh a set of word combinations which comprise more frequent words. Thus, those with higher t-scores inevitably featured function words (articles, pronouns, prepositions, and conjunctions) which could have many other partner words. Since MI-scores weigh a set of word combinations which are exclusively related to each other (with a limited number of partner words), those with higher MI-scores included content words (nouns, verbs, adverbs, and adjectives), and by extension appeared to be more structurally and semantically complete (Jeong & Jiang, 2019).

Table 4 *Summary of Bigram Examples with Higher T-Scores and MI-Scores*

t-scores	Bigram (> 250): Of the, in the, I think, on the, and then, kind of, at the
	Trigram (> 150): It be a, I do not, a lot of, be able to, do not know, there be a
MI-scores	Bigram (> 5.0): Little bit, each other, few days, pick up, exact same, take place
	Trigram (> 3.0): Depend on what, woman and man, around the corner, it looks like, in the middle, walk away from, go back home

Other Vocabulary Measures

To examine the *relative* predictive power of the collocation measures (logarithmic frequency, t-scores, MI-scores), the lexical characteristics of the same speech samples were also analyzed using single-word indices that have been adopted and found to show significant correlations with L2 oral proficiency in the previous literature. According to Crossley and colleagues' computational modeling of L2 lexical sophistication (Crossley et al., 2011; Kyle & Crossley, 2015), L2 proficiency can be defined as "both the depth and breadth of lexical knowledge available to speakers" (Kyle & Crossley, 2015, p. 759). The depth dimension was conceived as the extent to which a sample featured more abstract and complex words, while the breadth dimension was conceived as the extent to which more infrequent and specific words were used (see below).

Meaningfulness (depth). This index refers to how strongly words are associated in meaning with other words. Based on perceived meaningfulness scores available in the MRC psycholinguistic database (Coltheart, 1981) and two recently-added databases (Brysbaert & New, 2009; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), TAALES provides an average meaningfulness score for each sample. Salisbury, Crossley, and McNamara's (2011) have found that L2 learners' meaningfulness scores were predictive of L2 lexical development over time with their vocabulary use being less meaningful and more abstract.[4]

Hypernymy (depth). Depth of lexical knowledge can also be conceptualized in terms of the semantic hierarchy where a word is located (Crossley, Salsbury, & McNamara, 2009). For example, "color" is more abstract than "green" because the former is a superordinate term for the latter. In TAALES, hypernymy scores are calculated by dividing the total number of superordinate terms by the total number of words.

Frequency (breadth). An oft-used index of breadth knowledge is word frequency. While word frequency has been found to have strong associations with written lexical proficiency (Laufer & Nation, 1995), there is some evidence that this variable is linked to L1 speakers' judgements of L2 oral proficiency (Crossley et al., 2011). In TAALES, word frequency was calculated by dividing the total sum of frequency scores (in reference to COCA) by the number of all the words which were assigned a frequency score. Similar to n-gram frequency measures, raw word frequency scores were logarithmically transformed to control for Zipfian effects.

Range (breadth). Another crucial variable of breadth concerns how often texts feature words which occur across a broad or narrow range of sources in COCA. This variable is crucial, since more proficient L2 learners are assumed to use more specific and less general words which have been more narrowly used and observed in certain contexts and genres. Prior research identified range as a secondary predictor of speaking proficiency (Kyle & Crossley, 2015). In this study, logarithmically-transformed range scores were used for the analyses.

[4] Meaningfulness scores refers to the extent to which a word is connected with others. In the relevant database (MRC), meaningfulness was assessed by native speakers on a 7-point scale. While more meaningful words (e.g., *man*, *woman*, *city*) evoke many other related words and collocations, less meaningful words (e.g., *on, the, a*) result in limited links.

Results

Comprehensibility Scores

Similar to previous research (e.g., Saito et al., 2016a), the results of Cronbach alpha analyses demonstrated relatively high agreement for comprehensibility (α = .92). In light of the consistency among the five raters, scores were averaged across to provide a single comprehensibility score per talker. The comprehensibility scores widely ranged from 201 to 908 (*M* = 527, *SD* = 176); the results of a one-sample Kolmogorov-Smirnov test confirmed that the comprehensibility scores were normally distributed (*p* = .404).

Lexical Appropriateness Scores

The five raters' appropriateness judgements yielded relatively high agreement (Cronbach α = .90), which is comparable to previous studies (e.g., Foster & Wigglesworth, 2016 for .90; Saito et al., 2017 for .95). To check the reliability of the raters' lexical appropriateness evaluations, at the endpoint of the rating session the raters assessed the extent to which they understood the rated categories on a 9-point scale (*1 = "I did not understand at all", 9 = "I understand this concept well"*). Their understanding of the rated category (lexical appropriateness) was relatively high (*M* = 8.7). Similar to the L2 comprehensibility judgements, the raters' lexical appropriateness scores were averaged for each transcript. The lexical appropriateness scores ranged from 108 to 823 (*M* = 444, *SD* = 190) and followed a normal contribution (a one-sample Kolmogorov-Smirnov test: *p* = .538).

Comprehensibility vs. Lexical Appropriateness

The averaged comprehensibility and lexical appropriateness scores were strongly correlated with each other, *r* = .815, *p* < .001. This suggests that comprehensibility and lexical appropriateness overlap to a great degree (66.4% of variance), but that there is distinct variation that can be uniquely explained by either (33.6%). In essence, comprehensibility and lexical appropriateness overlap with each other to a great degree, but may also represent slightly different constructs of L2 oral proficiency.

Lexical Correlates of L2 Comprehensibility and Lexical Appropriateness

According to the results of Kolmogorov-Smirnov tests, the distribution patterns of Bigram t-scores were found to be negatively skewed (*p* = .027). Therefore, these scores were transformed using the Log10 function for subsequent analyses. The log-transformed t-scores demonstrated acceptable normality (*p* > .05). A set of Pearson correlation analyses were then performed to analyze how the lexical variables (collocation, depth, breadth) were associated with the raters' L2 comprehensibility and lexical appropriateness scores. The strength of correlations was evaluated in conjunction with Plonsky and Oswald's (2014) field-specific benchmarks (*r* = .25 for small, .40 for medium, .60 for large).

As shown in Table 5, all the lexical variables except word frequency were significantly correlated with L2 comprehensibility and lexical appropriateness scores ($p < .05$). Whereas the strength of the lexis-proficiency correlations was generally small-to-medium, the low-frequency collocation index (MI scores) demonstrated relatively strong predictive power (*r* = .676-.734 for comprehensibility; *r* = .641-.755 for lexical appropriateness). It is also worth nothing that single-word depth and breadth indices were negatively correlated with L2 comprehensibility and lexical appropriateness scores (more infrequent, specific and abstract words positively impacted L2 comprehensibility and lexical appropriateness), while the collocation indices yielded positive correlations (more frequent, more strongly associated collocations related to better L2 comprehensibility and lexical appropriateness).

Table 5 *Correlations between Lexical Variables and L2 Comprehensibility and Lexical Appropriateness Scores*

	Comprehensibility		Lexical Appropriateness	
	r	*p*	*r*	*p*
Collocation: Bigram				
Frequency (log-transformed)	.226	.037	.270	.012
High-frequency associations (t-scores)[a]	-.333	.002	-.322	.003
Low-frequency associations (MI-scores)	.734	< .001	.755	< .001
Collocation: Trigram				
Frequency (log-transformed)	.262	.015	.197	.071
High-frequency associations (t-scores)	.408	< .001	.335	.002
Low-frequency associations (MI-scores)	.676	< .001	.641	< .001
Depth				
Meaningfulness	-.340	.001	-.445	< .001
Hypernymy	-.322	.003	-.374	< .001
Breadth				
Frequency	-.099	.368	-.099	.369
Range	-.230	.035	-.212	.051

Note. [a]bigram t-scores were transformed through the Log10 function

Relationships among Collocation, Depth and Breadth Measures

To further examine how the collocation indices related to the other major dimensions of L2 lexical knowledge (depth, breadth), the 10 lexical scores were submitted to a factor analysis with promax rotation and Kaiser normalization. The objective of the factor analysis was to identify the larger lexical categories underlying the lexical variables. Following Loewen and Gonulal's (2015) field-specific recommendations, the suitability of the factor analysis for the current dataset (*n* = the 10 lexical scores relative to 85 samples) was carefully checked. First, the sample size ($n = 85$) could be considered beyond the minimum threshold ($n = 50$) (see also de Winter, Dodou, & Wieringa, 2009). Second, the factorability of the entire dataset was examined and validated via two tests: Bartlett's test of sphericity ($\chi^2 = 491.755$, $p < .001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (.655). Third, the model explained a total of 83.3% of variance in the outcomes of the L2 vocabulary analyses, which is beyond the norm in the field of L2 research (60-70%). Because we identified four factors with eigenvalues above 1, the decision was made to specify a four-factor solution.

As summarized in Table 6, Factor 1 included both low-frequency collocations (MI scores of bigram and trigram) and lexical depth (meaningfulness, hypernymy), indicating that this factor captured the extent to which L2 users could access less frequent combinations of more abstract and less frequent words (i.e., Advanced Collocation Use). Factor 2 encompassed two dimensions of lexical breadth: word frequency and range. Since both dimensions tap into learners' ability to access single-word units, this factor was labelled as "Single-Word Use". This factor was thought to represent the extent to which L2 users could use more infrequent and specific words in their speech. Finally, Factors 3 and 4 clustered the high-frequency trigram and bigram indices into two different groups. Therefore, they were labeled as "Three-Word Frequency" and "Two-Word Frequency," respectively. These factors together were presumed to illustrate the extent to which L2 users could access more frequent combinations of three words (Factor 3) and two words (Factor 4) in speech production.

Table 6 *Factor Analyses of Lexical Variables*

	Factor 1 (Advanced Collocation Use)	Factor 2 (Single Word Use)	Factor 3 (Frequent Three Words)	Factor 4 (Frequent Two Words)
Collocation: Bigram				
Frequency (log-transformed)	.099	.155	.096	**.901**
High-frequency associations (t-scores)[a]	-.113	.069	-.262	**-.867**
Low-frequency associations (MI-scores)	**.700**	-.325	.301	.404
Collocation: Trigram				
Frequency (log-transformed)	.067	.121	**.913**	.217
High-frequency associations (t-scores)	.304	-.009	**.888**	.171
Low-frequency associations (MI-scores)	**.764**	-.308	.401	-.040
Depth				
Meaningfulness	**-.762**	-.239	-.161	-.009
Hypernymy	**-.709**	-.386	.142	-.272
Breadth				
Frequency	.015	**.932**	-.031	.031
Range	.079	**.916**	.133	.034

Note. All loadings > .7 were highlighted in bold; [a]bigram t-scores were transformed through the Log10 function

Lexical Predictors of L2 Comprehensibility and Lexical Appropriateness

The next objective of the statistical analyses was to examine the relative influence of the four lexical factors—Advanced Collocation Use, Single-Word Use, Three-Word Frequency, and Two-Word Frequency—on the L1 raters' comprehensibility and lexical appropriateness judgements. Accordingly, two sets of stepwise multiple regression analyses were performed with comprehensibility scores and lexical appropriateness scores as the dependent variables and the four lexical factor scores as independent variables.

To determine the suitability of conducting multiple regression analyses, we first checked several assumptions. First, as explained in the manuscript, the 10 raw lexical variables originally included in the aforementioned correlation analyses were reduced to 4 predictors through Factor Analyses. Second, the normality of each dependent and independent variable was confirmed by Kolmogorov-Smirnov tests ($p > .05$).

As summarized in Table 7, the first regression model explained 53.1% of variance in naive raters' L2 comprehensibility judgements, $R = .729$, $F(4, 80) = 22.688$, $p < .05$. According to this model, L2 comprehensibility was most strongly associated with Advanced Collocation Use (35.3%), followed by Three-Word Frequency (7.3%), Two-Word Frequency (5.8%) and Single-Word Use (4.8%). Similarly, the second regression model accounted for 54.7% of variance in raters' L2 lexical appropriateness judgements, $R = .740$, $F(4, 80) = 32.575$, $p < .05$, with Advanced Collocation Use being the strongest predictor (42.8%) compared to Two-Word Frequency Use (5.9%) and Single-Word Use (5.9%).

The results here suggest (a) that raters' speech judgements greatly rely on the collocational information in terms of comprehensibility (48.4% for Advanced Collocation Use together with Two/Three-Word Frequency) and lexical appropriateness (48.8% for Advanced Collocation Use and Two-Word Frequency) relative to the single word information (4.8-5.9% for One Word Use); and (b) that the raters appeared to weigh low-frequency collocations more heavily during lexical appropriateness judgements (42.8%) compared to those during comprehensibility judgements (35.3%).

Table 7 *Results of Multiple Regression Analysis Using Lexical Variables as Predictors of L2 Comprehensibility and Lexical Appropriateness*

Predicted variable	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility	Advanced Collocation Use	.353	.353	45.222	< .001
	Three Word Frequency	.426	.073	30.412	< .001
	Two Word Frequency	.484	.058	25.281	< .001
	One Word Use	.531	.048	22.688	< .001
Lexical appropriateness	Advanced Collocation Use	.428	.428	62.169	< .001
	Two Word Frequency	.488	.059	39.036	< .001
	One Word Use	.547	.059	32.575	< .001

Effects of Text Length on Proficiency-Collocation Link

Given that the results hinted that some collocation measures (MI) could be strongly tied to perceived L2 oral proficiency, L2 oral proficiency ratings could also be influenced by the length of samples to some degree (e.g., Iwashita et al., 2008). Not surprisingly, therefore, the correlation analyses identified the significant relationship between text length and two different types of L2 oral ratings—i.e., comprehensibility ($r = .620, p < .001$) and lexical appropriateness ($r = .751, p < .001$). The results suggest that longer speech samples could be perceived more comprehensible and appropriate, simply because they may pack more linguistic information available for raters to better understand intended message.

Thus, it is important to test whether and to what degree text length can affect the strength of the proficiency-collocation link reported above. To this end, the two perceived oral proficiency scores, comprehensibility and lexical appropriateness, were regressed to text length. Next, these residual values were re-submitted to the same stepwise multiple regression models relative to the four lexical factor scores as predictor variables (Advanced Collocation Use, Single Word Use, Frequent Three Words, Frequent Two Words). As summarized in Table 8, the use of low-frequent collocations still remained as a statistically significant predictor, explaining a good amount of variance in comprehensibility (17.6%) and lexical appropriateness ratings (25.5%) at a $p < .001$ level, even after the text length factor was controlled for. The results here suggest (a) that both collocation and text length comprise an overlapping, composite factor that raters substantially rely on during their subjective judgements of L2 speech (explaining about 50-60%); and (b) that the collocation factor alone still makes an *independent* contribution to such perceived L2 oral proficiency (explaining about 20-30%).

Table 8 *Results of Multiple Regression Analysis Using L2 Comprehensibility and Lexical Appropriateness with Text Length Partialled out*

Predicted variable	Predictor variables	*Adjusted R^2*	*R^2 change*	*F*	*p*
Comprehensibility (residual values)	Advanced Collocation Use	.176	.176	17.716	< .001
	One Word Use	.243	.067	13.155	< .001
Lexical appropriateness (residual values)	Advanced Collocation Use	.255	.255	28.438	< .001
	One Word Use	.315	.060	18.882	< .001

Note. Comprehensibility and lexical appropriateness scores were regressed to text length

Discussion

The current study examined the extent to which collocation information, operationalized via different n-gram indices (high- and low-frequent bigrams and trigrams) could predict the comprehensibility and lexical appropriateness of the L2 speech of 85 L1 Japanese learners of English. The predictive power of the collocation measures was compared with that of single-unit measures tapping into depth (meaningfulness, hypernymy) and breadth (frequency, range) aspects of L2 vocabulary use to examine the *independent* contributions of collocation to perceived L2 oral proficiency beyond single-word units. As Gablasova et al. (2017) pointed out, the ratio of low-frequency collocation (Mutual Information scores) is likely correlated and confounded with single-word frequency factors. Therefore, it is crucial to statistically tease apart any effects of collocation (MI in particular) from other lexical factors. The current study solved this issue by revealing the factors underlying the 10 lexical factors via a factor analysis; and used the lexical factor scores to check the relative importance of collocation in perceived L2 oral proficiency scores via a multiple regression analysis.

To date, many scholars have examined the role of vocabulary in trained raters' judgements of low-, mid- and high-level L2 speaking proficiency, as determined by ACTFL (e.g., Crossley et al., 2011; Eguchi & Kyle, forthcoming) and TOEFL (Crossley & McNamara, 2013; Kyle & Crossley, 2015) descriptors and scoring rubrics. To my knowledge, the current study was the first attempt to re-examine the topic using L1 raters' *intuitive* and *subjective* judgements of comprehensibility (ease of understanding) and lexical appropriateness (adequate and natural choice of words), simulating what L1 and L2 users typically do when they interact with each other in real-life communication. Despite the methodological differences, the results of the current study match those of Kyle and Crossley (2015), confirming that L2 oral proficiency is primarily influenced by collocational qualities (48.4-48.8% of variance); and secondarily by single-word frequency and range (4.8-5.9% of variance). The strength of the collocation effects reported here (*r* = .641-.755 explaining 48% of variance) could be considered large in reference to Plonsky and Oswald's (2014) field-specific benchmarks.

It is important to emphasize here that one specific aspect of L2 collocation (i.e., Mutual Information rather than t-scores) was identified as a key component of the L2 oral proficiency judgements. While both t-scores and MI are mathematically designed to index distinctive meanings (rather than random co-occurrences), they correspond to two different types of collocational association. Whereas t-scores are sensitive to the use of high-frequent collocations, MI units favour less-commonly used combinations (i.e., combinations of more infrequent, abstract and complex words). Concurring with the existing L2 writing research (e.g., Durrant & Schmitt, 2009), the current study demonstrated that L1 speakers' judgements were more strongly linked to MI than to t-scores.

The argument is in line with the theoretical view that L1 speakers store word-combinations as memorized chunks in a way that allows them to develop sensitivity to the use of collocations in language, and to recognize and produce them more quickly and accurately (Ellis et al., 2008). Here, type of collocation associations matters. Given that multiword units with high t-scores likely consist of high-frequency words which could collocate with a number of other candidates (Gablasova et al., 2017), they may not necessarily help raters engage in efficient lexical selection and processing. Comparatively, the use of collocations with high MI scores may play a facilitative role in L2 assessment, arguably because their constituents are relatively infrequent, abstract and complex words (see also the results of the factor analyses). Since these words are exclusively tied to a limited number of collocates (without too many other competitors), low-frequency collocations may make L2 discourse more predictable, easier to follow and by extension more semantically precise.

Given that L2 learners' speech becomes increasingly comprehensible, natural, and appropriate after long periods of immersion (Derwing & Munro, 2015; Tavakoli, 2018), it is reasonable to assume that adult L2 speech learning will develop on a continuum of low, mid and high L2 comprehensibility and lexical appropriateness. Under this assumption, the cross-sectional findings of the current study can be interpreted as providing indirect support for frequency effects in adult L2 speech learning—i.e., L2 learners come to start using not only frequent and concrete, but also infrequent and abstract vocabulary in tandem with increased L2 proficiency and experience. Importantly, there is some evidence that such frequency effects may not be clearly observed when analyses focus only on single-word units (Crossley, Skalicky, Kyle & Montero, 2019). Rather, as suggested by the results of the current study, frequency may serve as a developmental index if interpreted in terms of collocations (two-to-three word units). The argument here echoes other cross-sectional findings that more advanced L2 learners produce more low-frequent collocations with higher MI-scores during L2 writing tasks (Durrant & Schmitt, 2009; Garner et al., 2018) and word association tasks (Clenton, 2015; cf. Zareva & Wolter, 2014).

The analysis also uncovered an interesting pattern regarding the role of low-frequency collocations in two different types of perceived L2 oral proficiency judgements—comprehensibility and lexical appropriateness. Specifically, raters seemed to rely more on mutual information during their lexical appropriateness judgements (42.8% of variance) than they did during their comprehensibility judgements (35.8% of variance). The relative weight of the low-frequency collocations in comprehensibility and lexical appropriateness (42.8% vs. 35.8%) suggests that these measures may tap into two overlapping yet different constructs of L2 oral proficiency. It is important to remember that lexical appropriateness by definition corresponds to one form of L2 accuracy (Saito et al., 2017), but that comprehensibility is equally tied to lexical appropriateness, fluency and richness (Crowther et al., 2017). What emerges from this observation is that since the low-frequency collocation factor is a primary determinant of lexical appropriateness (Siyanova & Schmitt, 2008), it is correspondingly a crucial component of the more global construct of comprehensibility (cf. Saito et al., 2016a).

Limitations

As the current investigation took a first step towards assessing the role of collocation in perceived L2 oral proficiency, the findings reported here need to be considered exploratory at best. With a view towards future replication endeavours, I would like to acknowledge a set of limitations of the current study. First, all the findings were solely based on the analyses of speech samples elicited from a single task (picture description). Previous research has shown that L2 learners' performance can greatly differ according to task-specific variables, such as planning time, repetition, and task complexity. In particular, it would be intriguing to explore the extent to which the predictive power of collocation could be generalizable especially when participants engage in more freely structured tasks (e.g., oral interviews, monologues). Different from picture narratives (accurately describing given information), these tasks are believed to prompt speakers to construct, elaborate and expand their own intended message by using a wide variety of words (Foster & Skehan, 1996).

Second, while collocation (MI scores in particular) demonstrated strong associations with L2 oral proficiency (comprehensibility, lexical appropriateness), the relationship between these factors was also confounded with the length of speech samples for obvious reasons: Less proficient L2 speakers tend to produce less words so that they have less opportunities to use low-frequent multiword combinations (see Iwashita et al., 2008). Although the results pointed out that the collocation-proficiency link remained significant, even after text length was statistically factored out, the strength of its predictive power became readily weaker. The findings indicate that both collocation and text length should be considered as one overlapping phenomenon underlying L2 oral proficiency; and that using statistically combined scores (collocation plus text length) could be an optimal option to explain the utmost amount of variance in L2 oral proficiency. Alternatively, future studies need to suggest, test and validate an adequate length threshold for calculating collocation factors (cf. In'nami & Koizumi, 2013 for lexical diversity).

Third, the current study has exclusively focused on comprehensibility and lexical appropriateness as an index of L2 oral proficiency. These two measures have one methodological feature in common—i.e., raters are asked to evaluate the global qualities of L2 speech based on their own standard of comprehensibility or appropriateness without any reference to pre-existing rubrics. Future studies should further pursue the collocation effects in such L2 speech assessments by including other well-researched subjective measures of L2 oral proficiency. One such example is communicative adequacy. Given that Révész et al. (2016) found that raters appeared to rely on a range of vocabulary information (e.g., diversity) during L2 communicative adequacy judgements, it would be intriguing to explore the extent to which communicative adequacy can be related not only to single-word, but also multiword indices.

Fourth, although the cross-sectional findings suggest that more proficient L2 speakers may use more low-frequent collocations, they need to be replicated with a longitudinal research design. There have been a growing number of empirical studies showing that L2 learners can enhance their collocation use as their written proficiency increases over time (e.g., Garner et al., 2018). When it comes to L2 oral proficiency development, previous longitudinal research has focused only on the lexical analyses of single-word units (e.g., Crossley et al., 2009 for depth knowledge; Crossley et al., 2019 for breadth knowledge). It would be interesting to extend this vein of L2 speech research by corroborating the relationship between collocation knowledge and vocabulary use (cf. Garner & Crossley, 2018; Kim et al., 2018), and its ultimate impact on perceived L2 oral proficiency development, over a longer period of time (e.g., Derwing & Munro, 2015 for comprehensibility; Foster & Wigglesworth, 2016 and Tavakoli, 2018 for weighted accuracy).

Last, the current study asked the raters to assess for global and lexical accuracy of transcribed L2 speech samples so as to control for the influence of phonological errors on their judgements. The methodology here has been widely used in order to examine the role of lexicogrammar use in rater behaviours during their L2 speech evaluations (e.g., Crossley et al., 2015 for lexical proficiency; Foster & Wigglesworth, 2016 for weighted accuracy). The method has also been recently extended to the paradigm of L2 comprehensibility (e.g., Saito et al., 2016a). However, it still remains open to further investigation whether and to what degree the lexical correlates of L2 speech assessments differ when they read transcripts (where all phonological errors removed) or when they actually listen to audio samples (where the influence of lexical errors interacting with those of phonological deviations). While transcript reading allows raters to have time to re-read and decode texts when needed, such reflection is not possible during listening (more fleeting in nature).

At the same time, the relationship between lexicogrammar, phonology and L2 global assessment is highly complex. For example, Issacs and Trofimovich (2012) found that listeners' judgements of L2 comprehensibility could be equally tied to vocabulary (type and token frequency) and phonology (prosody, vowel reduction), suggesting that L2 vocabulary use does play a crucial role in L2 comprehensibility of this kind, even if a speech sample is phonologically intelligible, and vice versa. Other follow-up studies have pointed out that the ratio of phonological and lexicogrammar influences on L2 comprehensibility greatly vary according to task conditions (e.g., Crowther et al., 2017 for simple vs. complex tasks), rater type (e.g., Saito & Shintani, 2016 for monolinguals vs. multilinguals; Saito, Train, Suzukida, Sun, Magne & Ilkan., 2019 for L1 vs. L2 raters) and proficiency levels (Saito, Trofimovich, & Isaacs, 2016 for stronger lexicogrammar effects in low-to-mid comprehensibility and stronger phonological effects in mid-to-high comprehensibility). In this regard, future studies should revisit the relatively strong effects of collocation in perceived L2 oral proficiency from different methodological perspectives. It would be interesting to compare the findings resulting from raters' evaluations of transcripts and audio samples (cf. Saito et al., 2016a for transcript judgements vs. Saito, Webb, Trofimovich, & Isaacs, 2016b for audio judgements).

Conclusion

The current study provides further evidence for the role of multiword units in determining L2 speaking proficiency (e.g., Kyle & Crossley, 2015), extending the findings to cover two different types of intuitive and subjective L2 speech judgements: comprehensibility (ease of understanding) and lexical appropriateness (adequate and natural choice of words in contexts). The study possesses a certain degree of ecological validity given that both of these constructs are assumed to mirror what L1 and L2 speakers do when interacting with L2 users (Crowther et al., 2017; Derwing & Munro, 2015; Saito et al., 2017). Taken together, the findings showed that the natural use of more infrequent, abstract and complex multi-word units has a strong, direct impact on perceived L2 oral proficiency. While the conclusion here parallels the recent research evidence on the role of collocation in L2 writing proficiency (e.g., Garner et al., 2018 for r = .20-.30 in CEFR Writing), it is important to add that the collocation-proficiency link could be relatively large in the context of perceived L2 comprehensibility and lexical appropriateness (r = .30-.70 for comprehensibility).

The study has several crucial implications for future research on L2 speech assessment and teaching. First, recent research has emphasized the use of human judgments in determining the global accuracy of L2 speech (e.g., Foster & Wigglesworth, 2016). While subjective judgments provide invaluable information about L2 speaking proficiency and collocation (cf. for rater judgements of formulaic sequences, see Boers et al., 2006; Stengers et al., 2011), the current study shows the value of complementing them with certain forms of more objective and automated analyses, namely, indices of low-frequency collocation (i.e., Mutual Information) available in TAALES (Kyle & Crossley, 2015).

Second, much attention has been given to exploring how to best measure L2 speakers to use vocabulary (i.e., productive vocabulary knowledge) (Fitzpatrick & Clenton, 2017). Some studies have begun to show that L2 users' responses to word association tasks could predict how they use collocations during actual L2 speech performance. In conjunction with the strong associations between collocation and perceived L2 oral proficiency reported in the current study, it would be interesting to use the results of MI-scores measured through a word association task or a related research tool (e.g., Lex30: Meara & Fitzpatrick, 2000) as a rough index of perceived L2 oral proficiency (Clenton, De Jong, Clingwall, & Fraser, forthcoming; Uchihara & Saito, 2018; Uchihara, Eguchi, Kyle, Clenton & Saito, forthcoming). If L2 learners' word association scores can predict collocational qualities of their L2 speech, and by extension perceived L2 oral proficiency, it will greatly help teachers diagnose L2 learners' speaking proficiency quickly, adequately and automatically.

Finally, the current study brought to light the possibility that collocation use may be an anchor of L2 oral proficiency development. It is noteworthy, however, that L2 learners' collocation knowledge has been reported to be limited in many L2 classrooms (e.g., Nguyen & Webb, 2017). In this regard, teachers and learners are advised to pay more attention to how L2 learners' collocation knowledge can be enhanced through research-based instructional techniques, such as intentional teaching in reference to lists of multiword expressions (identified by researchers' intuition, Wray, 2000, or/and corpus-based data, Ellis et al., 2008), or providing typographically enhanced input during incidental learning (e.g., reading) (Szudarski & Carter, 2016). For a comprehensive review on this topic, see Pellicer-Sanchez and Boers (2018).

In L2 speech research, it has been shown that interaction with L1 and advanced L2 speakers is instrumental to L2 oral proficiency development (Flege, 2016). Interaction is believed to allow L2 learners to receive not only positive evidence (exposure to more L1-like, natural and adequate word combinations), but also negative evidence (receiving a signal of errors especially in the case of communication breakdown) (Mackey, 2012). For example, Saito and Akiyama (2017) demonstrated the effectiveness of the negotiation-for-comprehensibility training on L2 lexical appropriateness development, whereby L1 speakers are trained to provide feedback when their L2 interlocutors make lexical errors hindering successful communication and comprehensibility. The pedagogical potential of such activities should be further examined with a greater focus on the treatment of word and collocation choice errors, repair and acquisition.

References

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

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

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65-78.

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

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

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

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

Clenton, J., De Jong, N., Clingwall & Fraser (forthcoming). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*, 497-505.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171-192.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009) Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*, 307–34.

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., & Skalicky, S. (2017). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, 1-21.

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*. DOI: 10.1017/S0272263118000268

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of second language accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40,* 443-457.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*, 159-190.

de Winter, J. D., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147-181.

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

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*, 157-177.

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

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of Applied Linguistics*, *32*, 17-44.

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

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*, 29-62.

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly*, *51*, 844-867.

Flege, J. (2016, June). *The role of phonetic category formation in second language speech acquisition*. Plenary address delivered at New Sounds, Aarhus, Denmark.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75-94). Harlow, UK: Pearson Education.

Foster, R., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98-116.

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

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

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176-187.

Hardie, A., Baker, P., McEnery, T., & Jayaram, B. D. (2006). Corpus-building for South Asian languages. In A. Saxene, & L. Borin (Eds.), *Lesser-known languages in South Asia: Status and policies, case studies and applications of information technology* (pp. 211-242). Mouton de Gruyter.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. John Benjamins Publishing.

Ishikawa, S. (2014). Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2* (pp. 63-76). Kobe, Japan: Kobe University

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 29–49.

Jeong, H., & Jiang, N. (2019). Representation and processing of lexical bundles: Evidence from word monitoring. *System, 80*, 188-198.

Kang, O., Rubin, D., Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English, *Modern Language Journal, 94,* 554–566.

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

Koizumi, R. (2012). Vocabulary and speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell: Wiley-Blackwell.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*, 554–564.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978-990.

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

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12-24.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, *50*, 675-724.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In Plonsky, L. (Ed), *Advancing quantitative methods in second language research*. New York: Routledge.

Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation*, *3*, 167-198.

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

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520–531.

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

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, *63*, 1-24.

Pellicer-Sánchez, A., & Boers, F. (2018). Pedagogical approaches to the teaching and learning of formulaic language. In A. Siyanova & A. Pellicer-Sánchez. (Eds.), *Understanding formulaic language: A second language acquisition perspective* (Chapter 8). Routledge.

Pennycook, A. (2017). *The cultural politics of English as an international language*. Routledge.

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

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*, 828-848.

Ruivivar, J., & Collins, L. (2018). The effects of foreign accent on perceptions of nonstandard Grammar: A pilot study. *TESOL Quarterly*, *52*, 187-198.

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning, 67*, 43-74.

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly, 50*, 421-446.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity and metacognition. *Studies in Second Language Acquisition*. DOI: 10.1017/S0272263119000226

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. (2016a). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition, 37*, 677-701.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016b). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition, 19*, 597-609.

Salsbury, T., Crossley, S. A, & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27*, 343–360.

Schmitt, N. (2008). State of the art: Instructed second language vocabulary acquisition. *Language Teaching Research, 12,* 329–363.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*, 429-458.

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*, *36*, 549-569.

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

Suzuki, S., & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in* Second *Language Acquisition.* DOI: https://doi.org/10.1017/S0272263119000421

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the Functional Load principle. *Language Teaching Research*. DOI: https://doi.org/10.1177/1362168819858246

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics*, *26*, 245-265.

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

Tavakoli, P., & Uchihara, T. (in press). To what extent are multiword sequences associated with oral fluency? *Language Learning.*

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes, 30,* 392–408.

Tomasello, M. (2003). *Constructing a Language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K., & Saito, K. (forthcoming). To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association.

Uchihara, T., & Saito, K. (2018). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal, 47*, 64-75.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463-489.

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