To be published in *Second Language Research* (SAGE)

Roles of Collocation in L2 Oral Proficiency Revisited: Different Tasks, L1 vs. L2 Raters, and Cross-Sectional vs. Longitudinal Analyses

Kazuya Saito[1]

Yuwei Liu

Abstract

There is emerging evidence that collocation use plays a primary role in determining various dimensions of L2 oral proficiency assessment and development (e.g., Eguchi & Kyle, 2020; Kyle & Crossley, 2015; Saito, 2020). The current study presents the results of three experiments which examined the relationship between the degree of association in collocation use (operationalized as t and mutual information scores) and the intuitive judgements of L2 comprehensibility (i.e., ease of understanding). The topic was approached from the angles of different task conditions (Study 1), rater background (L1 vs. L2) (Study 2) and cross-sectional vs. longitudinal analyses (Study 3). The findings showed that (a) collocation emerged as a medium-to-strong determinant of L2 comprehensibility in structured (picture description) compared to free (oral interview) oral production tasks; (b) with sufficient immersion experience, L2 raters can demonstrate as much sensitivity to collocation as L1 raters; and (c) conversational experience is associated with more coherent and mutually-exclusive combinations of words in L2 speech, resulting in greater L2 comprehensibility development.
.

Keywords: collocation, comprehensibility, speech, vocabulary, pronunciation

[1] We are grateful to Masaki Eguchi, *Second Language Research* reviewers, and the journal editor, Alice Fort for their insightful comments on earlier versions of the manuscript. The project was funded by Arnold Bentley New Initiatives Fund, Leverhulme Trust Research Grant (RPG-2019-039), and Spencer Foundation Research Grant (202100074). Corresponding Author: Kazuya Saito, University College London, Institute of Education, 20 Bedford Way, WC1H 0AL, United Kingdom. Email: k.saito@ucl.ac.uk

Highlights

- We examined the role of collocation use in perceived L2 comprehensibility.
- Raters tend to sensitize the mutual exclusivity of word associations.
- L2 development is tied to the use of more mutually-excusive combinations of words.
- Collocation effects are strong when speech is elicited via well-structured tasks.
- The findings suggest collocation as a cornerstone of L2 speech assessment/development.

Introduction

Whereas scholars have begun to examine the lexical characteristics of second language (L2) speech which can be judged to be fluent (Tavakoli & Uchihara, 2020), comprehensible and contextually appropriate (Saito, 2020), and highly proficient (Kyle & Crossley, 2015), the existing literature has suggested the use of multiword units (collocations) as a key factor when assessing L2 vocabulary. In this investigation, we aim to examine the variance in the predictive power of collocation under different task and rater conditions from both cross-sectional and longitudinal perspectives.

Background

Collocation, N-Gram, and L2 Proficiency Judgements

According to usage-based accounts of L2 acquisition, language is formulaic in nature, with language exemplars stored as "chunks" in the mental lexicon (Bybee & Hopper, 2001). Such chunks, which represent a single or function, are termed as *formulaic sequences* (Wray, 2002). It is thought that sufficient exposure to these chunks in a variety of different contexts can help learners automatize their access to them in response to any relevant contextual and linguistic cues (Ellis, 2012). There is ample research evidence that both L1 and L2 speakers receive and produce collocations more rapidly, accurately, and subconsciously than novel strings of words (e.g., Ellis, Simpson-Vlach, & Maynard, 2008; Sonbul, 2015). There is some corpus research showing that multiword combinations make up approximately half of written and spoken English (Erman & Warren, 2000), and are particularly characteristic of oral discourse among native

speakers (Biber, Johansson, Leech, Conrad, & Finegan, 1999), although such estimates may vary as per scholars' definitions and operationalizations of collocation.

To date, researchers have illustrated the formulaicity of language using various different constructs, such as collocation, n-grams, and lexical bundles (for comprehensive reviews, see Wood, 2019). To further understand the various illustrations of formulaicity, one useful notion in corpus linguistics concerns (a) co-occurrence and (b) recurrence (Paquot & Granger, 2012).

Co-occurrence "consists in the co-selection of (usually) two lexical items, which may be, but are not necessarily, contiguous." (Paquot & Granger, 2012, p. 136). Broadly speaking, one such form of co-occurrence is collocation, defined as adjacent word pairs that co-occur repeatedly in certain contexts (e.g., "reception desk" "information desk" "sports desk"), and/or as a part of linguistic functions (e.g., phrasal verb plus object for "think of him"). In the field of L2 vocabulary research, however, the precise definition of collocation widely varies across studies (see also Boers & Webb, 2018 for the phraseology vs. corpus-based approaches towards collocation).

The other dimension, reoccurrence refers to "the repetition of contiguous strings of words of a given length (e.g., bigrams, trigrams)" (Paquot & Granger, 2012, p. 138). Following this line of thought, the *lexical bundle* approach conceptualizes formulaic language based only on the frequency of discoursal rather than semantic functions (Biber et al., 1999). Thus, any combinations of words are allowed as long as they frequently occur in a reference corpus regardless of the semantic partnership of word combinations ("think highly of" and "think of the"). Under this lexical bundle approach, a growing number of scholars have focused on the frequency of multiword expressions consisting of specific numbers ("n") of words (i.e., *n-grams*) in a reference corpus as an index of collocation. Raw n-gram frequency scores include not only semantically and structurally complete sequences (e.g., "think" and "of"), but also random co-occurrences of incomplete lexical items (e.g., "think" and "desk").

Numerous adjusted measures have been devised to index the strength of meaningful associations (i.e., greater than chance). They differ in terms of how they capture three dimensions of formulaicity—dispersion (the extent to which a particular combination of words occurs *across* a reference corpus), exclusivity (the extent to which one word occurs exclusively with specific partner words but not with others), and directionality (the extent to which words in a collocation are asymmetrically attracted to each other). In this paper, we focus on t-scores and

mutual information (MI) as there is psycholinguistic evidence that MI significantly relates to native speakers' recognition and production of formulaic sequences, while t-scores reflect L2 speakers' collocation processing (e.g., Ellis et al., 2008). Thus, using t-scores and MI allows us to assess the nativelikeness of L2 collocation use, respectively. Furthermore, most L2 collocation research has extensively focused on t-scores and MI (Gablasova, Brezina, & McEnery, 2017). By using the same indices of collocation, we ensure the comparability of the current study and its findings.

T-scores highlight the use of high-frequency collocations. These typically consist of high-frequency words which may have multiple potential partner words (e.g., function words). MI indexes the *mutual exclusivity* of word associations, weighing combinations of less frequent, more abstract, and more complex words which likely have fewer partner words. Collocations with higher MI scores entail greater coherence, more distinctive meaning and clearer discourse functions due to the limited number of partner words (for details of the calculation procedure for t and MI scores and their examples, see the Method section below).

To date, there is some research evidence that the MI scores of the collocations speakers use are weakly but significantly associated with global L2 *written* proficiency (e.g., Kyle & Crossley, 2016 for r = .10-.20 in TOEFL Writing; Garner, Crossley, & Kyle, 2019 for r = .20-.30 in CEFR Writing). Scholars have begun to investigate the relationship between collocation and L2 *speaking* proficiency assessment and development. For example, Kyle and Crossley (2015) examined how both single-word and collocation measures related to holistic proficiency scorings on TOEFL iBT Speaking tasks. Results of the statistical analyses indicated that trigram frequency (MI) explained the largest amount of variance in L2 speaking proficiency (r = .59; see also Eguchi & Kyle, 2020). Though revealing, one critique regarding this line of research is that the findings have exclusively relied on trained raters' judgements of general proficiency test performance. In such high-stake assessment settings, raters receive extensive training in order to score each L2 sample consistently and reliably with reference to pre-existing and detailed descriptors. However, as pointed out by Koizumi (2012), trained raters may pay attention to certain lexical factors (e.g., collocation) simply because they are explicitly asked to do so. This raises the question of whether collocation use impacts L1 listeners' *intuitive* judgements of the comprehensibility and appropriateness of L2 speech.

Intuitive Judgements of L2 Speech

In L2 speech research, many scholars have emphasized the importance of probing how listeners *intuitively* comprehend foreign-accented speech[2] without any reference to predetermined descriptors. They are also interested in how such L2 speech judgements vary according to rater background (e.g., monolinguals vs. bilinguals; linguists vs. non-linguists; musicians vs. non-musicians) (for an overview, see Derwing & Munro, 2015). Understanding the behaviors underlying intuitive judgements of this kind is crucial, arguably because such intuitions ultimately matter in real-world L2 communication (Levis, 2018). To date, much scholarly attention has been given to the concept of comprehensibility, defined as "how easily a listener can understand L2 speech" (Isaacs, Trofimovich, & Foote, 2017). Upon hearing a sample of spontaneous L2 speech, raters are asked to assess it in terms of ease of understanding on a 9-point scale.[3] According to the existing literature, L2 comprehensibility judgements can be greatly influenced by a range of phonological factors, such as segmental details (e.g., Suzukida & Saito, 2019), adequate prosody (e.g., Kang, Rubin, & Pickering, 2010), and temporal fluency (e.g., Suzuki & Kormos, 2019; for a meta-analytic review, see Saito, forthcoming-a). However, a growing number of studies have delved into how L2 comprehensibility could be influenced by other linguistic features, such as lexicogrammar appropriateness, fluency and sophistication. Little is known about which vocabulary factors make certain L2 speech samples easier to understand despite foreign accentedness, and which vocabulary factors are crucial to successful L2 comprehensibility development.

In previous L2 comprehensibility research, raters listened to and assessed audio recordings. In contrast, Saito, Webb, Trofimovich, and Isaacs (2016) proposed a different methodological paradigm to examine the lexical profiles of comprehensible L2 speech, where raters read and evaluate the comprehensibility of speech transcripts rather than audio files (for a similar methodology, see Crossley, Salsbury, & McNamara, 2015 for "collocational accuracy";

[2] In this paper, the main focus concerns the examination of how raters process lexical characteristics of foreign-accented speech. In L2 speech literature, speech can be accented at various levels including not only phonology, but also lexicogrammar (e.g., inaccurate vocabulary and morphological markers, wrong word order, L1 insertion). Thus, the term, foreign-accented speech, is used throughout the paper while introducing a range of topics and existing studies related to vocabulary aspects of L2 speech.

[3] Although some studies have recently begun to use a moving slider to rate L2 comprehensibility as a continuous phenomenon (recorded on a 1000-point scale), such findings appear to be comparable to those of 9-point scale (for a review on methodological variation in L2 comprehensibility research, Saito, Trofimovich, & Isaacs, 2017).

Foster & Wigglesworth, 2016 for "weighted accuracy"). Using this method, it has been shown that raters attend to the appropriate and fluent use of diverse vocabulary items during L2 comprehensibility judgements (Saito et al., 2016); that the raters' behaviors could vary according to their backgrounds (e.g., Saito, Trofimovich, Isaacs, & Webb, 2016), and that the comprehensibility of spoken L2 vocabulary continues to develop as long as L2 learners continue to practice the target language in classroom and naturalistic settings (Saito, 2015, 2019, forthcoming-b).

More recently, Saito (2020) explored the role of collocation in L2 comprehensibility in the context of 85 Japanese learners of English with varied proficiency levels. According to the results, the collocation factor (MI scores) explained a medium to large amount of the variance in L2 comprehensibility ratings (40-50%), confirming the generalizability of Kyle and Crossley's (2015) earlier findings in TOEFL iBT Speaking. These findings bring to light multiple research avenues to explore the complex relationship between multiword factors and raters' intuitive judgements of L2 comprehensibility. In this paper, we extend the scope of this topic by focusing on the two predictor variables—task effects (Study 1) and rater effects (Study 2)—from both cross-sectional and longitudinal perspectives (Study 3).

Study 1: Task Effects

In the precursor research (Saito, 2020), participants were given an eight-frame picture cartoon and asked to describe the events that occur therein (for details, see Derwing & Munro, 2013). According to Skehan's (1998) model of task complexity, this format (i.e., picture description) can be considered as more formal and structured, since speakers are not given much freedom to conceptualize the productive content of the task. They are rather asked to explain information which is already known to the listener and is not personal. As such, picture description tasks are thought to induce speakers to prioritize producing accurate language without much stress on conceptualization (i.e., what to say). This allows raters to focus on how speakers accurately describe the sequence of events without needing to evaluate the content, creativity, and organization of each speaker's performance. In this context, it is possible that individual differences in speakers' targetlike and accurate use of collocations may be salient and thus serve as a good predictor of L2 oral proficiency.

To test the presence/absence of the task effects, the current study re-examined the relationship between collocation and rater behavior using speech samples elicited from both structured (picture description) and free (oral interview) speaking tasks. The latter format was believed to allow L2 speakers to discuss a familiar topic and elaborate on their own ideas with some level of freedom, creativity and organization of the content. According to Skehan's (1998) model, oral interview tasks could be considered as informal, personal, less structured, as they promote the ability of individual speakers to conceptualize and produce speech. In the process, they may risk using more complex and sophisticated words that they may not have full control over at the expense of accurate and controlled production (see Skehan & Foster, 1999 for empirical evidence).

Method

Participants

A total of four native speakers of English recruited from a university in the USA participated as raters (M_{age} = 24.8 years). Each rater was individually interviewed to confirm that none had any experience in linguistics nor in teaching English as a foreign language. Their backgrounds are similar as none of them reported any prior training in linguistics and they can all be considered linguistically naïve under Isaacs and Thomson's (2013) definition. In previous comprehensibility research, the number of raters has substantially varied. Instead of recruiting multiple raters with diverse backgrounds which inevitably affect L2 comprehensibility judgements, efforts were made to recruit a small number of raters with relatively homogeneous backgrounds. We found this to be reliable as their scores are relatively consistent (see below).

Speech Materials

The same dataset from Saito (2020) was used for the comprehensibility judgments. This dataset consisted of speech samples from 85 Japanese speakers of English with different levels of L2 proficiency and immersion experience. As such, the data was assumed to provide a general index of the collocation effects in L2 proficiency (without the findings being limited to either

beginner or advanced L2 proficient users). Sixty-one of these speakers completed not only the picture description, but also the oral interview task, and served as the main data in the current study. Twenty-seven participants were university students in Tokyo, Japan who had no experience overseas. The remaining 34 participants were mid- to long-term residents in the USA ($M_{\textit{length of residence}}$ = 15.3 years; *Range* = 1-28 years).

Each speech recording session took place individually with a researcher. All the speakers engaged in the picture description and oral interview tasks in this order. For the picture description task, participants were asked to describe an eight-frame cartoon picture depicting an accidental exchange of suitcases on a busy street. For the oral interview task, participants were prompted to speak more freely about a personal and familiar topic. Following the procedures of the IELTS long-turn speaking task, participants received a card which described the assigned topic (i.e., *What was the hardest and toughest change in your life?*). This came with a set of possible discussion points for participants to extend and elaborate on their speech (e.g., *Why was it so challenging?*). The participants first spent one minute familiarizing themselves with the content of the task. Then, they spoke for two minutes. Finally, the researcher asked one follow-up question in response to the content of their speech (e.g., *What did you learn from the experience?*) (for the materials used in the study, see **Supporting Information-A**).

For transcription, two research assistants participated. Both were Japanese native speakers with high-level L2 English proficiency and an extensive amount of experience on L2 speech analyses of this kind. They separately transcribed the same 10 samples (not included in the main dataset) as a part of their training then compared their transcriptions for consistency. They resolved any disagreements before continuing. The first and second coders transcribed about 50% of the dataset for the picture description and oral interview tasks, respectively. The length of speech was shorter in the former (*M* = 228.3 words; *SD* = 96.1 words; *Range* = 95-424 words) than the latter task (*M* = 424.8 words; *SD* = 198.7 words; *Range* = 189-939 words).

Comprehensibility Judgements

Comprehensibility is one of the most extensively researched topics in L2 speech research (Derwing & Munro, 2015). Comprehensibility is typically operationalized via raters' intuitive judgements of ease of understanding. While many previous studies have been concerned with the

role of phonological errors in perceived comprehensibility (e.g., Kang et al., 2010), the current study aimed to explore the relationship between vocabulary (collocation) use and L2 comprehensibility. As initially proposed and developed in Saito et al. (2016a), and later extended in Saito (2019, 2020), the raters read transcripts instead of listening to audio samples. Using this framework, we intended to look at the raters' reactions to the lexical characteristics of speech while controlling for phonological factors.[4]

All rating sessions were conducted individually with a trained research assistant. The raters were first explained the objective of the study: to explore linguistically naïve raters' intuitive judgements of comprehensibility (ease of understanding) while reading transcribed L2 speech samples. In order to tap into such intuitions, comprehensibility was given a simple definition that did not mention language accuracy, vocabulary, or collocation use (summarized in Figure 1). This procedure is essentially different from high-stakes L2 proficiency assessments, where linguistically experienced raters receive much training on specific evaluation criteria in accordance with detailed rubrics and descriptors (e.g., IELTS).

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.

Figure 1 Training Scripts and Onscreen Labels for Comprehensibility Ratings

All transcripts were displayed to the raters in a randomized order on a computer screen using a MATLAB-based program. The raters read and rated the comprehensibility of each transcript using a moving slider. Each end of the continuum featured a smiling or frowning face

[4] The methodology here (raters' assessment of transcripts) has been widely used in L2 vocabulary research (e.g., Crossley et al., 2015), L2 speech research (e.g., Foster & Wigglesworth, 2016), and in age-related research (e.g., Patkowski, 1990).

to clearly indicate each end of the 0 to 1000-point continuum (*0 = difficult to understand, 1000 = easy to understand*; see Figure 1). To reduce fatigue and any extraneous distractions, the entire session (90+ minutes in total) was administered across two days (Day 1 for 61 picture descriptions and Day 2 for 61 oral interviews).

Interrater agreement was calculated using Cronbach's alpha. As found in Saito (2020), the four raters demonstrated relatively high inter-rater agreement for the picture description ($\alpha = .879$) and oral interview tasks ($\alpha = .883$). Therefore, the four raters' comprehensibility scores were averaged to generate a single score for each sample under each task condition.

Collocation Measures

Following Saito (2020), the collocation use of L2 speech was analyzed via bigram and trigram association measures—i.e., t- and mutual information (MI) scores—using the Tool for the Automatic Analysis of Lexical Sophistication 2.0 (TAALES) (Kyle & Crossley, 2015). Since the speakers and raters used General American English, we chose the spoken dimension of the Corpus of Contemporary American English (Davies, 2009) as the reference corpus.

In TAALES, random co-occurrences of words were first calculated by dividing the number of any possible combinations within a fixed window size of five words by the total number of tokens in the reference corpus. To generate t-scores, the difference between raw frequency and random co-occurrence frequency were divided by the square root of the raw frequency. To generate MI scores, the frequency of collocations is divided by the frequency of random co-occurrences of the words, and then logarithmized. Whereas t-scores are thought to index how much a sample is made up of combinations of relatively frequent words (function words in particular), which are likely to have many other partner words (i.e., high-frequency associations), MI scores are thought to reflect the extent to which a sample features combinations of mutually exclusive words that do not have many other partner words (i.e., low-frequency associations). For examples of bigrams and trigrams, see **Supporting Information-B**.

Results

Constructs of Comprehensibility and Collocation Measures

The results of normality tests (a one-sample Kolmogorov-Smirnov test) indicated that the comprehensibility scores were not significantly different from a normal distribution in both task contexts ($p > .05$). According to the results of independent sample *t*-tests, the averaged comprehensibility scores did not significantly differ between the picture description (M =573; SD = 174; *Range* = 262-910) and oral interview tasks (M =513; SD = 217; *Range* = 172-861), t = 1.675, p = .096, d = 0.34. As for the collocation measures, a series of Kolmogorov-Smirnov tests found the bigram t-scores in the picture description task to be positively skewed (p = .015) and the trigram t- and MI scores in the oral interview task to be negatively skewed (p = .007, and .016). After these scores were transformed using the log10 function, they were shown to follow a normal distribution ($p > .05$). To facilitate the interpretability of the data, the directions of all the factor scores were kept consistent (larger values indicating stronger associations).

Relationship Between Comprehensibility and Collocation

To examine the role of collocation in L2 comprehensibility judgements, a set of Pearson correlation analyses were performed with alpha set to .01 (Bonferroni corrected). As summarized in Table 1, both bigram and trigram MI scores demonstrated significant, moderate-to-strong associations with comprehensibility in the picture description task (r = .356 to .713). Only bigram t-scores demonstrated significant correlations with comprehensibility in the oral interview task (r = .343).

Table 1

Summary of Simple and Partial Correlations Between Collocation and Comprehensibility

	Comprehensibility (Picture Description)		Comprehensibility (Oral Interview)	
	r	*p*	*r*	*p*
Bigram				
t scores	.356	.005*	.343	.006*
MI scores	.713	< .001*	.237	.066
Trigram				
t scores	.431	.001*	.243	.059
MI scores	.488	< .001*	.246	.096

* *for* $p < .01$

In the precursor study (Saito, 2020), it was suggested that text length could be related to the collocation-proficiency link (i.e., longer speech samples tend to feature more likelihood of targetlike collocation use). Indeed, significant correlations between comprehensibility and text length were found in both the picture description (r = .356, p = .005) and oral interview (r = .643, p < .001) tasks. To investigate the relative weights of collocation, text length and comprehensibility, stepwise multiple regression analyses were performed with comprehensibility scores as a dependent variable relative to five predictor variables (bigram and trigram t- and MI scores, the number of words per sample). As summarized in Table 2, the collocation factor (bigram MI scores) accounted for 50.8% of the variance in the comprehensibility judgements in the picture description task. Collocation effects were much weaker in the oral interview task (accounting for 8.6% of the variance). There was no clear instance of strong multicollinearity in any model (Variance Inflation Factor < 1.412).

Table 2

Summary of Multiple Regression Analyses of the Relationship Between Collocation and Comprehensibility

	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility	Bigram MI scores	.508	.508	60.929	< .001
(Picture Description)	Text length	.624	.116	48.129	< .001
Comprehensibility	Text length	.413	.413	41.560	< .001
(Oral Interview)	Bigram MI scores	.499	.086	28.875	< .001

Discussion

The results of the precursor research showed that L2 speakers' collocation use (operationalized as t- and MI scores) was a primary determinant of native raters' intuitive comprehensibility judgements (ease of understanding) (Saito, 2020). The primary aim of the current study was to examine the generalizability of collocation effects across different task conditions (i.e., picture description and interview task). While the findings showed clear collocation effects in the picture description task (accounting for 50.8% of the variances), the

predictive power of collocation was smaller in the oral interview task (explaining 8.6% of the variances).

As predicted earlier, this could be due to the nature of the tasks themselves. From raters' perspectives, the picture description remained the same across all participants. Once the raters knew the story, they could focus more on the linguistic characteristics than semantic content/details of their speech. In contrast, the content of the interviews inevitably varied according to each speaker. Thus, the raters had to pay more attention to the content of their speech (for similar discussion on the role of task structure in intuitive L2 speech judgements, see Crowther, Trofimovich, Isaacs, & Saito, 2016; Crowther, Trofimovich, Saito, & Isaacs, 2018; Derwing, Rossiter, Munro, and Thomson, 2004).

As stated in Skehan's (1998) task complexity framework, the picture description task has been found to induce speakers to focus on processing already-given information whose structure is well-known (the eight-frame cartoon picture). Thus, speakers may prioritize accuracy and fluency in conveying their message (see Skehan & Foster, 1999 for empirical evidence). In light of these phenomena (speech being more accurate and fluent in picture description than interview), raters may pay much attention to assessing linguistic accuracy and fluency, while simultaneously understanding what the speaker intends to say (Derwing & Munro, 2015).

Since there is emerging evidence that collocation is linked to accuracy (Saito, 2020) and fluency (Tavakoli & Uchihara, 2020) aspects of L2 speech, it is understandable that collocation use could be most clearly and strongly predictive of L2 oral proficiency performance and assessment when tasks are well-structured with known content. Comparatively, the oral interview task was assumed to induce speakers and assessors to focus on elaborating on familiar and personal topics (*What was the hardest challenge in your life?*) with ample room for conceptualization (what to say). Under this task condition, speakers and assessors are likely to prioritize content rather than form. Since the task format inevitably results in more diverse, unpredictable word choice, speakers and assessors are likely to rely on the amount of information (i.e., text length) as a primary cue, and collocation use as a secondary cue.

Study 2: Rater Effects

It is noteworthy that all the findings so far have been exclusively based on the intuitive judgements of native English raters. In Study 2, therefore, we explored the role of rater background (L1 vs. L2 raters) in L2 comprehensibility judgements. Although there has been ample research examining the mechanisms underlying various raters' comprehensibility judgements (for an overview, see Derwing & Munro, 2015), the existing literature has exclusively relied on intuitive evaluation of *audio* samples. To our knowledge, our study is the first attempt to pursue this topic in the context of intuitive judgements of *transcript* samples. In what follows, we first briefly review a set of studies on the relationship between rater backgrounds and *audio* L2 speech assessment. Accordingly, we introduce some studies which have provided some insights on how different types of raters evaluate the use of collocation in L2 speaking *and* writing.

Within the L2 speech assessment literature, there is ample evidence that biographical background affects rater behavior. For example, certain native raters have been shown to evaluate familiar foreign accents more leniently because of their language experience (Winke, Gass, & Myford, 2013), linguistics training (Isaacs & Thomson, 2013), bilingual experience backgrounds (Saito & Shintani, 2016), and/or professional ESL/EFL teaching experience (Saito, Trofimovich et al., 2016; for a meta-analysis, see Saito, forthcoming-a).

Given that English is used as a lingua franca in today's globalized world, an increasing amount of attention has been given to the mutual comprehensibility of L2 English speakers (Pennycook, 2017). Though limited, the findings have thus far been mixed. Some studies have shown that L1 and L2 raters assess foreign-accented speech similarly (Crowther, Trofimovich, & Isaacs, 2016; Munro, Derwing, & Morton, 2006). Other studies have demonstrated that L1 and L2 raters' evaluation of accented speech behaviors could be substantially different (Foote & Trofimovich, 2018; Ludwig & Mora, 2017). It could be argued that the variation in results is caused by the large degrees of individual variation in L1-L2 distance, L2 proficiency, experience, attitude, and familiarity with particular foreign-accents, and therefore that L2 users cannot be treated as a single group.

For example, some L2 raters show much difficulty understanding other foreign-accented speech (resulting in stricter L2 comprehensibility judgements) due to the lack of enough

conversation experience with a wide range of foreign language speakers. In contrast, certain L2 raters may be capable of paying attention to both the form and meaning aspects of language and provide more lenient comprehensibility judgements. The perceptual representations of these raters can flexibly accommodate and decode a wide range of novel voices, and by extension foreign-accented speech (see Witteman, Weber, & McQueen, 2013 for a comprehensive review on the psycholinguistic phenomenon of perceptual adaptation). Such lenient raters regularly use the target language with different types of interlocutors with a clear appreciation of and positive attitude towards foreign-accented speech (Saito, Tran, Suzukida, Sun, Magne, & Ilkan, 2019).

It is noteworthy that all the aforementioned literature on rater effects has been exclusively concerned with audio samples. However, surprisingly little is known about how L1 and L2 speakers and assessors differentially process collocation during speaking, writing, and assessment tasks. Some empirical evidence has indicated that L2 users overly rely on a combination of high-frequency words during writing tasks (resulting in greater t-scores) (Durrant & Schmitt, 2009) and demonstrate less sensitivity to the way patterns of low-frequency words are used together (indexed as MI scores) (Ellis et al., 2008). In the current study, we would like to further pursue whether and to what degree L1 and L2 raters differentially attend to collocation use when making intuitive judgements of L2 comprehensibility. Corresponding to two different learning contexts (learning L2 English through foreign language instruction vs. through immersion), we recruited two different groups of L2 raters— (a) 34 Chinese English-as-a-Foreign-Language (EFL) students (without any experience abroad) and (b) 28 Chinese English-as-a-Second-Language (ESL) students in the UK. We then compared their rating behaviors with those of five native speaking raters.

Method

Participants

The objective of Study 2 was to analyze the L2 comprehensibility judgement patterns of two groups of L2 raters with diverse bilingual experience profiles (i.e., experienced vs. inexperienced L2 raters).[5] Rater background was operationalized via the presence of immersion

[5] Previous L2 speech literature widely creates a distinction between “Experienced” and “Inexperienced” based on the presence and length of naturalistic, intensive immersion experience (e.g., study-abroad), when researchers are

experience (rather than any professional speech assessment experience). The immersion experience variable was chosen for the following reasons.

First, it is easy to quantify whether L2 raters have ever had any immersion experience in English-speaking environments. Secondly, the quantity and quality of L2 learning is substantially different between immersion vs. non-immersion (i.e., foreign language contexts). In the former case, L2 learning takes places in various social settings where learners process language for both meaning and form (e.g., social conversations, ESL classrooms, content-based classes). In the latter case, while some students do have opportunities to participate in conversation-based English classes or/and subject matter education in English, many EFL classrooms are form-oriented (Nishino & Watanabe, 2008). More importantly, EFL learners' access to a target language (either through form or meaning-oriented instruction) is severely limited outside classrooms (see Muñoz, 2014 for further discussion on the contextual differences between immersion vs. foreign language settings). Third, the presence/absence of meaning-oriented, conversational experience (characteristic of immersion) has been found to affect L2 listeners' behaviors: L2 raters who have used their target language on a daily basis likely have more flexible mental representations, providing higher and more lenient comprehensibility scores to foreign-accented speech (Saito et al., 2019).

Efforts were made to maximize between-group distinction (Chinese learners of English in English-as-a-Second-Language [ESL] settings vs. EFL settings) as much as possible and minimize within-group variation (the homogeneity within each group condition). A total of 28 Chinese postgraduate students in London, UK were recruited as the experienced ESL raters. They were relatively homogeneous in terms of the quantity and quality of L2 immersion experience. They had approximately one year of study abroad experience ($M_{\text{length of immersion}}$ = 9.6 months, *SD* = 3.4, *Range* = 6-24 months) and similar levels of general L2 English proficiency (M_{IELTS} = 7.4 out of 9 points, *SD* = 0.3, *Range* = 7-8 points). The inexperienced L2 raters (*n* = 34) were carefully recruited from a university in China by screening for any experience travelling to English-speaking countries (i.e., immersion experience). All of them had relatively high-levels of

interested in group comparison. For example, more experienced L2 users tend to demonstrate more advanced phonological proficiency (e.g., Trofimovich & Baker, 2006); more experienced L2 users better attend to and understand various types of foreign-accented speech (e.g., Saito et al., 2019); and L2 speech development is relatively limited and subject to a great deal of individual variation among inexperienced L2 users in classroom settings (e.g., Mora & Valls-Ferrer, 2012). To be consistent with the standard definitions within the field, we used "Experienced" (with immersion) and "Inexperienced" (without immersion) in the current study.

L2 English proficiency (M_{IELTS} = 7.2 points, *SD* = 0.2, *Range* = 7-8 points), but without any experience of living or studying abroad. They also reported that their classroom experience was mainly form-oriented with little experience in conversation- and content-based classes at the time of the project.

As a result, the biographical backgrounds of the two rater groups (ESL vs. EFL raters) were different in terms of the presence/absence of L2 immersion experience, but generally comparable in many other respects, such as L2 English proficiency, age of learning, and familiarity with Japanese accented English. For a summary of rater backgrounds, see **Supporting Information-C**.

For the purpose of comparison, a total of five native speakers of English (M_{age} = 25.4 years) were recruited at an English-speaking university in Montreal, Canada. Similar to Study 1, they were naïve raters (no experience in linguistics training and EFL/ESL teaching) who communicated mainly in English (95+% per day).

Speech Materials

A total of 50 samples were randomly selected from the 85 picture descriptions used in the precursor study (Saito, 2020). The length of each speech sample varied from 89 to 221 words (*M* = 123.6 words; *SD* = 34.3 words). We consider the length of the samples sufficient as the range (89-221) was comparable to Koizumi and In'nami's (2012) guidelines for robust vocabulary analyses (i.e., 100 words).

Comprehensibility Judgements

The raters comprised 34 EFL raters, 28 ESL raters and 5 native speaking raters. Following the same procedure in Study 1, they read 50 transcripts (displayed on a computer screen in a randomized order via a MATLAB-based software), and rated them for comprehensibility using a moving slider (recorded on a 1000-point scale). All rating sessions took place individually in the presence of a researcher who provided a brief explanation of the project, described the rating construct, and explained the procedures. After the raters practiced with three samples (not included in the main dataset), they proceeded to assess the main dataset

(n = 50 picture descriptions). Similar to Study 1, the raters showed relatively high Cronbach alpha in accordance with their group category: α = .899 for EFL Group, .946 for ESL Group, and .905 for Native Baselines. Thus, the raters' comprehensibility scores were averaged across the raters and group conditions, respectively.

Collocation Measures

The same four collocation measures (bigram t- and MI scores, trigram t- and MI scores) were employed to index the collocation quality of each transcript sample.

Results

Constructs of Comprehensibility and Collocation

According to the results of normality tests (Kolmogorov-Smirnov), all comprehensibility scores appeared to follow a normal distribution, $p > .05$. The results of one-way ANOVAs demonstrated that the three groups assigned significantly different comprehensibility scores to the 50 picture description samples, $F(2, 147) = 14.304$, $p < .001$, $\eta^2 = .163$. Post-hoc multiple comparison analyses showed that the ESL raters' comprehensibility scores (M = 567, SD = 103, *Range* = 339-796) were significantly greater (and thus more lenient) than the EFL raters (M = 497, SD = 133, *Range* = 273-794) with moderate effects (d = 0.67); and the Native Baselines (M = 440, SD = 166, *Range* = 238-774) with large effects (d = 1.23). However, the EFL and Native Baselines appeared to be comparable ($p > .05$). In terms of collocation measures, the results of Kolmogorov-Smirnov tests found only Bigram MI scores to be positively skewed (p = .006). After transforming the values via the log10 function, their distribution pattern became normal (p = .067). To facilitate the interpretability of the findings, the directionality was set positive (larger values indicating stronger associations).

Relationship Between Comprehensibility and Collocation

The role of collocation in L2 comprehensibility judgments was examined via simple Pearson correlation analyses with alpha set to .01 (Bonferroni corrections). As shown in Table 3, bigram MI scores demonstrated significant associations with L2 comprehensibility for all three groups of raters (ESL, EFL, Native Baselines), suggesting that both L1 and L2 raters used collocation information similarly during their intuitive judgments of L2 speech.

Table 3 *Summary of Simple Correlations Between Collocation and Comprehensibility*

	Comprehensibility (ESL raters)		Comprehensibility (EFL raters)		Comprehensibility (Native Baselines)	
	r	*p*	*r*	*p*	*r*	*p*
Bigram						
t scores	.315	.026	.328	.128	.218	.128
MI scores	.598	< .001*	.552	< .001*	.600	< .001*
Trigram						
t scores	.126	.381	.230	.108	.184	.200
MI scores	.068	.640	.130	.368	.152	.293

* *for* $p < .01$

Interestingly, we also found that the raters' comprehensibility scores were significantly associated with the length of speech samples. Such length effects were clearly observed among the EFL raters ($r = .648$, $p < .001$) in contrast with the ESL raters ($r = .444$, $p = .001$) and the native baselines ($r = .573$, $p < .001$). A set of stepwise multiple regression analyses were performed to further examine how the three groups of raters—ESL, EFL and Native Baselines—differentially used collocation and text length information to assess L2 comprehensibility. The models featured L2 comprehensibility scores as a dependent variable and five predictor variables (bigram t and MI scores, trigram t and MI scores, text length). According to the results, summarized in Table 4, both ESL and native raters showed very similar patterns—i.e., using collocation (bigram MI) as a primary cue (accounting for 35.8-36% of the variances) and text length as a secondary cue (6.2-15%). However, the EFL raters' comprehensibility judgements were mainly determined by the text length factor (42.0%) followed up by the collocation factor (bigram MI) (12%). There was no indication of strong multicollinearity (VIF < 1.23).

Table 4 *Summary of Multiple Regression Analyses of the Relationship Between Collocation and Comprehensibility*

	Predictor variables	*Adjusted* R^2	R^2 *change*	*F*	*p*
Comprehensibility (ESL Raters)	Bigram MI scores	.358	.358	26.770	< .001
	Text length	.420	.062	17.045	< .001
Comprehensibility (EFL Raters)	Text length	.420	.420	34.711	< .001
	Bigram MI scores	.540	.120	12.271	< .001
Comprehensibility (Native Baselines)	Bigram MI scores	.360	.360	26.953	< .001
	Text length	.509	.150	24.399	< .001

Discussion

Much scholarly attention has been directed towards examining the diverse rating behaviors of L1 and L2 raters when assessing the comprehensibility of foreign-accented speech (e.g., Ludwig & Mora, 2017 vs. Crowther et al., 2016; for a meta-analysis, Saito, forthcoming-a). In the context of 50 picture description samples produced by Japanese learners of English, Study 2 examined how the three different groups of raters—(a) experienced Chinese users of English in the UK (ESL raters), (b) inexperienced Chinese users of English in China (EFL raters), and (c) native speakers of English (Native Baselines)—differentially rely on collocation information during L2 comprehensibility judgements.

The experienced L2 users (ESL raters) assigned higher scores than the other groups, which may indicate a more lenient attitude towards evaluating other foreign-accented speech. They also seemingly adopted the same strategy as the native baseline raters in their evaluations—prioritizing the frequency of mutually exclusive combinations of words (i.e., mutual information) over other lexical factors (e.g., text length). The results here line up with previous evidence showing that (a) L2 raters are more adaptable, flexible, and lenient as a result of more experience with and exposure to different types of foreign accents (Kang & Lu, 2019); and (b) that MI scores serve as a primary correlate of L2 oral/speaking proficiency (Kyle & Crossley, 2015; Eguchi & Kyle, 2020).

By contrast, our results indicate that inexperienced L2 users (EFL raters) made their comprehensibility judgements by focusing on the quantity (text length) rather than quality (collocation) of L2 speech. This lack of sensitivity to collocation among L2 users has been reported in the dimensions of writing (Garner et al., 2019), word recognition and production

(Ellis et al., 2008), acceptability judgments (Wolter & Gyllstad, 2013), and speaking (Kyle & Crossley, 2015). As pointed out by many scholars, this could be due to the fact that the majority of EFL learners do not have sufficient opportunities to access authentic input in their L2 English classrooms (see Biber et al., 2004; Boers, Dang, & Strong, 2017 for the analyses of collocation use in ESL and EFL textbooks).

Although the current study used a cross-sectional dataset (i.e., comparing the rater behaviors of experienced and inexperienced L2 users vs. native listeners), the findings shed light on how L2 users develop, revise and elaborate their ability to comprehend other L2 speakers. To grasp what other L2 users say, L2 users may initially prioritize the amount of information delivered without paying attention to any collocational aspects of use (i.e., the mutual exclusivity of word strings). This is arguably because these inexperienced L2 users cannot afford to allocate sufficient cognitive resources to such sophisticated vocabulary analyses, especially when they focus on using language for meaning rather than form (Skehan, 1998); and/or because they have yet to develop sufficiently robust collocation knowledge (Boers et al., 2017). With increasing L2 experience, L2 users can start to encode language into multiword units, and store frequently occurring combinations as chunks which can be accessed more accurately, fluently, and automatically (Ellis et al., 2008).

These tentative suggestions are reminiscent of the psycholinguistic phenomenon of perceptual adaptation (Witteman et al., 2013). Perceptual adaptation occurs when listeners revise their existing perception systems upon engaging in intensive, systematic and repeated exposure to novel sounds, words and sentences. For example, there is empirical evidence that brief listening experience helps L1 listeners adjust to unfamiliar acoustic signals that they have never heard before, and integrate them into their phonetic systems such as acoustically manipulated sounds (Norris, McQueen, & Cutler, 2003), or foreign-accented speech (Bradlow & Bent, 2008). The findings of the current study suggest that perceptual adaptation could also occur in L2 listeners', facilitating understanding of foreign-accented speech after a certain amount of immersion experience (e.g., one year of study abroad in the current study) has taken place. They also provide insight into the mechanisms of perceptual adaptation on a micro level—i.e., obtaining more robust sensitivity to the association of low-frequent, exclusive multiword networks (mutual information).

Here, it is important to stress that the current investigation adopted an exploratory methodology by asking raters to read transcripts (rather than listening to audio samples) in order to factor out the influence of phonology on the relationship between collocation and comprehensibility judgements. Although many studies in Second Language Acquisition and psychology have extensively examined the role of rater experience in L2 speech assessment, such literature has been exclusively concerned with listening rather than reading (cf. Saito, forthcoming-a). Due to the lack of literature on unique methodology that we used in the current study (reading rather than listening), we discussed our findings in line with the relevant theories of human speech perception in psychology (perceptual adaptation). Our assumption is that the way raters evaluate L2 speech shares similar mechanisms, even though the modality is different (reading vs. listening) and the domain is different (vocabulary vs. phonology). However, we acknowledge that such assumption needs further empirical investigation which will scrutinize whether, to what degree and how raters differently (or similarly) process lexical information in L2 speech when reading transcripts vs. listening to audio samples (e.g., Saito, Webb et al., 2016a vs. Saito, Webb et al., 2016b).

Study 3: Collocation, Comprehensibility, and Longitudinal L2 Speech Development

Recently, scholars have begun to show that comprehensibility can serve as a developmental index of L2 speech learning. With sustained use of the L2, learners can continue to enhance the comprehensibility of their speech even while remaining foreign-accented (Derwing & Munro, 2013). Study 3 adopts a longitudinal perspective to further examine the *causal* effect of collocation use on the development of L2 comprehensibility. Following the assumption that collocation knowledge and use could drive L2 speech assessment *and* development, we made two predictions. First, if L2 learners practiced the target language over time, they would improve their L2 speech, especially in terms of comprehensibility. Secondly, such enhanced L2 comprehensibility could be linked to the development of their L2 collocation use, and vice versa.

Method

Participants

Speakers. As a part of a larger project, we invited interested L1 Japanese university students in Tokyo, Japan to participate in a semester-long language exchange project. All participants were paired with native speakers of English who were enrolled in college-level schools in the USA. The pairs used a video-conferencing tool installed on their computers to engage in 10 individual meetings over the course of 10 weeks in accordance with their schedules and time differences between Japan and the USA. For each meeting, they chatted for one hour, 30 minute in Japanese and 30 minutes in English. As prompts for conversation, they were asked to bring two images that corresponded to a weekly theme (e.g., sports, pop culture). The current study focuses on a cohort of 28 Japanese students who completed the project in Spring 2014. All participants engaged in the same picture description task used in Studies 1 and 2 one week before and one week after the language-exchange project (28 speakers × 2 pre/post-tests = 56 samples). We have reported some parts of these results elsewhere (e.g., the participants' L2 English pronunciation and fluency performance and development; Saito & Akiyama, 2017).

The participants varied in terms of their general L2 English proficiency at the time of the project (measured via TOEIC) (M = 681.8 out of 990, SD = 165.5, *Range* = 350-900), indicating that their general proficiency spanned Basic (B1) and Proficient Users (C2) as per CEFR benchmarks. Similarly, their amount of immersion experience in English-speaking countries was varied substantially (M = 8.3 months, SD = 15.97, *Range* = 0-48 months). While they took a few hours of English classes per week at their university-level schools at the time of the project (M = 3.9 hours, SD = 2.3, *Range* = 1.5-6), they reported having limited opportunities to practice English outside of the classroom (M = 0.45 hours, SD = 1.05, *Range* = 0-3 hours).

Raters. A total of five native speakers of English were recruited in Montreal, Canada (M_{age} = 23.8 years) to rate the comprehensibility of the speech samples. They all reported English as their L1 and primary language of communication (100% per day). All of them were undergraduate and graduate students at an English-speaking university at the time of the project.

Like Studies 1 and 2, they were naïve raters as they lacked any experience in linguistics and ESL/EFL teaching.

Speech Materials

The participants engaged in a range of different speaking tasks one week before the outset of the project (T1) and one week after the end of the project (T2). As reported in our earlier study (Saito & Akiyama, 2017), the participants' English speech was elicited using a timed picture description task and analyzed for phonological accuracy and fluency. However, such short, fragmented speech samples (20-30 words per speaker) may not be adequate for robust vocabulary analyses. Thus, the samples used for phonological analyses and reported in the earlier study were not used for the current investigation.

As in Studies 1 and 2, the current study reports on participants' speech elicited from the picture description task (*M* = 115 words, *SD* = 38.9, *Range* = 75-226 words). To ensure comparability between participants' performance at T1 and T2, we used the same picture cartoon twice.[6] Like before, two coders first separately transcribed the same 10 out of the 56 samples (18% of the entire dataset: 28 speakers × T1/T2) to check for agreement. There were few disagreements between their transcriptions. After agreeing on transcription standards, they proceeded to transcribe 28 samples.

Comprehensibility Judgements

Following the same rating procedure in Studies 1 and 2, all the sessions took place individually with a trained research assistant. After the five raters familiarized themselves with the same picture cartoon featured in the speech samples, they received a brief explanation of the construct of comprehensibility and the rating procedure (for details, see Studies 1 and 2). They first practiced with three transcripts not included in the current dataset, and then rated the main dataset (56 transcripts). All the transcripts were displayed in a randomized order. The raters

[6] One obvious limitation of using the same test format twice concerns test-retest effects. However, research has shown that such test-retest effects may be minimum when test interval is spaced out (e.g., Derwing, Thomson, & Munro, 2006).

evaluated each transcript for comprehensibility on a 1000-point scale using a moving slider via a MATLAB-based program. Like Studies 1 and 2 (and Saito, 2020), the raters' agreement was relatively high, Cronbach α = .910. Thus, their scores were averaged to generate a single score for each speaker at each testing point (T1, T2).

Collocation Measures

The same four collocation measures (bigram t- and MI scores, trigram t- and MI scores) were employed to index the collocation quality of each transcript sample.

Results

According to the results of Kolmogorov-Smirnov tests, the pattern of comprehensibility and collocation scores were not significantly different from normal distribution ($p > .05$). In the main analysis, we first examined the predictive role of four different collocation measures (bigram, trigram*, t, MI* scores) in L2 comprehensibility judgements at T1 and T2 by conducting a set of partial correlations controlling for text length. As summarized in Table 5, the participants' MI scores were significantly correlated with their L2 comprehensibility scores at both T1 ($r = .517$, $p = .006$) and T2 ($r = .495$). Echoing the findings of Studies 1 and 2, collocation appeared to serve as a relatively strong predictor of L2 comprehensibility, even when the length of each transcript was factored out.

Table 5 *Summary of Partial Correlations Between Collocation and Comprehensibility With Text Length Factored Out*

	Comprehensibility (T1)		Comprehensibility (T2)	
	r	*p*	*r*	*p*
Bigram				
t scores	.124	.539	.341	.082
MI scores	.517	.006*	.495	.009*
Trigram				
t scores	.034	.866	.112	.577
MI scores	.033	.869	.164	.414

Note. * for $p < .01$.

To examine the collocation-proficiency link from a longitudinal perspective, we probed whether and to what degree participants' collocation proficiency and comprehensibility changed over time. On the whole, the results of paired-sample t tests did not find significant improvement in participants' comprehensibility over time ($t = -0.580$, $p = .567$, $d = 0.11$), suggesting that the extent to which they had benefitted from video-based conversation activities was subject to a great deal of individual variation. The participants were subsequently divided into two groups: (a) those who demonstrated positive change in L2 comprehensibility between T1 and T2 ($n = 15$); and (b) those whose comprehensibility levels did not show any improvement within the timeframe of the project ($n = 13$). According to the results of paired-sample t-tests (summarized in Table 6), participants in the Improvement Group significantly enhanced their bigram MI scores over the course of the project with medium effects ($p = .024$, $d = 0.63$).

Table 6

Summary of Participants' Comprehensibility and Collocation Performance Over Time (T1 → T2)

	T1		T2		Improvement (T1 → T2)		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
A. Improvement Group (*n* = 15)							
L2 Comprehensibility	395	113	512	101	-6.555	< .001*	1.09
Bigram (t scores)	48.23	22.98	54.14	20.11	-1.160	.255	0.41
Bigram (MI scores)	1.17	0.22	1.31	0.23	-2.538	.024*	0.63
Trigram (t scores)	14.70	7.64	18.53	5.30	-1.601	.132	0.50
Trigram (MI scores)	2.09	0.48	2.21	0.39	-1.138	.274	0.27
B. Non-Improvement Group (*n* = 13)							
L2 Comprehensibility	474	117	427	150	1.593	.137	0.34
Bigram (t scores)	55.13	16.43	54.92	21.13	.030	.977	0.01
Bigram (MI scores)	1.33	0.15	1.30	0.20	.398	.698	0.01
Trigram (t scores)	18.03	4.46	18.59	3.93	-0.281	.784	0.12
Trigram (MI scores)	2.15	0.32	2.19	0.23	-0.382	.709	0.12

Note. * for $p < .025$

Discussion

The goal of Study 3 was to test the predictive role of collocation use in native speakers' intuitive judgments of L2 comprehensibility from a longitudinal perspective. Our assumption was that improvements in L2 comprehensibility would be accompanied by improvements in collocation use. Focusing on a group of Japanese EFL learners who engaged in a series of video-based conversation sessions (1 hour per week × 10 weeks) with American conversation partners, we found evidence that L2 comprehensibility development could be related to, and by extension driven by, their ability to access and sensitize more frequent and targetlike use of mutually exclusive word combinations (collocations with higher MI scores). The results here confirm that native speakers' processing of second language speech could be significantly determined by collocation qualities (as they process first language speech) (Kyle & Crossley, 2015), and that L2 learners improve their speaking proficiency thanks to an increasing amount of attention and control over their collocation knowledge and use (Kim, Crossley, & Kyle, 2018).

Conclusion and Future Directions

In the field of L2 speech, there is a consensus that L2 oral proficiency should be evaluated in terms of comprehensibility rather than nativelikeness (Derwing & Munro, 2015). Some scholars have begun to examine which lexical factors make L2 speech more easily understood (via comprehensibility judgements), and which factors underlie comprehensibility improvement as a function of increased experience (comprehensibility development). Extending the precursor work (Saito, 2020), the current study reports the results of three different experiments. In discussing them, we aim to provide further empirical support for the role of collocation use in L2 comprehensibility judgements and speech development. Overall, the findings align with the emerging findings that collocation acts as a primary cue during L2 speaking assessment (Kyle & Crossley, 2015) and L2 speaking development (Kim et al., 2018).

More specifically, we found that raters rely substantially on collocations while making intuitive judgements, particularly when the lexical context of speech is relatively limited and predictable (structured picture description rather than freely-constructed oral interview; Crowther et al., 2016, 2018), as long as they have a sufficient amount of collocation knowledge (native speakers and experienced L2 users but not inexperienced L2 users) (Saito et al., 2019). These overall findings led us to make tentative conclusions regarding how L2 learners develop oral comprehension (better ability to understand others) and production (making themselves more easily understood) skills. Although L2 learners rely on different strategies to understand other foreign-accented speech, relying on the length of speech rather than details of speech, they may begin to analyze, sensitize and attend to the quality of multiword units with increasing amounts of L2 learning experience. Similarly, their production becomes more comprehensible and thus more advanced despite non-nativelike use of language as they refine their control over the use of multiword units which have very limited sets of collocates (collocation with higher MI scores).

In closing, we call for more studies which delve into the complex mechanisms underlying the development of L2 lexical networks. It is important to remember that the findings in Studies 1, 2, and 3 were significant when collocation was operationalized in terms of mutual information scores (the degree to which words pairings are mutually exclusive) rather than t-scores (how often word combinations co-occur). That is, L2 speech was judged to be, and became more comprehensible when it included more distinctive and coherent combinations of content words. As for single word units, Crossley and colleagues have

conducted a range of longitudinal studies showing that L2 learners' use of content words becomes more diverse, abstract, infrequent, and complex in nature with more conversational experience (e.g., Crossley & Skalicky, 2019). Following this line of thought, it is possible that more experienced and proficient L2 learners are able to strengthen their lexical networks in terms of both function words (which are relatively frequent) and content words (which are relatively infrequent). Our argument echoes a usage-based account of language acquisition, which assigns a central role to formulaic sequences in language analysis, processing, comprehension, production, and acquisition (Ellis, 2012). In some studies, the extent to which single word frequency measures were associated with the development of L2 oral proficiency remains unclear (Crossley et al., 2015; see also Crossley, Skalicky, Kyle, & Monteiro, 2019). However, if we focus on multiword units, we can predict that more infrequent combinations of content words will be produced in the later stages of acquisition—an assumption that our findings suggest and future studies should test (cf. Kim et al., 2018).

References

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

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*, 362-380.

Bradlow, A. & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106, 707–729.

Bybee, J. L., & Hopper, P. J. (Eds.). (2001). *Frequency and the emergence of linguistic structure* (Vol. 45). John Benjamins.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570-590.

Crossley, S. A., & Skalicky, S. (2019). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, *52*, 385-405.

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*, 721-744.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*, 160-182.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility?. *The Modern Language Journal*, *99*, 80-95.

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.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*, 163-185.

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

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). L2 fluency: Judgments on different tasks. *Language Learning, 54,* 655–679.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, *47*, 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*, 381-400.

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.

Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*, 253-278.

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

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

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*, 193-216.

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

Kang, O., & Lu, D. (March, 2019). World Englishes and learners' listening comprehension skills in the context of EFL classrooms. AAAL 2019, Atlanta, GA.

Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The 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.

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

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.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics, 24*, 578-596.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111-131.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204-238.

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

Patkowski, M. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics*, *11*, 73–89.

Pennycook, A. (2017). Translanguaging and semiotic assemblages. *International Journal of Multilingualism*, *14*, 1-14.

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning, 65*, 563-595.

Saito, K. (2019). To what extent does long-term foreign language education improve spoken second language lexical proficiency? *TESOL Quarterly, 53,* 82-107.

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

Saito, K. (forthcoming-a). What characterizes comprehensible and nativelike pronunciation among English-as-a-Second-Language speakers? Meta-analyses of phonological, rater, and instructional factors.

Saito, K. (forthcoming-b). Age effects in spoken second language vocabulary attainment in adulthood.

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*, *41*, 1133-1149.

Saito, K., Trofimovich., P., Isaacs, T., & Webb., S. (2016). Re-examining phonological and lexical correlates of second Language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.). *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141-156). Bristol, UK: Multilingual Matters.

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*, *28*, 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.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*, 93-120.

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

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*. https://doi.org/10.1177/1362168819858246

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*. DOI: https://doi.org/10.1111/lang.12384

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*, 231–252.

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, *75*, 537-556.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, *35*, 451-482.

Wood, D. (2019). Classifying and Identifying Formulaic Language. In S. Webb (Ed.). *The Routledge Handbook of Vocabulary Studies* (pp. 30-45). London, UK: Routledge.

Supporting Information-A: Speaking Tasks

A. Picture Description (adopted from Derwing & Munro, 2013, *Language Learning*)

B. Oral Interview (adopted from Crowther, Trofimovich, Isaacs, & Saito, 2015, *Modern Language Journal*)

Describe the hardest and toughest challenge in your life.

Your story should start with the following words:

One of the hardest/toughest challenges in my life was _____

- Discussion points
  - ✓ When? How old and where were you?
  - ✓ Why did you encounter this challenge?
  - ✓ Why was it so challenging?
  - ✓ Did anybody (e.g., friends, parents) help you?

- Rounding off questions
  - ✓ What did you learn from this experience?
  - ✓ Would you like to go through the same experience again?

Supporting Information-B: Summary of Bigram and Trigram Examples with Higher T Scores and MI Scores

t scores	Bigram (> 250): Of the, in the, be a, it be, one of Trigram (> 100): a lot of, I think it, it be a, a couple of, be going to
MI scores	Bigram (> 5.0): years ago, an ambulance, even though, every day, years old Trigram (> 3.0): know anything about, get married and, for the entire, there be many, whether or not

To be published in *Second Language Research* (SAGE)

Supporting Information-C: Summary of Biographical Backgrounds of 34 EFL Students and 28 ESL Students

	EFL Group (n = 34)				ESL Student (n = 28)			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chronological age	20.8	7.1	18	25	24.5	3.5	22	39
Age of learning	7.9	1.9	5	13	8.1	2.9	4	15
L2 English proficiency (IELTS)	7.2	0.2	7	8	7.4	0.3	7	8
Length of immersion in the UK (months)	0	0	0	0	9.6	3.4	6	24
Familiarity with Japanese accented English (*6 = very much; 1 = not at all*)	2.5	1.0	1	4	3.1	1.2	1	6
Daily L2 English use (speaking)	11.0%	9.3%	1%	40%	45.8%	23.1%	10%	95%