8 Re-examining Phonological and Lexical Correlates of Second Language Comprehensibility: The Role of Rater Experience

Kazuya Saito, Pavel Trofimovich, Talia Isaacs and Stuart Webb

Introduction

Few researchers and teachers would disagree that some linguistic aspects of second language (L2) speech are more crucial than others for successful communication. Underlying this idea is the assumption that communicative success can be broadly defined in terms of speakers' ability to convey the intended meaning to the interlocutor, which is frequently captured through a listener-based rating of comprehensibility or ease of understanding (e.g. Derwing & Munro, 2009; Levis, 2005). Previous research has shown that communicative success – for example, as defined through comprehensible L2 speech – depends on several linguistic dimensions of L2 output, including its segmental and suprasegmental pronunciation, fluency-based characteristics, lexical and grammatical content, as well as discourse structure (e.g. Field, 2005; Hahn, 2004; Kang *et al.*, 2010; Trofimovich & Isaacs, 2012). Our chief objective in the current study was to explore the L2 comprehensibility construct from a language assessment perspective (e.g. Isaacs & Thomson, 2013), by targeting rater experience as a possible source of variance influencing the degree to which raters use various characteristics of speech in judging L2 comprehensibility. In keeping with this objective, we asked the following question: What is the extent to which linguistic aspects of L2 speech contributing to comprehensibility ratings depend on raters' experience?

Linguistic correlates of L2 comprehensibility

The relationship between L2 comprehensibility and the linguistic content of L2 speech (e.g. in terms of its segmental, suprasegmental or fluency-based characteristics) has been a productive area of research. For instance, L2 comprehensibility appears to be related to various linguistic dimensions of L2 speech, including individual sounds with high functional load, such as those that distinguish word meaning across many word pairs (Munro & Derwing, 2006), and those that represent 'lingua franca core' sounds, such as vowels and consonants which frequently lead to miscommunication in interaction between L2 speakers (Jenkins, 2000). Beyond segmentals, understandable L2 speech also seems to be linked to the production of word stress (Field, 2005), sentence stress (Hahn, 2004), and such aspects of fluency as pausing frequency (Kang *et al.*, 2010).

Perhaps a question that is more relevant to both language researchers and teachers is not which linguistic dimensions of speech contribute to L2 comprehensibility, but rather which linguistic dimensions are relatively more important for comprehensibility, compared to other dimensions. For example, Derwing *et al.* (1998) showed that a 12-week course for ESL learners in Canada with an explicit focus on prosody (i.e. suprasegmentals, such as word stress) and fluency resulted in more gains in learners' comprehensibility compared to instruction targeting individual segments. It appears, then, that an instructional focus on L2 prosody and fluency may lead to a greater impact on comprehensibility than a focus on individual segments (see also Derwing *et al.*, 2014). In another study, Isaacs and Trofimovich (2012) examined how various segmental, prosodic and temporal characteristics of L2 speech (18 speech measures in total) interact to determine the comprehensibility of 40 native French speakers of English. Their findings showed that word stress (prosody) distinguished speakers of low, mid and high levels of comprehensibility, while speech rate (fluency) discriminated between low and intermediate levels, and vowel and consonant errors (segmental accuracy) distinguished intermediate from high levels. Similar findings were reported in follow-up studies featuring 60 ESL learners from various native language (L1) backgrounds (Crowther *et al.*, 2015b) and 120 Japanese learners of English (Saito *et al.*, 2016).

A growing number of studies have recently focused on vocabulary–comprehensibility links, targeting lexical profiles of advanced, intermediate and beginner level learners' spontaneous production. In these studies, L2 speech is often evaluated from written transcripts rather than from audio-recordings, to minimize pronunciation and fluency influences on speech assessments. For example, transcript-based ratings of lexical proficiency (ranging from 'high' to 'low') have been shown to be related to lexical sophistication (in terms of word frequency counts), abstractness (measured as lexical hierarchy), and lexical appropriateness (defined through collocation

accuracy) (Crossley *et al.*, 2011, 2015). In our own recent study (Saito *et al.*, in press), we also asked raters to judge comprehensibility in transcribed L2 speech samples. We found that lexical appropriateness (number of lexical errors), variation (lexical diversity), fluency (number of fillers produced), and abstractness (word imageablity) were crucial for distinguishing between beginner and intermediate comprehensibility levels. When it came to advanced comprehensibility, raters also seemed to attend to morphological appropriateness (morphology errors) and were sensitive to the use of semantically complex words with multiple senses.

Motivation for the current study

Apart from the linguistic characteristics of the speech itself, other variables can influence L2 comprehensibility. One source of such influences relates to various listener characteristics, which include the amount of listeners' exposure to and experience with L2 speech, the extent of their linguistic training, or the degree to which their own language background overlaps with that of the speaker. For example, listeners who are familiar with accented L2 speech or those who share the speakers' L1 tend to rate L2 speech differently, demonstrating more lenient attitudes towards accented speech, compared to listeners without relevant experience (e.g. Kennedy & Trofimovich, 2008; Winke *et al.*, 2013). Prior linguistic training also appears to matter for speech ratings. As shown by Isaacs and Thomson (2013), unlike novice raters who do not possess sufficient knowledge to articulate their rating decisions, experienced raters (ESL professionals) can explain their judgements by drawing on their linguistic knowledge and access to vocabulary or applied linguistics jargon with which to express themselves.

While it is clear that listeners' characteristics influence how they evaluate global aspects of L2 performance (e.g. in terms of overall accent or comprehensibility), the extent to which listeners with different experience profiles attend to similar linguistic characteristics of L2 speech to arrive at their rating decisions is unknown. This issue is important because understanding expert and novice judges' rating processes can inform the training of raters, particularly in the context of high-stakes language proficiency tests (e.g. IELTS), where all participating raters are expected to demonstrate consistent L2 speech assessments (Winke *et al.*, 2013). Recently, Saito and Shintani (2016) took an exploratory approach towards examining linguistic correlates of L2 comprehensibility, as perceived by listeners from different backgrounds (Singaporean and Canadian raters). The Singaporean raters, who had access to various native and nonnative models of English and also spoke a few L2s in a multilingual environment, tended to assign more lenient comprehensibility judgements compared to raters in Canada. Singaporean raters paid attention not only to pronunciation aspects of L2 speech but also to its lexical and grammatical content. In contrast, the comprehensibility

judgements of the Canadian raters, who used only North American English in a monolingual environment, were mainly determined by the pronunciation accuracy and fluency of L2 speech. However, the raters in Saito and Shintani's study evaluated relatively short samples (30 seconds), which may have been too short to capture various linguistic aspects of L2 speech (especially those specific to its lexis and grammar content) so that their relative contribution to L2 comprehensibility could be determined.

The current study extended previous research investigating rater influences on L2 speech assessment (Isaacs & Thomson, 2013; Saito & Shintani, 2016) by focusing on expert and novice raters' assessments of L2 comprehensibility. Two separate rater sessions with expert and novice raters were conducted to examine the role of pronunciation and lexis in L2 comprehensibility. In the first session, the raters evaluated audio samples so that their ratings could be related to extensive analyses of the same samples for several pronunciation variables (i.e. segmental and syllable structure errors, word stress, intonation and speech rate). In the second session, they evaluated transcribed speech, and their ratings were compared to extensive lexical analyses of the same speech samples (i.e. in terms of frequency, diversity, polysemy, hypernymy, text length, lemma and morphology). In line with previous L2 vocabulary research (e.g. Crossley *et al.*, 2011, 2015), the targeted speech samples were relatively long (about 3 minutes), which maximized the likelihood that they included a variety of pronunciation and lexical phenomena that could be linked to L2 comprehensibility.

Pronunciation Aspects of Comprehensibility

Rating materials

The speech samples consisted of 40 native French speakers' descriptions of an eight-frame cartoon narrative from our previous research (e.g. Isaacs & Trofimovich, 2012; Saito *et al.*, 2015). The speakers represented a range of L2 speaking ability, from complete beginners to simultaneous French-English bilinguals. The length of the recorded audio samples varied from 55 to 351 seconds (M = 146 seconds), which corresponded to 75–485 words produced (M = 209.2).

Audio-based comprehensibility analyses

Expert and novice raters

We recruited: (a) five expert raters who were graduate students in applied linguistics at an English-speaking university in Montreal, Canada (Mage = 28.0); and (b) five novice raters who were undergraduate or graduate students with non-linguistic majors at the same school (Mage = 22.6).

As residents in a bilingual (French-English) city, the raters were comparable in terms of their high familiarity with French-accented English (Kennedy & Trofimovich, 2008). However, in line with Isaacs and Thomson's (2013) definition of experienced and inexperienced raters, the two rater groups differed in their familiarity with L2 speech assessment. The expert raters had all taken graduate-level linguistics courses where they had received training in pronunciation, vocabulary and grammar analyses. These raters additionally reported on average 3.5 years (8 months–12 years) of prior English teaching experience, where they were tasked with evaluating their own students' L2 proficiency. In contrast, the novice raters had not completed any courses in linguistics and had no experience with teaching English.

Procedure

Following Derwing and Munro (2009: 478), comprehensibility was defined as 'the listener's perception of how easy or difficult it is to understand a given speech sample', and measured via scalar judgements. As described in Saito *et al.* (2015), the raters used a moving slider to provide a comprehensibility score on a scale between 0 = 'hard to understand' and 1000 = 'easy to understand'. If the slider was placed at the leftmost end of the continuum, labelled with a frowning face (indicating the negative endpoint), it was recorded as 0; if it was placed at the rightmost end of the continuum, labelled with a smiley face (indicating the positive endpoint), it was recorded as 1000. The raters first received brief instruction from a trained research assistant (via training scripts and onscreen labels, as shown in the Appendix). After familiarizing themselves with the rating procedure by rating three practice samples, they proceeded to the main dataset, with all 40 samples played randomly through a MATLAB interface. To ensure that raters' judgements reflected their intuitions, resembling real-life experiences with speech, the raters listened to each sample only once but were required to listen to each sample in its entirety. To reduce fatigue, the rating took place in two one-hour sessions.

Pronunciation analyses

As reported in Isaacs and Trofimovich (2012), the speech samples were analyzed for five pronunciation variables, with all analyses carried out and verified by two trained coders. The intraclass correlations were > 0.90. The five pronunciation variables were operationalized as follows:

(1) *Segmental error ratio*, defined as the total number of phonemic substitutions (e.g. 'put' spoken with /u/ in place of /ʊ/) divided by the total number of segments articulated.
(2) *Syllable structure error ratio*, defined as the total number of vowel and consonant epenthesis (insertion) and elision (deletion) errors (e.g. 'have' spoken without the initial /h/) divided by the total number of syllables articulated.

(3) *Word stress error ratio*, defined as the total number of instances of word stress errors (misplaced or missing primary stress) in polysyllabic words (e.g. 'SUIT-case' spoken as 'suit-CASE') over the total number of polysyllabic words produced.
(4) *Intonation error ratio*, defined as the number of correct pitch patterns at the end of phrases (syntactic boundaries) over the total number of instances where pitch patterns were expected (e.g. 'In a busy street [level tone], there is a businessman and a businesswoman [falling tone]').
(5) *Articulation rate*, defined as the total number of syllables produced excluding dysfluencies (e.g. filled pauses, repetitions, self-corrections, false starts) over the total speech sample duration.

Results

We first calculated Cronbach's alpha to check inter-rater agreement in raters' comprehensibility judgements. The expert raters showed higher consistency ($\alpha = 0.91$) than the novice raters ($\alpha = 0.81$). Since these indexes exceeded benchmark consistency values ($\alpha = 0.70$; Larson-Hall, 2010), mean comprehensibility ratings for each L2 speaker were computed by pooling the data across the five expert and five novice raters, respectively (see Table 8.1 for descriptive statistics).

Next, we compared the expert and novice raters' comprehensibility scores using a matched-samples *t*-test, which showed that the expert raters assigned significantly higher (and thus more lenient) comprehensibility scores compared to the novice raters ($t(39) = 3.05$, $p = 0.004$, $d = 0.21$). Finally, we examined how the expert and novice raters' comprehensibility scores were related to the five pronunciation variables in L2 speakers' speech, using correlation and regression analyses. As summarized in Table 8.2, correlation analyses showed that both expert and novice raters' comprehensibility scores were significantly associated with segmental, word stress and intonation errors, and nearly to the same degree.

We then performed two sets of multiple regression analyses to explore the degree to which the three pronunciation variables (segmental, word stress and intonation errors) predicted the expert and novice raters' comprehensibility scores. These analyses (summarized in Table 8.3) revealed that

Table 8.1 Descriptive statistics for expert and novice raters' comprehensibility scores on a 1000-point scale

Speaking dimension	*Mean*	*SD*	*Range*
Comprehensibility (expert)	713	196	267–998
Comprehensibility (novice)	667	233	214–1000

Note: 0 = 'hard to understand', 1000 = 'easy to understand'.

Table 8.2 Pearson correlations between the five pronunciation variables and expert and novice raters' comprehensibility ratings

Pronunciation variable	*Comprehensibility*	
	Expert raters	*Novice raters*
Segmental errors	–0.51*	–0.51*
Syllable structure errors	–0.36	–0.36
Word stress errors	–0.80*	–0.76*
Intonation errors	–0.51*	–0.51*
Articulation rate	0.32	0.38

Note: *$p < 0.01$ (Bonferroni adjusted).

Table 8.3 Results of multiple regression analyses using pronunciation variables as predictors of comprehensibility

Predicted variable	*Predictor variable*	*Adj. R^2*	*R^2 change*	*F*	*p*
Comprehensibility (expert)	Word stress	0.63	0.63	66.02	0.001
Comprehensibility (novice)	Word stress	0.56	0.56	50.78	0.001

Note: The variables entered into the regression included segmental, word stress and intonation errors; no evidence of strong collinearity was found (*VIF* < 1.259).

the number of word stress errors was the only significant predictor of the expert and novice raters' comprehensibility scores (accounting for a total of 62.5% and 56.1% of shared variance, respectively).

Lexical Aspects of Comprehensibility

Rating materials

To examine lexical contributions to expert and novice raters' comprehensibility judgements, the speaking materials used in the pronunciation analyses were transcribed and then rated by novice and expert raters for comprehensibility and analyzed for seven lexical variables.

Transcript-based comprehensibility analyses

Expert and novice raters

Following the same criteria used in the first analysis, we recruited five expert and five novice raters (*M*age = 29.3 years). None of these raters was involved in the investigation of the pronunciation aspects of comprehensibility. The expert raters (graduate students in applied linguistics) reported

having linguistic training and familiarity with pronunciation, vocabulary and grammar analyses, as well as a mean of 5.2 years of language teaching experience (2–10 years). The novice raters had not taken any linguistic courses nor taught language and thus had never experienced formal assessment of learner language.

Procedure

As with the previous analysis, the raters first received a brief explanation of comprehensibility (i.e. defined as effort in understanding what someone is trying to convey) from a trained research assistant (see Appendix to this chapter). Then the raters practised by evaluating three sample transcripts (not included in the main dataset), after which they proceeded to evaluate the 40 target transcripts. The transcripts were randomly presented on a computer screen through a MATLAB interface, and the raters used a free-moving slider to assess comprehensibility on a scale between 0 = 'hard to understand' and 1000 = 'easy to understand'. To ensure that the raters paid close attention to the transcripts, they were only allowed to make their judgements after spending at least five seconds reading each transcript.

Lexical analyses

Following Saito *et al.* (2015), the transcripts were analyzed for five lexical variables using the *Coh-Metrix* software (Graesser *et al.*, 2004) and for two additional variables (lexical appropriateness and morphological accuracy) through the coding of two trained coders. The intra-class correlations were beyond 0.90. The seven lexical variables were operationalized as follows:

(1) *Frequency* was calculated as the average frequency of vocabulary in the texts, using the word frequency scores included in the CELEX Lexical Database.
(2) *Diversity*, defined as 'the range and variety of vocabulary deployed in a text by either a speaker or a writer' (McCarthy & Jarvis, 2007: 459) was calculated using McCarthy's (2005) Measure of Textual Lexical Diversity (MTLD). MTLD derives diversity scores that are mathematically adjusted for varied text length (McCarthy & Jarvis, 2010).
(3) *Polysemy* was defined as the number of related senses in a single lexical entry. For example, 'man' has several meanings, such as 'an adult male person', 'humankind', 'husband', 'a male lover' and 'a subordinate'. Yet 'car' has fewer meanings, and these are primarily limited to either 'automobile' or 'a vehicle running on rails'.
(4) *Hypernymy* was defined as the hierarchical connections between general and specific lexical items, which facilitate the efficient processing and generalization of word knowledge. For example, words like 'transportation' and 'parents' are considered to be more general and less specific than words like 'car' and 'mother'.
(5) *Text length* was defined as the total number of words in each text.

(6) *Lexical appropriateness* was defined as the number of inaccurate and inappropriate words used, including L1 substitutions.
(7) *Morphological accuracy* was defined as the number of morphological errors including verb (i.e. tense, aspect, modality and subject-verb agreement), noun (i.e. plural usage related to countable and uncountable nouns), derivation (i.e. wrong derivational forms, such as 'surprised' instead of 'surprise'), and article (i.e. article usage in terms of finite, infinite and non-articles, and possessive determiners) errors.

Results

Analyses of rater consistency (Cronbach's alpha) revealed higher agreement for the expert ($\alpha = 0.93$) than for the novice raters ($\alpha = 0.86$). Again, because these indices exceeded the threshold of rating consistency typically assumed to be acceptable ($\alpha = 0.70$; Larson-Hall, 2010), the comprehensibility scores for each speaker were averaged across the expert and novice raters, respectively (see Table 8.4 for descriptive statistics). A comparison of the expert and novice raters' comprehensibility scores using a paired-samples *t*-test showed that the expert raters assigned significantly higher (more lenient) comprehensibility scores, compared to the novice raters ($t(39) = 3.104$, $p = 0.004$, $d = 0.23$).

We also performed correlation analyses to explore the relationship between the expert and novice raters' comprehensibility judgements and the seven lexical variables in L2 speakers' speech. As summarized in Table 8.5, comprehensibility scores were associated with the diversity, polysemy and lexical appropriateness variables for both groups of raters. However, a significant link between the morphological accuracy and comprehensibility variables was found only among the expert raters.

The four lexical variables significantly associated with comprehensibility were subsequently submitted to multiple regression analyses to examine the extent to which these variables predicted the expert and novice raters' comprehensibility ratings. Both the expert and novice raters' comprehensibility scores were equally predicted by the lexical appropriateness and diversity measures (Table 8.6). However, lexical appropriateness explained much of the variance in the expert raters' scores (71%), whereas diversity accounted for most of the variance in the novice raters' judgements (50%).

Table 8.4 Descriptive statistics for expert and novice raters' comprehensibility scores on a 1000-point scale

Speaking dimension	*Mean*	*SD*	*Range*
Comprehensibility (expert)	633	263	87–987
Comprehensibility (novice)	575	235	72–952

Note: 0 = 'hard to understand', 1000 = 'easy to understand'.

Table 8.5 Pearson correlations between the seven lexical variables and expert and novice raters' comprehensibility ratings

Lexical variable	*Comprehensibility*	
	Expert raters	*Novice raters*
Frequency	0.25	0.38
Diversity	0.55*	0.47*
Polysemy	0.57*	0.49*
Hypernymy	0.20	0.04
Text length	0.20	0.03
Lexical appropriateness	0.84*	0.71*
Morphological accuracy	0.52*	0.39

Note: *$p < 0.01$ (Bonferroni adjusted).

Table 8.6 Results of multiple regression analyses using lexical variables as predictors of comprehensibility

Predicted variable	*Predictor variable*	*Adj. R^2*	*R^2 change*	*F*	*p*
Comprehensibility (expert)	Appropriateness	0.71	0.71	63.47	0.001
	Diversity	0.77	0.06		
Comprehensibility (novice)	Diversity	0.50	0.50	50.78	0.001
	Appropriateness	0.64	0.14		

Note: The variables entered into the regression included diversity, polysemy, lexical appropriateness and morphological accuracy; no evidence of strong collinearity was found (*VIF* < 1.259).

Discussion

The current study was designed to examine whether and to what degree expert and novice raters (i.e. raters with linguistic and pedagogic backgrounds versus raters without professional experience in L2 classroom teaching) perceive the comprehensibility of L2 speech as a function of its pronunciation and lexical content. The global analyses showed that the expert raters assigned higher (more lenient) comprehensibility scores than the novice raters when evaluating both audio samples and transcripts of speech. These findings are in line with previous L2 speech research which shows that raters with L2 teaching experience and/or enhanced familiarity with particular L2 accents tend to be more lenient in their assessments of L2 speech relative to untrained teachers who have less exposure to accented speech (e.g. Isaacs & Thomson, 2013; Kennedy & Trofimovich, 2008; Winke *et al.*, 2013).

Additionally, the pronunciation and lexical analyses of L2 speech revealed that the expert and novice raters attended to overlapping yet somewhat distinct linguistic dimensions of L2 speech in rating comprehensibility. With respect to pronunciation variables, both expert and novice raters similarly relied on acoustic-phonetic information in L2 speech, in this case prioritizing the prosodic factor (word stress) over segmental accuracy (Crowther *et al.*, 2015a; Derwing & Munro, 2009; Derwing *et al.*, 1998; Field, 2005; Isaacs & Trofimovich, 2012; Kang *et al.*, 2010). With respect to lexical variables, the two sets of raters also seemed to attend to comparable domains of L2 vocabulary use, such as diversity (Koizumi & In'nami, 2012), polysemy (Crossley *et al.*, 2010), lemma appropriateness (Crossley *et al.*, 2015) and morphological accuracy (Yuan & Ellis, 2003). However, the relative weights of these lexical influences differed between the expert and novice raters. Unlike the novice raters, whose comprehensibility judgements were primarily linked to lexical diversity, the expert raters attended not only to how many different words L2 speakers used but also to whether they used them in a contextually appropriate manner.

In essence, these findings support Saito and Shintani's (2016) suggestion that more experienced raters' leniency towards L2 speech may be attributed to their sensitivity to, in particular, lexical content of L2 speech. More specifically, the expert raters seem to make a greater effort to understand what L2 speakers intend to convey, at least in terms of the lexical composition of speech, perhaps despite the fact that some of the spoken words are used contextually and conceptually inappropriately. In contrast, the less experienced raters appear to attend to surface-level L2 lexical characteristics such as lexical diversity, and focus less on the appropriateness of word use, which would make understanding of L2 speech more effortful. This difference in rater behaviour could be attributed to the expert raters' L2 teaching experience as language teachers, as well as to their expertise in applied linguistics (Isaacs & Thomson, 2013).

Implications for Second Language Assessment

The findings in this study offer several implications for rater training, particularly in high-stakes assessment contexts targeting the evaluation of L2 proficiency, where all raters should have an understanding of possible sources of rater bias to minimize individual differences among potentially heterogeneous raters (Xi & Mollaun, 2011). For example, raters with little linguistic or teaching experience could be informed that experienced raters judge L2 speech by attending not only to form (pronunciation and diversity), but also to meaning (appropriateness of word use). Based on previous research targeting listener recognition of L2 speech (Bradlow & Bent, 2008), it is possible that exposing raters with little linguistic or teaching experience to a

variable, diverse set of L2 speech (e.g. in terms of accents, speech rates or proficiency levels) can improve rater consistency in speech assessment, particularly with respect to L2 comprehensibility.

Since successful L2 communication can (and should be) treated as a consequence of joint action between the speaker and the listener (Jenkins, 2000; Levis, 2005), it is noteworthy that much research to date has largely focused on the L2 speaker, highlighting problematic areas in need of improvement in terms of their pronunciation. Relatively few studies have examined how listeners should accommodate their listening strategies to better understand accented L2 speech (see Derwing *et al.*, 2002; Jenkins, 2000; Kang *et al.*, 2014). Assuming that listeners' assessments of L2 accent are largely based on pronunciation aspects of L2 speech, while their evaluations of L2 comprehensibility draw on a variety of linguistic dimensions (Crowther *et al.*, 2015b; Isaacs & Trofimovich, 2012; Saito *et al.*, 2016), raters might need to be told that L2 comprehensibility ratings capture listeners' ability to extract word- and discourse-level meaning from L2 speech. To avoid being distracted by nonnative pronunciation patterns, raters might also need to be made aware of perceptually salient characteristics of L2 speech which do not necessarily hinder understanding. As such, raters can focus on evaluating the comprehensibility of their speech without penalizing L2 speakers for their nonnative-like use of phonological features with little communicative value (Derwing & Munro, 2009), such as segments with low functional load (Munro & Derwing, 2006), *schwa* insertion in complex syllables (Lin, 2003), and monotonous (but not necessarily erroneous) prosody (Jenkins, 2000).

Limitations

Due to the exploratory nature of this study, several limitations need to be acknowledged. One obvious limitation is the small sample size of L2 speakers (40) limited to a single linguistic background (French), and native-speaking raters (10 for audio- and transcribed-based comprehensibility analyses, respectively). Secondly, this study focused on only one rater characteristic, namely raters' experience with L2 assessment through graduate-level linguistic training and/or language teaching. Thus, it would be important to examine how other rater background variables, such as the amount of familiarity with L2 speech (Kennedy & Trofimovich, 2008) and L2 learning background (Winke *et al.*, 2013), can influence raters' sensitivity to linguistic information during L2 speech assessment. Thirdly, the current findings were based on raters' judgements of L2 speech elicited via a single task (picture description). Because the same L2 users' speaking performance tends to vary (e.g. in terms of linguistic complexity, accuracy and fluency) across tasks (Robinson, 2011), future research needs to examine how rater experience influences the assessment of L2 speech elicited under various task conditions, including the

availability of planning time (Yuan & Ellis, 2003), task repetition (Ahmadian & Tavakoli, 2011), story complexity (Tavakoli & Foster, 2010), and the presence or absence of an interlocutor (Crowther *et al.*, 2015a). Finally, the study only relies on quantitative data and, thus, is not able to probe rater perceptions and triangulate these with correlations between listener-coded measures and the scores they assign, as in the Isaacs and Thomson's (2013) study. Also, it is unclear whether the measures that were identified for the study are actually those that raters attend to during normal operational ratings in research contexts.

Conclusion

The current study investigated the role of rater experience in listener-based judgements of L2 comprehensibility, focusing on two groups of native-speaking raters with and without classroom teaching experience. Results showed that expert raters (graduate students in applied linguistics and teaching professionals) provided more lenient comprehensibility ratings than novice raters. Secondly, the study demonstrated that raters with and without professional experience in L2 teaching and (by implication) experience in assessment were both similar and different in the extent to which they relied on various linguistic dimensions of L2 pronunciation in relation to comprehensibility. While both expert and novice raters processed pronunciation information in a comparable fashion (by drawing particularly on prosody), they revealed different patterns of behaviour with regard to lexical dimensions of speech. For novice raters, comprehensibility was linked to the number of different words used by L2 speakers; for expert raters, comprehensibility was largely associated with the appropriateness of word use. Building on previous comprehensibility research (e.g. Derwing & Munro, 2009; Isaacs & Trofimovich, 2012) as well as rater-focused studies (e.g. Saito & Shintani, 2016; Winke *et al.*, 2013), the current findings highlight the importance of studying the complex relationship between rater background, linguistic composition of speech, and L2 comprehensibility, with the goal of improving both the success of L2 communication and a better understanding of the linguistic constructs being measured in order to enhance the validity of the assessment.

References

Ahmadian, M.J. and Tavakoli, M. (2011) The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research* 15, 35–59.

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

Crossley, S.A., Salisbury, T. and Mcnamara, D.S. (2010) The development of polysemy and frequency use in English second language speakers. *Language Learning* 60, 573–605.

Crossley, S.A., Salsbury, T., Mcnamara, D.S. and Jarvis, S. (2011) What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly* 45, 182–193.
Crossley, S.A., Salsbury, T. and Mcnamara, D.S. (2015) Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics* 36, 570–590.
Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015a) Does speaking task affect second language comprehensibility? *Modern Language Journal* 99, 80–95.
Crowther, D., Trofimovich, P., Saito, K. and Isaacs, T. (2015b) Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly* 49, 814–837.
Derwing, T. and Munro, M. (1997) Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition* 12, 1–16.
Derwing, T. and Munro, M. (2005) Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly* 39, 379–397.
Derwing, T.M. and Munro, M.J. (2009) Putting accent in its place: Rethinking obstacles to communication. *Language Teaching* 42, 476–490.
Derwing, T., Munro, M. and Wiebe, G. (1998) Evidence in favor of a broad framework for pronunciation instruction. *Language Learning* 48, 393–410.
Derwing, T.M., Munro, M.J. and Rossiter, M.J. (2002) Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingualism and Multicultural Development* 23, 245–259.
Derwing, T.M., Rossiter, M.J., Munro, M.J. and Thomson, R.I. (2004) L2 fluency: Judgments on different tasks. *Language Learning* 54, 655–679.
Derwing, T.M., Munro, M.J., Foote, J.A., Waugh, E. and Fleming, J. (2014) Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning* 64, 526–448.
Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.
Graesser, A.C., Mcnamara, D.S., Louwerse, M.M. and Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36, 193–202.
Hahn, L.D. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38, 201–223.
Isaacs, T. and 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. and 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.
Kang, O., Rubin, D. and Pickering, L. (2010) Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *Modern Language Journal* 94, 554–566.
Kang, O., Rubin, D. and Lindemann, S. (2014) Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly* 49, 681–706. doi:10.1002/tesq.192.
Kennedy, S. and Trofimovich, P. (2008) Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64, 459–489.
Koizumi, R. and In'nami, Y. (2012) Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System* 40, 554–564.

Larson-Hall, J. (2010) *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

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

Lin, Y. (2003) Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics* 24, 439–464.

McCarthy, P.M. (2005) An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Unpublished PhD dissertation, University of Memphis.

McCarthy, P.M. and Jarvis, S. (2007) vocd: a theoretical and empirical evaluation. *Language Testing* 24, 459–488.

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

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

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

Robinson, P. (2011) *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. Amsterdam: John Benjamins.

Saito, K. and 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. and Isaacs, T. (2015) Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 29 September. doi:10.1093/applin/amv047.

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

Tavakoli, P. and Foster, P. (2010) Task design and second language performance: The effect of narrative type on learner output. *Language Learning* 58, 439–73.

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

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

Xi, X. and Mollaun, P. (2011) Using raters from India to score a large-scale speaking test. *Language Learning* 61, 1222–1255.

Yuan, F. and 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 and Onscreen Labels for Comprehensibility Judgement

Training script

Comprehensibility 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.

Onscreen labels