Accepted for publication in *Journal of Second Language Pronunciation* (John Benjamins)

Linguistic Correlates of Comprehensibility in Second Language Japanese Speech

Kazuya Saito & Yuka Akiyama

Abstract

This study examined phonological, temporal, lexical and grammatical correlates of native speakers' perception of second language (L2) comprehensibility (i.e., ease of understanding). L2 learners of Japanese with various proficiency levels engaged in oral picture description tasks which were judged by native speaking raters for comprehensibility, and then submitted to pronunciation, fluency, and lexicogrammar analyses. According to correlation analyses and linear mixed-models, the native speaking judges' comprehensibility ratings were significantly linked not only with actual usage of words in context (lexical appropriateness) but also with the surface details of words (pitch accent, speech rate, lexical variation). Similar to previous L2 English studies (e.g., Isaacs & Trofimovich, 2012), the influence of segmental and morphological errors in the comprehensibility of L2 Japanese speech appeared to be minor.

Key words: Comprehensibility, Second Language Speech, Japanese, Pronunciation, Fluency, Vocabulary, Grammar

Linguistic Correlates of Comprehensibility in Second Language Japanese Speech

In second language acquisition (SLA), a number of researchers (Derwing & Munro, 2015; Jenkins, 2000; Levis, 2005) have emphasized the importance of assessing overall second language (L2) oral proficiency based on what actually matters in real-life situations—native speakers' intuitive judgements of comprehensibility (how easily L2 utterances can be understood) rather than accentedness (i.e., how much L2 utterances approximate the native speaker norm). Consistent with the agenda, a growing body of SLA studies have corroborated what kinds of pronunciation, fluency, vocabulary, and grammar errors are relatively crucial (or irrelevant) for native speakers' L2 comprehensibility judgements under various task conditions (e.g., Crowther, Trofimovich, Isaacs, & Saito, 2015 for IELTS; Kang, Rubin, & Pickering, 2010 for TOEFL iBT; Isaacs & Trofimovich, 2012 for picture description). Given that these studies have exclusively dealt with English as an L2, the current study was designed to examine the generalizability of the topic—the linguistic correlates of L2 comprehensibility—by focusing on a different L2 context: Learners of Japanese as an L2.

Background

Second Language Comprehensibility

Over the past 40 years, one of the most extensively studied topics in the field of L2 speech research has been how to conceptualize, define and measure L2 oral proficiency. Whereas native speaking raters' impressionistic judgements of foreign accentedness have often been used for this purpose (for a review, Piske, McKay, & Flege, 2001), much research evidence has pointed out that few late L2 learners can attain nativelike speaking proficiency without any detectable L1-related accent (e.g., Flege, Munro, & McKay, 1995). Consequently, a number of L2 education researchers have emphasized the importance of analyzing L2 speech with more realistic and practical standards, such as for comprehensibility and intelligibility rather than accentedness (Derwing & Munro, 2015; Jenkins, 2000; Levis, 2005).

According to Derwing and Munro's seminal work on native speakers' comprehensibility and accentedness judgments (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995), some L2 learners, perceived as heavily accented, can be considered to be highly comprehensible, indicating that the two rating constructs are partially overlapping but essentially independent (see also Jenkins, 2000; Levis, 2005). It is certainly understandable that a number of students will still seek to attain full, accent-free mastery of the target language and should not be discouraged from pursuing their personal goals. However, teachers need to elaborate on a syllabus (i.e., what and how to teach) to help students attain adequate L2 oral ability, especially by improving certain linguistic features with high communicative value in an efficient and effective manner within a limited amount of classroom time (Foote, Holtby, & Derwing, 2012).

Several attempts have been made to identify which linguistic properties of L2 speech are relatively crucial for native speakers' comprehensibility evaluation. For example, it has been shown that comprehensibility is positively related to L2 pronunciation, fluency, lexicon and grammar, including segmental contrasts with high functional load (Munro & Derwing, 2006), L2 specific segmental errors (Jułkowska & Cebrian, 2015), prosody (Derwing, Munro, & Wiebe, 1998), optimal speech rate (Munro & Derwing, 2001), and appropriate lexicon/grammar usage (Derwing, Rossiter, & Ehrensberger-Dow, 2002). More recently, Isaacs and Trofimovich (2012), studying 40 Francophone learners of English, examined which areas of language (pronunciation, fluency, vocabulary, grammar) influence native speakers' judgments of L2 comprehensibility.

The results of this study showed that raters' comprehensibility scores were influenced by a wide range of phonological (segmentals, prosody), temporal (speech rate), lexical (appropriateness, variation) and grammatical (accuracy) factors. By using trained linguistic coders' subjective judgments, Crowther et al. (2015) found that L2 learners' pronunciation, fluency, vocabulary and grammar scores significantly predicted overall L2 comprehensibility especially under a cognitively difficult task condition (i.e., TOEFL iBT) relative to a cognitively easy task condition (i.e., IELTS).

Rater Backgrounds

Another variable potentially affecting L2 speech assessment relates to different types of raters. Certain L2 speech studies have shown that some native listeners tend to show more lenient attitudes towards accented speech, resulting in more positive evaluations, especially when they have relevant experience with foreign accented speech (Saito & Shintani, 2016); have learned the L2 learners' first language (Winke, Gass, & Myford, 2013); interact with the non-native speakers on a daily basis (Kennedy & Trofimovich, 2008); or have taught non-native speakers in classroom settings (Saito, Trofimovich, & Isaacs, 2015).

At the same time, however, other studies have failed to find such significant predictive power of rater background for L2 speech assessment. For example, Isaacs and Thomson (2013) showed that experienced (ESL professionals) and inexperienced (graduate students from non-linguistic fields) similarly rated the comprehensibility of L2 speech without any significant group difference. Their individual difference in rater behaviours was revealed based on the results of verbal protocols and posttask interviews.

Investigating the effects of different rater backgrounds on L2 speech assessment is especially crucial for the development of rater training materials in high-stakes testing environments, where any individual variability among professional raters needs to be minimized as much as possible. Providing explicit information regarding the source of rater effects may help raters in these contexts to obtain and demonstrate shared assessment patterns with little variance in attitudes towards the same accented-speech samples (Winke et al., 2013).

Motivation for Current Study

According to the growing amount of research evidence regarding native speakers' intuitive assessment of L2 English speech on the continuum of comprehensibility, native speakers tend to pay attention to every aspect of language (segmentals, prosody, fluency, lexical appropriateness and variation, morphosyntactic accuracy) to understand what talkers intended to say. This indicates that native speakers' successful understanding of the content of L2 speech can be achieved by gathering as much linguistic information as possible in accented L2 speech. Though revealing, it is noteworthy that most of the existing literature on comprehensibility has exclusively highlighted native speakers' judgments of L2 English speech.[1] To date, it remains open to empirical inquiry whether and to what degree such pedagogically crucial findings can be replicated in other cross-linguistic contexts. For similar findings and discussion, see Crowther et al., 2015 for 45 ESL learners; Saito, Trofimovich, & Isaacs, 2015 for 40 Francophones; Saito, Trofimovich, & Isaacs, 2016 for 120 Japanese learners of English.

[1] It is noteworthy here that a growing number of researchers have examined how not only native but also non-native speakers assess L2 comprehensibility (e.g., Crowther, Trofimovich, & Isaacs, 2016; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002).

To our knowledge, two empirical studies to date have examined phonological, temporal, lexical and grammatical correlates of L2 German (O'Brien, 2014) and Spanish (McBride, 2015) comprehensibility. In O'Brien's study, native speakers of English learning L2 German were recruited to judge the overall comprehensibility of L2 German speech. McBride (2015) examined both native and non-native raters' comprehensibility judgement of L2 Spanish speech. Despite some methodological differences in these individual studies (native vs. non-native judges), the findings have suggested that raters use varied linguistic information to arrive at their comprehensibility judgements (similar to the L2 English studies reviewed above).

The current investigation was designed to examine the generalizability of previous findings in conjunction with native speakers' assessment of L2 Japanese speech. We also examine whether and to what degree such rater behaviours would differ between experienced and inexperienced native raters (for definition, see below). By comparing the results of the current study with previous findings (e.g., Isaacs & Trofimovich, 2012), we reveal which aspects of comprehensibility judgment processes can be generalizable and/or specific to certain cross-linguistic contexts. The study's research questions were thus formulated as follows:

1. Which linguistic features predict the overall comprehensibility of L2 Japanese speech?
2. How do native judges with different rater backgrounds (i.e., ± knowledge of L2 Japanese production) perceive L2 comprehensibility?

Method

Participants

Speakers. Thirty learners of Japanese in six American universities participated in the study (M_{age} = 20.1, *Range* = 19-27). According to the language background questionnaire, their L1 backgrounds widely varied, including native speakers of English (n = 21), Chinese-English bilinguals (n = 5), and heritage learners (n = 4). Despite some difference in their L1 backgrounds, all the participants demonstrated native and near-nativelike proficiency in English and used it as a main language of communication during their school time.

The participants were registered in Japanese courses at the time of the project. To ensure a wide range of L2 Japanese proficiency levels (representing beginner to advanced oral ability), these participants were carefully selected from various instructional levels (n = 10 elementary, n = 6 intermediate, n = 8 advanced, and n = 6 post-advanced). Following L2 research standards (Ortega, Iwashita, Norris, & Rabie, 2002), the participants' proficiency was measured via the Elicited Imitation Test, a test that assesses oral performance of participants who repeat a series of sentences in L2 after a short pause. The test has been proven to correlate with other criterion measures such as the ACTFL Oral Proficiency Interview (Ortega et al., 2002). The results confirmed their varied proficiency skills (M = 80.67 out of 120 points, *Range* = 44-119).

Raters. eight L1 Japanese raters were recruited (M_{age} = 29, *Range* = 25-30 The raters were all born and raised in Japan, with both of their parents being native speakers of Japanese. To recruit native judges with varied rater backgrounds—one potential factor affecting the judgement of L2 oral proficiency (e.g., Winke et al., 2013), we carefully recruited these raters who differed in terms of relevant L2 judgement experience (Isaacs & Thomson, 2013; Saito et al., 2016). The first four raters were considered "experts," as they were Japanese-as-a-foreign-language professionals at a university in the US with a great amount of teaching experience at university-level schools (M_{years} = 4.25, *range* = 1.5-9.5). They were also graduate students in applied linguistics and were familiar with various kinds of L2 analyses. We asked the raters to judge (a) their familiarity with accented Japanese and (b) the amount of contact with L2 Japanese

Table 1
Characteristics of the Learners of Japanese (N = 30)

Instructional Experience (years of Japanese learning)		~ 1 year (*n* =10)	~ 2 years (*n* = 6)	~ 3 years (*n* = 8)	3 years+ (*n* = 6)	All (*N*=30)
EIT scores	*M* (*SD*)	82.1 (16.34)	86.5 (14.73)	81.50 (26.66)	101.71 (7.94)	88.67 (20.95)
	Min-Max	56-104	71-104	44-118	99-119	44-119
Age	*M* (*SD*)	19.7 (1.06)	21.17 (2.48)	20.88 (0.83)	20.67 (1.03)	20.50 (1.46)
	Min-Max	19-21	19-26	20-22	19-22	19-26
Sex	Male	6	5	2	1	14
	Female	4	1	6	5	16
L1 background		English (8), English & Chinese (2)	English (6)	English (4) English & Chinese (3), Heritage (1)	English (3), Heritage (3)	English (21), English & Chinese (5), Heritage (4)

speakers on a 6-point Likert scale (*1 = none, 6 = a lot*). The mean scores for these questions were 5.25 and 6, respectively. The other four raters were considered as "novices," as they were residents in Japan with few opportunities to be exposed to any non-native speakers of Japanese. According to a language background questionnaire, they had no previous teaching and linguistic background at the time of the project. They rated both their familiarity and amount of contact with L2 speakers of Japanese as 0.5 out of 6. None of the participants reported any hearing problems. In the Results, we explore if the rater background (experts vs. novices) affected their comprehensibility ratings.

Speech Materials

As a part of a larger project, the participants in the study were involved in 30-min dyadic conversations with native speakers of Japanese via a video-conferencing tool (i.e., *Google Hangout*) (for details, see Akiyama & Saito, 2016). The conversation exchanges took place in quiet rooms using their own computers. To prepare for the interaction activity, the L2 learners of Japanese were first asked to bring a photo related to one of the following cross-cultural topics: life styles, pop culture, or school life in Japan vs. US. After small talk, the learners described the content of the photo and asked questions related to the photo. To avoid any task effects, the three topics were randomized across the 30 learners. We carefully watched all of the video-recorded conversational sessions, and isolated the first 30 seconds of each participant's photo description.

Unlike previous studies in which participants all described the same pictures (e.g., Derwing & Munro, 1997; Isaacs & Trofimovich, 2012), which in turn resulted in somewhat homogeneous linguistic usage among the participants—speech materials that may not be ideal for robust lexicogrammar analyses, the relatively loose, flexible and free task format (describing a photo of choice related to one of the themes) was adopted in the study. As such, we were able

to elicit L2 Japanese learners' diverse use of vocabulary during spontaneous speech. Similar tasks have been used in L2 vocabulary and grammar research, such as monologues (e.g., Koizumi & In'nami, 2013) and oral interviews (e.g., Segalowitz & Freed, 2004). The length of each speech sample (30sec) provided an appropriate amount of linguistic information for native raters' assessments (see Derwing & Munro, 1997 for the relationship between sample lengths vs. raters' judgement scores).

Global Analyses

As operationalized in previous studies (e.g., Derwing & Munro, 1997; Isaacs & Trofimovich, 2012), comprehensibility was measured via native speakers' *intuitive* judgements. The raters listened to speech samples played in a randomized order via a custom program, which we developed using the commercial software package MATLAB 8.1 (The MathWorks Inc., Natick, MA, 2013). Upon hearing each sample, raters used a free-moving slider on a computer screen based on a 1000-point scale to evaluate comprehensibility (*0 = hard to understand, 1000 = easy to understand*), with the leftmost corner labeled with a frowning face and the rightmost corner with a smiling face. To ensure that ratings tapped into intuitions equivalently, the raters listened to each speech sample only once.

First, the eight raters received brief instruction in Japanese on the definitions of comprehensibility. Following Derwing and Munro's (1997) definition, comprehensibility referred to how much effort it takes to understand an utterance (for Japanese translations of onscreen labels and training scripts, see the Appendix). After familiarizing themselves with the procedure by practicing with four speech samples (not included in the main analysis), they rated the 30 speech samples in the main dataset. The session took approximately one hour.

Linguistic Analyses of L2 Japanese

In order to explore the linguistic influences on comprehensibility, the precursor research (e.g., Isaacs & Trofimovich, 2012) analyzed the phonological, temporal, lexical and grammatical aspects of L2 English speech samples, and examined their correlations with native raters' global (comprehensibility) judgement scores. Following the research standards for L2 English, two linguistically trained coders conducted linguistic analyses on 10 different dimensions of L2 speech spanning pronunciation (segmentals, pitch accent, intonation), fluency (mora per minute, filled and unfilled pauses), vocabulary (appropriateness, variation, sophistication) and grammar (morphological accuracy). Subsequently, the second coder double-checked the validity of the linguistic analyses. If there was any disagreement, the coders discussed the disagreements and came to agreement by conducting reanalyses. Both of them were native speakers of Japanese and graduate students in applied linguistics with extensive experience in L2 speech analyses of this kind.

Segmentals. L2 learners of Japanese are reported make a range of segmental errors in vowels and consonants specific to the Japanese phonetic system (Toda, 2003). For example, English learners can substitute English /ɹ/ for the Japanese tap sound (i.e., L1 substitution) (e.g., /ɹiŋgo/ instead of /ɾiŋgo/ for "apple"). Another segmental problem could be the long vs. short vowel distinction (e.g., /oba:saɴ/ instead of /obasaɴ / for "grandmother") as well as the deletion or addition of sounds (/kite/ instead of /kit:e/ for "stamp," /anai/ instead of /aɴnai / for "guidance"). The segmental category was analyzed based on the number of all substitution, long/short vowel distinction, and addition/deletion errors over the total number of moras.

Pitch accent. Another pronunciation feature difficult for L2 Japanese learners relates to the correct place of pitch in a word. In Japanese, one of the moras in accented words will be pronounced with a high falling pitch while the rest of the moras are pronounced with a low falling pitch. In contrast, accentless words will receive little pitch change. Pitch accents can distinguish phonologically similar words, such as /ha$^{\downarrow}$si/ for "chopsticks" and /hasi$^{\downarrow}$/ for "bridge". The Japanese pitch system is somewhat different from the prosodic system in English, wherein one syllable in a bisyllabic/multisyllabic word is pronounced longer and louder than the others. Non-native speakers of Japanese often place pitch accent on the wrong mora, or put equal stress on all of the moras in a word (Akagi, Uchida, & Furuichi, 2010). The pitch accent category was analyzed based on the amount of misplaced, or absence of correct pitch over the total number of content words.

Intonation. Finally, non-nativelike use of L2 Japanese pronunciation can be determined based on the degree of deviation from the pitch changes that occur while speaking at the sentence level (i.e., intonation). To mark any type of question (wh-questions, yes-no, tag), for example, native speakers' pitch goes up at the end of the sentence. Non-native speakers may speak Japanese without any emphasis of intonation (i.e., flat speech) and/or demonstrate deviant intonation patterns based on L1 strategies (e.g., L1 English learners using a falling pitch for wh-questions). The intonation category was analyzed based on the total number of intonation errors over the total number of sentences where different pitch patterns are expected.

Fluency. In the previous literature, fluency has traditionally been conceptualized as pause frequency (the number of filled/unfilled pauses) and tempo (speech/articulation rate) across various L1/L2 contexts (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). Filled pauses are fillers that are used between meaningful utterances, and can be universal (e.g., ah, eh, umm) or Japanese-specific ones (eto, ano, nanka, ma). Unfilled pauses are silences that are longer than 0.5 sec when the speaker has the floor.[2] As is the case with L2 English, non-native Japanese speakers' speech can be considered as "disfluent" when it includes a large number of filled/unfilled pauses and repetitions. As for measuring tempo in Japanese, since Japanese is a mora-timed language as opposed to a stress-timed language such as English (Warner & Arai, 2001), speech rate can be measured via moras per minute. Therefore, the fluency category was measured by the total number of filled pauses and unfilled pauses as well as the number of moras per minute.

Vocabulary. According to L2 vocabulary research, the lexical quality of speech production has been extensively analyzed from three different perspectives: (a) the contextually and conceptually appropriate use of the vocabulary words used by L2 learners (i.e., appropriateness); (b) learners' access to a wide range of lexical items (i.e., variation); and (c) their ratio of infrequent word usage (i.e., sophistication). There is some evidence that non-native speakers' Japanese speech is characterized by a large number of lexical choice errors as well as the overuse of frequent words (Hatakeyama, 2014).

In this study, the lexical appropriateness category was operationalized as the number of lexical choice errors (e.g., "e:ga" [movie] for "eigo" [the English language]) and English

[2] As mentioned in Derwing, Rossiter, Munro, and Thomson (2004), there has been no "correct" standard with regard to pause lengths (e.g., 0.2 sec vs. 0.4 sec vs. 0.5 sec).

substitutions (e.g., "public" instead of "ko:kyo") over the total number of running words; the lexical variation category as the total number of different words (i.e., types) by the square root of the number of tokens (types/√tokens), calculated via the Guiraud index; and the lexical sophistication category as the total number of Japanese Language Proficiency Test Level 1-2 vocabulary items[3] over the total number of running words.

Grammar. In previous research, this domain has been analyzed based on the number of morphological errors in the domain of verbs (tense/aspect, subject/verb agreement), nouns (plurality), determiners, and prepositions (e.g., Yuan & Ellis, 2003). Different from L2 English, where grammatical functions are mainly determined by word order, morphological suffixes in L2 Japanese (e.g., particles) play an important role in marking a range of grammatical functions of words (e.g., case, sentence ending, binding, adverbial, conjunctive) in a sentence. The morphological accuracy category was thus calculated according to the total number of morphological errors in the conjugations of verbs/adjectives/nouns (e.g., te-form) and derivational forms (e.g., hayaku vs. hayai), tense/aspect, voice (e.g., "causative" and "causative passive"), modality (e.g., "tewaikenai"), particles, and transitivity over the total number of running words. The linguistic characteristics of the 30 speech samples are summarized in Table 2.

Table 2
Pronunciation, Fluency and Lexicogrammar Aspects of the 30 Speech Samples

Linguistic dimensions	Calculation method	*M*	*SD*	*Min*	*Max*
A. Pronunciation					
Segmental errors	No. of errors per mora	.008	.012	.000	.045
Pitch accent errors	No. of errors per content word	.099	.105	.000	.364
Intonation errors	No. of errors per obligatory context	.078	.213	.000	1.000
B. Fluency					
Filled pauses	No. of filled pauses per minute	4.96	3.96	0	16
Unfilled pauses	No. of unfilled pauses per minute	.1.16	1.62	0	6
Speech rate	No. of moras per minute	144.0	37.7	78	242
C. Lexicogrammar					
Lexical error ratio	No. of errors per word	.043	.046	0	.133
Lexical variation	Via Guiraud index	3.879	0.708	1.964	5.746
Lexical sophistication	No. of infrequent words per word	.041	.052	.000	.171
Morphosyntactic error ratio	No. of errors per word	.023	.040	0	.182

[3] The Level 1-2 items consist of advanced-level vocabulary that predicts the ACTFL OPI rating higher than High-Intermediate (Hatakeyama, 2014).

Results

Inter-rater reliability

Cronbach's alpha was used to examine the raters' behaviours during comprehensibility judgements with only minimal instruction (on a 1000-point scale). The raters demonstrated consistent agreement ($\alpha = .82$), suggesting that they generally agreed on their intuitive notion of what it meant by ease of understanding and linguistic nativelikeness in L2 Japanese speech. Cronbach's alpha also revealed that the four expert raters, who had taught Japanese at a college in the US, demonstrated slightly more consistent ratings ($\alpha = .83$) compared to the four novice raters, who were residing in a monolingual environment in Japan with little familiarity with accented Japanese ($\alpha = .69$).

Experts vs. Novices

To further examine the influence of rater background in their comprehensibility assessment, we performed an independent samples *t*-test to analyze if they assigned different comprehensibility scores to the same 30 L2 Japanese samples. The descriptive results are summarized in Table 3.

Table 3
Summary of Comprehensibility Scores among the Expert vs. Novice Raters

Rated categories	Rater type	*M*	*SD*	*Range*
Comprehensibility	Experts (*n* = 4)	584	246	131-960
	Novices (*n* = 4)	572	270	101-986

The results did not yield any statistical significance, $t(58) = .176$, $p = .861$, indicating that the raters reacted to accented Japanese samples in a similar way, irrespective of their background. Unlike previous L2 English research (e.g., Winke et al., 2013), we did not find enough statistical evidence for significantly different rater behaviours between the experts and novices (for similar results, see Isaacs & Thomson, 2013). Thus, in the following analyses, all raters' scores were averaged to derive a single score for the perceived comprehensibility of each speech sample.

Linguistic Correlates of Comprehensibility

To explore how the raters differentially used linguistic information during their comprehensibility evaluations, we conducted a set of correlation analyses to examine how their rating scores were related to the 10 linguistic domains of the speech samples. Due to the small sample size (n = 30 speech samples), a more conservative method than the Pearson's, Spearman nonparametric correlation, was selected. The results of the linguistic influences on their comprehensibility evaluation are summarized in Table 4. Comprehensibility was mildly correlated with pitch accent ($r = -.40$) and lexical variation ($r = 41$), and weakly correlated with speech rate ($r = .36$) and lexical appropriateness ($r = -.37$). In other words, more comprehensible speech samples were comprised of fewer pitch accent errors and more varied and appropriate vocabulary items and delivered at a faster speed. This in turn suggests that the raters similarly relied on prosody and fluency information as well as vocabulary (variation but also appropriateness) during their comprehensibility judgements.

Table 4
Spearman Correlations between the Linguistic Variables and Mean Comprehensibility Ratings

Linguistic variables	Comprehensibility	
	r	*p*
Pronunciation		
Segmental errors	-.20	.29
Pitch accent errors	-.40	.02*
Intonation errors	-.16	.39
Fluency		
Filled pauses	-.27	.14
Unfilled pauses	-.23	.20
Speech rate	.36	.04*
Lexicogrammar		
Lexical error ratio	-.37	.04*
Lexical variation	.41	.02*
Lexical sophistication	.15	.42
Morphosyntactic error ratio	-.19	.29

* denotes $p < .05$

Linear Mixed Model Analyses

Finally, we approached the linguistic correlates of comprehensibility by using a different statistical method—a linear mixed model analysis. Taking random effects of the raters' potentially different behaviours into account (rather than averaging all of their comprehensibility scores), this mixed-model analysis allowed us to analyze the relationship between the raters' comprehensibility scores as a dependent variable relative to a total of 10 linguistic predictors (segmentals, pitch, intonation, filled pauses, unfilled pauses, speech rate, lexical appropriateness, variation, lexical sophistication and morphology). All of the necessary conditions (homogeneity, normal distribution) were met for the analysis. According to the results of the analyses (summarized in Table 4), significant fixed effects were found for (a) pitch, F (1, 17) = 14.02, p = .002, (b) lexical appropriateness, F (1, 17) = 5.07, p = .038, and (c) lexical variation, F (1, 17) = 5.56, p = .031.

In sum, the results presented here showed that the raters' comprehensibility judgements were significantly affected by the number of pitch accent/vocabulary errors and the total number of different words. Different from the correlation analyses mentioned above, the mixed analyses did not find speech rate to be a significant predictor for L2 comprehensibility. This indicates that the effect of fluency (speech rate) on raters' successful understanding of L2 Japanese could be subject to much rater variability.

Table 5
Summary of Significant Predictors for L2 Comprehensibility via Linear Mixed-Model Analyses

	Mixed-Model ANOVA	
Linguistic variable	$F(1, 17)$	p
Pronunciation		
Segmental errors	3.96	.098
Pitch accent errors	14.02	.002*
Intonation errors	1.73	.206
Fluency		
Filled pauses	.25	.622
Unfilled pauses	.58	.456
Speech rate	.12	.734
Lexicogrammar		
Lexical error ratio	5.07	.038*
Lexical variation	5.56	.031*
Lexical sophistication	.158	.696
Morphosyntactic error ratio	.782	.399

* indicates $p < .05$

Discussion

Due to the significant practical value of the topic, much research attention has been directed to corroborating how native speakers perceive non-native speakers' oral proficiency, especially along the continuum of ease of understanding (comprehensibility) (Derwing & Munro, 2015). To examine the generalizability of previous findings (exclusively specific to L2 English) (e.g., Isaacs & Trofimovich, 2012), we revisited the topic in another L1-L2 context—native speaking raters' assessment of L2 learners of Japanese.

In response to the first research question (the linguistic correlates of L2 comprehensibility), the results of the correlation analyses and mixed-model analyses showed that the raters' comprehensibility scores were related to adequate and varied prosody (pitch accent), type-frequency of words (lexical variation) and meaning aspects of words (lexical appropriateness). Our findings (for L2 Japanese) were generally consistent with the previous literature on L2 English, in that native speakers understand L2 speech in an optimal fashion by selectively attending to particular linguistic features affecting successful communication, such as prosody (Kang et al., 2010), fluent delivery of speech (Derwing et al., 2004), and diverse and proper word choice (Saito, Trofimovich, Isaacs, & Webb, 2017).

Furthermore, the findings here also echoed that the segmental and morphological accuracy seems to have less influence on L2 comprehensibility than errors at the suprasegmental level, as observed in previous L2 English studies (e.g., Derwing, Munro, & Wiebe, 1998). In the context of L2 English speech, scholars have indeed suggested that certain errors are believed to entail more communicative value than others, such as those with high functional load (Munro & Derwing, 2006) and communicative adequacy (Foster & Wigglesworth, 2016). In addition, previous research has also suggested that the role of segmental and morphological errors in L2 comprehensibility may vary according to learners' proficiency levels; whereas lexical and suprasegmental accuracy is fundamental at the initial stage of L2 comprehensibility development (beginner to intermediate), the segmental and morphological accuracy is necessary at the later

stage of L2 comprehensibility development (intermediate to advanced) (Isaacs & Trofimovich, 2012).

As for the second research question (the role of rater backgrounds), the results did not find any group difference between the expert and novice raters' comprehensibility scores. At the same time, the two different statistical analyses—correlation and mixed-model analyses—demonstrated slightly different results regarding the effect of speech rate on L2 comprehensibility. When all of the raters' scores were averaged across (without taking into account their potentially different rating behaviours), the correlation analyses found that speech rate was significantly predictive of L2 comprehensibility judgement. With such individual rater difference statistically controlled, however, the mixed-model ANOVAs did not find such significant fluency-comprehensibility link.

One possible implication from the mixed results is that the role of fluency in L2 Japanese comprehensibility could be subject to certain kind of individual variability among raters. Due to the lack of group differences between the expert and novice rater groups, however, this individual variability—whether raters attend to speech rate while assessing L2 comprehensibility—could not be attributed to the raters' relevant experience with accented L2 Japanese (cf. Winke et al., 2013). Thus, scrutinizing the source or rater difference may further require more qualitative analyses, such as think-aloud protocols (e.g., Isaacs & Thomson, 2013; Isaacs & Trofimovich, 2012).

Limitations

Given that the current study took an exploratory approach towards examining the linguistic correlates of L2 Japanese speech, we address here several topics worthy of future research attention based on some limitations. First, our interpretations were based on a relatively small dataset (30 L2 Japanese learners judged by eight native raters), and should thus be considered as tentative, because some of the findings could be simply artifacts of statistical analyses. Similarly, the number of raters was small ($n = 8$) relative to other L2 speech studies ($n = 60$ for Isaacs & Trofimovich, 2012). The findings reported here need to be replicated with larger samples in different cross-linguistic contexts.

Second, our study used a picture description task during video-based conversation. Given that L2 learners' comprehensibility widely varies according to task type (Derwing et al., 2004), future studies need to examine the linguistic influences on L2 comprehensibility by adopting different speaking tasks, especially more argumentative, formal and complex ones (see Hulstijn, Schoonen, de Jong, Steinel, & Florijn, 2012). In a related issue, the participants in the study provided free speech on three different topics (life styles, pop culture, school life in Japan vs. US). To our knowledge, no empirical studies have ever examined the effects of similarly familiar topic types on L2 comprehensibility judgements (cf. Varonis & Gass, 1982). Future research needs to further explore whether, to what degree and how topic choice affects linguistic qualities of L2 speech and thus impacts overall comprehensibility.

Third, we acknowledge that our findings need to be interpreted with caution, especially with regard to the lexicogrammar analyses. Although the length of each speech sample (30 sec) could be considered adequate in line with L2 pronunciation and fluency research (e.g., Derwing & Munro, 1997), it may fail to reach the standards of L2 vocabulary and grammar research (i.e., > 3 min) (e.g., Yuan & Ellis, 2003). Indeed, there is some evidence that a certain amount of text (more than 100 words) is needed for the purpose of the robust lexical analyses, at least in L2 English (Koizumi & In'nami, 2012).

Last, it is important to note that we restricted our literature review, method and discussion to comprehensibility, and eschewed any mention of intelligibility. Whereas comprehensibility is generally assessed via native speakers' scalar judgement (e.g., Trofimovich & Isaacs, 2012), intelligibility, defined as an actual product of understanding (Derwing & Munro, 2015), has been evaluated by way of a range of outcome measures, such as native speakers' transcriptions of speech samples (Derwing & Munro, 1997), comprehension questions (Hahn, 2004), and impressionistic judgments (Anderson-Hsieh, Johnson, & Koehler, 1992). As Isaacs (2008) argues, the methodological variability of intelligibility measures points to the strong call for more empirical studies which elaborate and validate adequate tasks and methods of analysis specific to various research contexts and goals. Thus, future studies should elaborate and validate reliable assessments to examine the intricate role of pronunciation, fluency and lexicogrammar in determining native speakers' intelligibility judgements.

Conclusion

Using 30 L2 learners of Japanese, the current study investigated the linguistic correlates of comprehensibility. According to the results of the correlation and mixed-model ANOVAs, the native speaking judges' comprehensibility ratings were significantly linked not only with actual usage of words in context (lexical appropriateness) but also with the surface details of words (pitch accent, speech rate, lexical variation). The findings here suggested native speaking raters tend to assess the perceived comprehensibility of accented speech by checking each domain of language with relatively high communicative value (prosody, fluency, vocabulary).

Building on previous studies concerning L2 learners of English (Derwing & Munro, 2015; Isaacs & Trofimovich, 2012; Saito et al., 2015, 2016, 2017), McBride (2015) and German (O'Brien, 2014), the findings presented here led us to postulate a tentative hypothesis on the underlying mechanisms of L2 comprehensibility assessment. Overall, comprehensibility is a unique index of how much progress L2 learners have made towards acquiring the communicatively adequate level of competence with multiple pronunciation, fluency, vocabulary and grammar elements which affect successful L2 communication. Given the great deal of theoretical and pedagogical importance of the topic, these tentative suggestions on L2 comprehensibility need to be further replicated in various cross-linguistic contexts.

References

Akagi, H., Uchida, N., & Furuichi, Y. (2010). *Rizumu de minitsuku nihongo no hatsuon [Learning Japanese pronunciation with rhythms]*. Tokyo: Three A Network.

Akiyama, Y., & Saito, K. (2016). Development of comprehensibility and its linguistic correlates: A longitudinal study of video-mediated telecollaboration. *Modern Language Journal, 100*, 585-609.

Anderson-Hsieh, J., Johnson K, & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning,* 42, 529-555.

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

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

Derwing, T. M.., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 12*, 1–16.

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

Derwing, T. M., & Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 393–410.

Derwing, T. M., Rossiter, M. J., & Ehrensberger-Dow, M. (2002). "They speaked and wrote real good": Judgements of non-native and native grammar. *Language Awareness, 11,* 84-99.

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

Flege, J., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97*, 3125-3134.

Foote, J., Holtby, A., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal, 29*, 1-22.

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.

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201–223.

Hatakeyama, M. (2014, March). *Corpus analysis of vocabulary by OPI proficiency levels: Interviews from Database of Japanese Language Learners Conversation (DJLLC)*. Poster session presented at the meeting of the American Association of Japanese Teachers, Philadelphia, PA.

Hulstijn, J.H., Schoonen, R., De Jong, N.H., Steinel, M.P, & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing, 29*, 203 - 221.

Isaacs, T., (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review 64*, 555–580.

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

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34,* 475-505.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Jułkowska, I. A., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, *1*, 211-237.

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.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64,* 459–489.

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

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 367–377.

Major, R., Fitzmaurice, S., Bunta, F., & Balasubramanian, C. (2002). The Effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36,* 173–190.

McBride, K. (2015). Which features of Spanish learners' pronunciation most impact listener evaluations? *Hispania*, *98*, 14-30.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*, 73-97.

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

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

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, *64*, 715-748.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research.* Paper presented at the Second Language Research Forum, Toronto.

Piske, T., MacKay, I., & Flege, J. (2001). Factors affecting degree of foreign accents in an L2: A review. *Journal of Phonetics*, *29*, 191-215.

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., Trofimovich, P., & Isaacs, T. (2015). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*. doi: 10.1093/applin/amv047

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

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs, P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives*. Bristol, UK: Multilingual Matters

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173-199.

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition, 4*, 114-136.

Warner, N., & Arai, T. (2001). The role of the mora in the timing of spontaneous Japanese speech. *The Journal of the Acoustical Society of America*, *109*(3), 1144-1156.

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.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1-27.

Appendix

Training materials (original text)

Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

Japanese translation (Japanese translation)

理解しやすさ	このカテゴリーでは、どれだけノンネィティブスピーカーの日本語が理解しやすいかを判断して下さい。もし内容を簡単に理解出来るのであれば、それは理解しやすさが高い日本語です。しかし内容を理解するのに注意深く聞き取らなければならないか、もしくは全く理解出来ない場合、それは理解しやすさが低い日本語です。