Studies in Second Language Acquisition, 2016, **38**, 677–701.
doi:10.1017/S0272263115000297

LEXICAL PROFILES OF COMPREHENSIBLE SECOND LANGUAGE SPEECH

The Role of Appropriateness, Fluency, Variation, Sophistication, Abstractness, and Sense Relations

Kazuya Saito
Birkbeck, University of London

Stuart Webb
University of Western Ontario

Pavel Trofimovich
Concordia University

Talia Isaacs
University of Bristol

This study examined contributions of lexical factors to native-speaking raters' assessments of comprehensibility (ease of understanding) of second language (L2) speech. Extemporaneous oral narratives elicited from 40 French speakers of L2 English were transcribed and

We are grateful to the *SSLA* reviewers for their constructive feedback on an earlier version of the manuscript and to George Smith and Ze Shen Yao for their help with data collection and analyses. The project was funded by the Grant-in-Aid for Scientific Research in Japan (No. 26770202).

Correspondence concerning this article should be addressed to Kazuya Saito, Birkbeck, University of London, The Department of Applied Linguistics and Communication, 30 Russell Square, London WC1B 5DT, UK. E-mail: k.saito@bbk.ac.uk

evaluated for comprehensibility by 10 raters. Subsequently, the samples were analyzed for 12 lexical variables targeting diverse domains of lexical usage (appropriateness, fluency, variation, sophistication, abstractness, and sense relations). For beginner-to-intermediate speakers, comprehensibility was related to basic uses of L2 vocabulary (fluent and accurate use of concrete words). For intermediate-to-advanced speakers, comprehensibility was linked to sophisticated uses of L2 lexis (morphologically accurate use of complex, less familiar, polysemous words). These findings, which highlight complex associations between lexical variables and L2 comprehensibility, suggest that improving comprehensibility requires attention to multiple lexical domains of L2 performance.

Many researchers investigating the development of second language (L2) speaking have emphasized the importance of setting realistic goals for learners, such as prioritizing being understandable to listeners over nativelikeness, to enable learners to communicate successfully in academic and business settings (e.g., Derwing & Munro, 2009; Levis, 2005). The construct of L2 oral ability can be defined as a componential phenomenon encompassing various linguistic domains including pronunciation, fluency, vocabulary, and grammar (De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). Despite the multidimensional nature of speaking, only some of its components have been extensively researched. For instance, several studies have examined which pronunciation and fluency aspects of L2 oral production relate to rater-based measures such as comprehensibility or ease of understanding (e.g., Isaacs & Trofimovich, 2012; Kang, Rubin, & Pickering, 2010). By comparison, little research has focused on lexical characteristics of L2 speech especially from the perspective of comprehensibility. Therefore, the goal of this study was to extend previous research on L2 comprehensibility by targeting a range of lexical measures in L2 speech, including lexical appropriateness, fluency, variation, sophistication, abstractness, and sense relations. The overall intent was to identify which lexical aspects of L2 speech are associated with different levels of oral ability, defined in terms of comprehensibility of L2 speech for native-speaking raters.

BACKGROUND

Vocabulary and Speaking

In the field of SLA, lexical knowledge is central to theoretical views of speaking ability, such as psycholinguistic models of L2 production

(De Bot, 1996; Kormos, 2006), and to practical issues of language learning, with patterns of vocabulary use linked to learners' speaking ability (Schmitt, 2008). However, as recently noted by Koizumi (2012, p. 1), "empirical studies on vocabulary and speaking proficiency are limited in scope." Indeed, most vocabulary research has exclusively focused on L2 listening (instead of speaking) using frequency-based analyses and examining the percentage of words needed for learners to achieve a certain level of comprehension of oral texts (van Zeeland & Schmitt, 2013) or investigating the number of word families that constitute various genres of spoken discourse such as daily conversations (Adolphs & Schmitt, 2003) or movies (Webb & Rodgers, 2009). For example, it has been shown that the knowledge of 3,000 to 4,000 of the most frequent word families may enable learners to reach the threshold for successful aural comprehension (Nation & Webb, 2011).

To date, several studies have examined various lexical aspects of L2 speech with the goal of understanding how they interact to impact native speakers' judgments of speaking ability (Crossley & McNamara, 2013; Crossley, Salsbury, & McNamara, 2014; Crossley, Salsbury, McNamara, & Jarvis, 2011; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lu, 2012). In this line of work, lexical profiles of L2 speech have been analyzed using six broad domains of word knowledge: (a) appropriateness (i.e., how accurately words are chosen and used), (b) fluency (i.e., how many words are produced per unit of speaking time), (c) variation (i.e., how many different words are produced), (d) sophistication (i.e., how many infrequent and unfamiliar words are used), (e) abstractness (i.e., how many abstract words are used), and (f) sense relations (i.e., how often polysemous words with multiple senses are used). For example, Iwashita et al. (2008) focused on lexical fluency and variation characteristics of L2 learners' Test of English as a Foreign Language internet-based test (TOEFL iBT) speaking test performance. Both sets of variables predicted native-speaking raters' judgments of five different levels of L2 speaking proficiency (advanced to beginner). In another study, Lu (2012) computationally analyzed L2 oral narratives for 25 lexical measures. Native-speaking raters' judgment of L2 speaking proficiency (ranging from excellent to fail) was mainly predicted by lexical variation (e.g., type-token ratio) and, to a lesser degree, by fluency (e.g., text length, speech rate), with no link found between proficiency rating and any lexical sophistication factors (e.g., ratio of infrequent words).

It is noteworthy, however, that neither of these two studies disentangled the effects of phonology and fluency variables (e.g., rates of segmental substitutions or frequency of pausing) from the effects of lexical variables on rater judgments of L2 speaking. For instance, even at advanced proficiency levels, at which learners' speech may feature accurate use of sophisticated and diverse vocabulary, pronunciation errors and

dysfluencies may negatively impact listeners' impressions of L2 ability. To sidestep this limitation, Crossley et al. (2014) had raters evaluate the overall proficiency of L2 oral production by rating transcriptions of learner speech using holistic rubrics adapted from the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines for speaking and writing (ranging from high to low proficiency). The resulting transcript-based ratings were associated with 5 out of 10 lexical variables, which included measures of appropriateness, diversity, frequency, imageability, concreteness, and hypernymy (see also Crossley & McNamara, 2013; Crossley et al., 2011).

Though revealing, these findings need to be interpreted with caution. One reason for this is that the raters in previous studies received training on how to categorize beginner, intermediate, and advanced levels of L2 speaking proficiency following prescribed rubrics from specific tests (e.g., TOEFL iBT, ACTFL). Therefore, it is possible that at least some lexical variables may have factored into raters' judgments simply because these variables were part of the assessment rubrics (such as the use of appropriate and diverse vocabulary) or because they were emphasized during rater training. To extend this line of L2 vocabulary and speaking research, the current study approached the same topic from a different angle—namely, by targeting judgments of L2 oral ability that are typical of assessments made by listeners in everyday communicative settings. Put simply, the current study investigated which lexical dimensions of L2 oral performance are associated with native speakers' intuitive judgments of L2 speech, such as perceived comprehensibility, in the absence of explicit rater training or the use of assessment rubrics associated with a particular testing system.

Human Ratings of L2 Speech

There is a long-standing tradition in L2 speech research to use human ratings as measures of various aspects of L2 oral production. For example, by using simple 7- or 9-point Likert-type scales, raters can reliably judge various linguistic domains of L2 speaking performance, including the quality of vowels and consonants (Piske, MacKay, & Flege, 2001), global aspects of L2 speech, such as comprehensibility and accent (Isaacs & Trofimovich, 2012), as well as fluency characteristics of L2 speech (Bosker, Pinget, Quené, Sanders, & de Jong, 2013; Derwing, Rossiter, Munro, & Thomson, 2004). What is common to this research is that raters generally show high interrater reliability (e.g., Cronbach alpha > .8–.9), suggesting that native speakers have an internalized notion of what constitutes proficient L2 speech and are able to achieve consensus in

rank ordering L2 speakers' ability without receiving much training and without using detailed assessment rubrics.

However, scalar ratings of L2 speech are rare in vocabulary and grammar studies, in which learners' oral production is mostly examined through lexical profiling and linguistic coding (e.g., Foster, Tonkyn, & Wigglesworth, 2000), using such variables as accuracy (e.g., number of error-free clauses) and complexity (e.g., ratio of subordinate clauses). To address this gap in the literature, Isaacs and Trofimovich (2012) recently examined the extent to which listeners can use Likert-type rating scales to evaluate not only phonological (segmentals, prosody) and temporal (speech rate) dimensions of L2 speech but also its lexical (appropriateness, richness) and grammatical (accuracy, complexity) characteristics in their evaluations of picture descriptions produced by native French speakers of English (beginner to advanced levels). Raters' intuitive ratings of vocabulary and grammar were found to be internally consistent and also closely related to relevant linguistic properties of oral narratives measured through acoustic and corpus analyses, suggesting that rating scales focusing on various aspects of speech represent a reliable and easy-to-use method of evaluating L2 oral performance.

Although these findings are promising, several methodological shortcomings need to be addressed before definitive conclusions can be reached. The most important shortcoming is that the L2 oral narratives targeted by Isaacs and Trofimovich (2012) were relatively short in length (about 50 words) and may, thus, have been insufficient for robust lexical analyses. Although short samples are appropriate for analyses of phonology (e.g., 15–30 s of speech in Derwing & Munro, 2009), they may be inadequate for lexical analyses, for which the threshold of minimum text length is established for certain domains in L2 vocabulary research (e.g., 100+ words for the diversity analysis; Koizumi & In'nami, 2012). In addition, Isaacs and Trofimovich's research featured a limited set of lexical measures, involving only accuracy (ratio of lexical errors), fluency (token frequency), and variation (type frequency).

To summarize, further research is needed to examine the contribution of multiple lexical variables to rater-based L2 speaking proficiency, especially because the results of previous research may have been influenced by raters relying on preexisting L2 proficiency descriptors from TOEFL (Iwashita et al., 2008) and ACTFL proficiency guidelines (Crossley et al., 2011). More importantly, the vocabulary-L2 proficiency link needs to be examined using measures of L2 ability that are more reflective of the judgments made by interlocutors communicating with learners, as compared to ratings assigned by trained raters. Therefore, to examine how multiple lexical characteristics of L2 speech contribute to human ratings of L2 speaking performance, the present study targeted comprehensibility as one dimension of L2 ability.

The Current Study

Comprehensibility, which refers to raters' impressionistic judgments about how easy or difficult it is for them to understand L2 speech, may be particularly useful as a measure of L2 speaking ability. A focus on comprehensibility allows researchers to move away from broad definitions of L2 speaking inherent in some assessments of oral performance, such as TOEFL or ACTFL proficiency guidelines, to focus on listeners' perceived effort in understanding a message. Comprehensibility is also consistent with views that posit that L2 skills must be defined independently from native-speaker norms (Cook, 2002; Jenkins, 2000). Indeed, it is comprehensible L2 speech, rather than nativelike or accent-free L2 oral performance, that is important for successful communication, given that even a substantial degree of accent is not necessarily detrimental to listener understanding (Derwing & Munro, 2009). In addition, the construct of comprehensibility is central to interactionist views of L2 development, which propose that learners make conscious or intuitive efforts to modify or repair nontarget utterances when faced with communication breakdowns, thereby making them more comprehensible to their interlocutors (Gass & Mackey, 2006; Long, 1996). Arguably, learners improve their L2 oral ability through negotiation for meaning as a way of promoting understanding in interaction. For instance, in Derwing and Munro's (2013) longitudinal study of L2 speaking, learners showed improvement in their oral performance after 7 years of residence in an English-speaking environment when their speaking was assessed through ratings of comprehensibility rather than nativelikeness, suggesting that learners may have selectively focused on aspects of language linked to interlocutor understanding such as adequate and varied prosody (Trofimovich & Baker, 2006) and proper lexicogrammar usage (Saito, 2015).

It may also be advantageous to target comprehensibility to understand how multiple lexical variables contribute to rater-based L2 speaking performance. This is because previous research on comprehensibility has chiefly focused on the phonology and fluency dimensions of L2 oral production. For instance, scalar ratings of comprehensibility appear to be associated with prosody (Kang et al., 2010) and segmental errors (especially those with high functional load; Munro & Derwing, 2006) and with pausing frequency and speaking rate (Derwing et al., 2004). There is also mounting evidence that comprehensibility is related to grammatical accuracy in L2 speech (Munro & Derwing, 1999; Saito, Trofimovich, & Isaacs, 2015) such that understanding is compromised when listeners are exposed to ungrammatical utterances (Varonis & Gass, 1982). However, it has yet to be determined which lexical variables in learner speech (e.g., lexical appropriateness, fluency, variation,

sophistication, abstractness, and/or sense relations) feed into listener perceptions of comprehensible L2 speech.

Therefore, the current study was conceptualized as a detailed investigation of lexical characteristics of comprehensible L2 speech. To address this goal, the original dataset from Isaacs and Trofimovich (2012) was revisited by targeting full-length extemporaneous oral narratives appropriate for robust lexical analyses (see the "Results" section). To control for the influence of pronunciation- and fluency-related variables, the narratives were transcribed and subsequently used for comprehensibility ratings and lexical analyses. In line with previous L2 vocabulary research, 12 measures encompassing different domains of lexical usage were examined including appropriateness (lemma, morphology), fluency (text length, filler ratio), variation (type-token ratio), sophistication (frequency, familiarity), abstractness (hypernymy, concreteness, imageability, meaningfulness), and sense relations (polysemy). The analyses, whose aim was to clarify which lexical aspects of L2 speech are associated with different levels of comprehensibility, were guided by the following two research questions:

1. Which lexical aspects of L2 speech are associated with raters' intuitive judgments of comprehensibility?
2. How do lexical correlates of comprehensibility differ as a function of speakers' comprehensibility level?

METHOD

L2 Speakers

The speakers were 40 native French speakers of L2 English (27 females, 13 males) from Quebec, Canada (M_{age} = 35.6 years, range = 28–61). All speakers started learning English in elementary school, except two early French-English bilinguals. At the time of the study, the speakers estimated using English to varying degrees (0–70% of the time daily) and reported a full range of self-rated English ability (1–9) in speaking, listening, reading, and writing using a 9-point scale (1 = *extremely poor*, 9 = *extremely proficient*). To ensure that the speakers represented various levels of L2 speaking ability, their oral performance was screened through a paragraph reading task (440 words) using several measures, which included perceived nativelikeness (foreign accent), segmental accuracy (pronunciation of /ð/, as in *brother*, a difficult consonant for French speakers), and fluency (articulation rate). In terms of accent ratings, the speakers' reading aloud was evaluated by 10 native-speaking judges using a 9-point scale (1 = *heavily accented*, 9 = *not accented at all*), with pooled scores across raters ranging between 1.8 and 9.0.

With respect to /ð/ production, the speakers' accuracy varied between a low of 7% and a high of 99% correct. Finally, in terms of articulation rate (total number of syllables, including repetitions and hesitations, divided by total sample duration), the speakers' output ranged between 0.4 and 3.4 syllables per second. Thus, the speakers represented a range of L2 speaking ability, from beginning to advanced.

Oral Narratives

Following earlier L2 speech research (e.g., Derwing & Munro, 2009), extemporaneous speech was elicited using a picture description task. The speakers described an eight-image picture sequence about two strangers bumping into each other on a busy street corner and inadvertently switching their suitcases, which were identical in appearance (Derwing et al., 2004). Whereas the previous study by Isaacs and Trofimovich (2012) focused only on the first 30 s of the recorded picture narratives to investigate the relationship between L2 comprehensibility and phonology and fluency variables, the present study targeted full-length recordings to analyze the same speech samples for a number of lexical characteristics. The narratives in the current investigation varied widely in duration (M = 2 min 26 s, range = 55 s–5 min 51 s). All but two samples exceeded the suggested threshold (i.e., 100 words) in terms of word length for the diversity analysis (Koizumi & In'nami, 2012).

These two shorter recordings, which were 75 and 81 words long, came from the speakers who appeared to have difficulty producing more than 100 words due to their limited linguistic abilities; however, the two samples were of sufficient length (1 min 15 s and 3 min 10 s). As such, these two samples represented the lower range of L2 oral ability, especially in the context of the picture task used in the present study, whereby those with limited linguistic knowledge had difficulty producing more than 100 words and made a number of filled and unfilled pauses. Because the goal of this study was to analyze L2 oral narratives spanning a wide range of L2 speaking ability, these samples were included in the final dataset. Although the inclusion of these two shorter samples may have contradicted Koizumi and In'nami's (2012) suggestion of using a minimum of 100 words per narrative, the extensive range of obtained narratives (75–485 words, corresponding to 55 s–5 min 51 s of speaking time) allowed us to examine how various other lexical variables, such as appropriateness, fluency, abstractness, and semantic relations, interact to affect raters' intuitive judgments of comprehensibility. The mean narrative length in the final dataset was 209.2 words (SD = 90).

Comprehensibility Analysis

The 40 oral narratives were rated for comprehensibility by 10 native speakers of English. The raters were all born and raised in English-speaking homes in Canada, with at least one parent being a native English speaker. The raters estimated using English more than 90% of the time daily and, as residents of Montreal (a bilingual French-English city), all reported high familiarity with French-accented English speech. Keeping rater familiarity with L2 speech constant (i.e., at high levels) was important because this factor can impact rater behavior (see Winke, Gass, & Myford, 2013).[1] The raters participated in individual rating sessions to evaluate L2 oral narratives for comprehensibility.

When L2 comprehensibility studies rely on native speakers to listen to and judge L2 speech samples for comprehensibility (e.g., Derwing & Munro, 2009), they often examine how various domains of language (e.g., pronunciation or fluency) relate to listener assessment (Isaacs & Trofimovich, 2012). Because the primary goal of this study was to examine how lexical (rather than pronunciation or fluency) factors influence native speakers' comprehensibility judgment, our raters read transcribed L2 speech, as opposed to listening to it. This methodological decision was made following previous SLA research standards that allow researchers to investigate lexical correlates of L2 speech ratings with pronunciation removed as a possible confound (Crossley et al., 2014; Crossley et al., 2011; Patkowski, 1980).

The recordings were transcribed, and the resulting transcripts were edited to remove spelling clues signaling pronunciation-specific errors (e.g., *hit*, although pronounced as *heat*, was still spelled as *hit*) and punctuation to avoid transcriber influence (Ochs, 1979). The raters received a brief explanation of comprehensibility—namely, that it refers to perceived effort in understanding what a language user is trying to convey (for training scripts and onscreen labels, see the Appendix).

In the rating sessions, each written transcript was presented on a computer screen one at a time in a unique, randomized order by way of a custom software, Z-Lab (Yao, Saito, Trofimovich, & Isaacs, 2013), developed using commercial software package (MATLAB [Version 8.1]). The raters used a free-moving slider, shown below the transcript, to assess the comprehensibility of each narrative. If the slider was placed at the leftmost end of the continuum, labeled with a frowning face (indicating the negative endpoint), the rating was recorded as zero. If the slider was placed at the rightmost end of the continuum, labeled with a smiley face (indicating the positive endpoint), the rating was recorded as 1,000. The raters were told that the narratives came from French speakers of L2 English representing a range of speaking ability, and they were encouraged to use the entire scale as much as possible. To ensure

that the raters carefully read each transcript, they were allowed to record their rating only after spending at least 5 s on each transcript. Before proceeding to rate the 40 transcripts, the raters performed a practice session consisting of three example transcripts drawn from the same population of L2 speakers.

Lexical Analyses

The 40 oral narratives were analyzed for 12 lexical variables: appropriateness (lemma, morphology), fluency (text length, filler ratio), variation, sophistication (frequency, familiarity), abstractness (hypernymy, concreteness, imageability, meaningfulness), and sense relations (polysemy). Whereas trained coders used the original transcripts to conduct appropriateness and fluency analyses, measurements of variation, sophistication, abstractness, and sense relations were carried out through the Coh-Metrix software (McNamara, Graesser, McCarthy, & Cai, 2014) using modified transcripts with French substitutions and fillers removed.

Appropriateness. Building on previous literature (e.g., Yuan & Ellis, 2003), two measures of lexical appropriateness were used. The first measure was lemma appropriateness, defined as the number of contextually and conceptually inappropriate words (including French substitutions) over the total number of words. Thus, all inappropriately used words (e.g., *walkside* [for sidewalk]) and French substitutions (e.g., *malette* [for suitcase], *ah mon Dieu les temps en plus* "Oh my God now and more")[2] were counted as lemma errors. The second measure was morphological appropriateness, computed as the number of morphological errors over the total number of words. These errors were related to verbs (i.e., tense, aspect, modality, and subject-verb agreement), nouns (i.e., plural usage related to count and noncount nouns), derivations (i.e., wrong derivational forms such as *confused* instead of *confuse*), articles (i.e., article usage in terms of definite, indefinite, and nonarticles), and possessive determiners (*her suitcase* instead of *his suitcase*). All 40 transcripts were first coded by a trained coder, and then another trained coder recoded 10 randomly chosen transcripts (i.e., 25% of all transcripts). The resulting intraclass correlations showed high consistency for both lemma (r = .97) and morphological (r = .88) appropriateness.

Fluency. Because lexical fluency refers both to how many words are produced and how effortlessly they are articulated (i.e., without undue pauses and hesitations), two fluency measures were computed. The first

measure was text length, defined as the total number of words in each narrative (Iwashita et al., 2008; Lu, 2012). The second measure was filler ratio, defined as the total number of fillers (e.g., *uh*, *ah*, *oh*) over the total text length (Lennon, 1990).

Variation. Lexical variation captures the diversity of words in a text. Although lexical variation is typically defined as the number of different words produced by a speaker or writer (e.g., type-token ratio), such measures are highly dependent on text length (with longer texts associated with lower values). Therefore, more accurate measures of lexical diversity, such as the measure of textual lexical diversity (MTLD), involve indices that are mathematically transformed to account for text length (McCarthy & Jarvis, 2010). In this study, lexical diversity was defined as the MTLD and was derived through Coh-Metrix (McNamara et al., 2014). Koizumi and In'nami (2012) considered the MTLD an appropriate measure of lexical variation, especially for oral texts of 100–200 words.

Sophistication. Lexical sophistication refers to the number of unusual or advanced words used by a speaker or writer (Read, 2000). In L2 vocabulary research, lexical sophistication is measured objectively through corpus-based lexical profiling (i.e., word frequency; Laufer & Nation, 1995) and also subjectively through native speakers' estimates of how commonly a given word is experienced (i.e., familiarity; Stadthagen-Gonzalez & Davis, 2006). In line with prior research, both frequency and familiarity indices of lexical sophistication were computed to determine the extent to which less common and more advanced words were used in oral narratives.

The first measure was *word frequency*, defined as the average frequency of all words in each narrative and derived through Coh-Metrix (McNamara et al., 2014) from the CELEX corpus of English (Baayen, Piepenbrock, & Gulikers, 1995). Word frequency may help differentiate output produced by learners of varying ability levels (Crossley et al., 2014; Crossley et al., 2011; Laufer & Nation, 1995).

The second measure of lexical sophistication was *word familiarity*, which refers to how commonly a word is experienced. Native speakers tend to report more familiarity with words like *window*, *city*, and *room* than *floor*, *direction*, and *tie*. Familiarity scores, derived for content words through Coh-Metrix from the Medical Research Council psycholinguistics database (Wilson, 1988), consisted of native speakers' subjective judgments using 7-point scales (1 = w*ord never seen*, 7 = *word seen every day*). Word familiarity may capture the extent to which learners encounter words through L2 experience (Schmitt & Meara, 1997) and may explain changes in word use as learners' L2 proficiency increases (Salsbury, Crossley, & McNamara, 2011).

Abstractness. Second language lexical use can be conceptualized from a developmental perspective that captures the extent to which abstract words are used (Crossley et al., 2011). Second language users may demonstrate enhanced lexical knowledge through their use of vocabulary that differs along the dimensions of hypernymy, concreteness, imageability, and meaningfulness, which represented the four measures of lexical abstractness computed for each oral narrative using Coh-Metrix (McNamara et al., 2014).

The category of *hypernymy* refers to hierarchical connections between general and specific lexical items that facilitate efficient processing and generalization of word knowledge. For example, words like *building* and *color* are considered to be more general and less specific than words like *library* or *hotel* and *green* or *red*. Second language learners tend to produce less specific words as their L2 experience increases (Crossley, Salsbury, & McNamara, 2009), which contributes to raters' judgment of overall lexical proficiency (Crossley et al., 2011). More proficient L2 learners likely rely on strategies by using more general or holistic terms to compensate for specific words that they may not know or have difficulty accessing (e.g., *water* vs. *pond*; Færch & Kasper, 1984).

The category of *concreteness* is concerned with how abstract a word meaning is. Words referring to an object, material, or person (e.g., *car*, *glass*, *people*) have greater concreteness scores than words referring to more semantically abstract constructs (e.g., *week*, *life*, *problem*). Second language learners tend to learn concrete words at earlier stages and with greater ease compared to more abstract words (Crossley et al., 2009; Ellis & Beaton, 1993).

The category of *imageability* refers to how easy it is to construct a mental image of a word. For example, native speakers create visual images more easily for certain words (e.g., *woman*, *green*, *telephone*) compared to others (e.g., *appointment*, *name*, *problem*). Second language learners appear to learn more imageable words more easily than less imageable words because they can visually experience and analyze imageable words (Ellis & Beaton, 1993). Second language learners also start using less imageable words as their proficiency increases, with utterances becoming less context dependent (Salsbury et al., 2011).

The final category of *meaningfulness* refers to the extent of interconnections between a given lexical item and other words. Whereas more meaningful words (e.g., *color*, *town*, *trip*) evoke many other related words, less meaningful words (e.g., *west*, *yellow*, *office*) result in limited links. As learners' proficiency improves, they tend to increase the number of known word associations (Zareva, 2007) and start using less meaningful words with fewer word associations (Salsbury et al., 2011).

Sense Relations. This measure, computed for each oral narrative through Coh-Metrix (McNamara et al., 2014), refers to the number of

related senses words have. For example, *case* has several senses such as an instance of something (e.g., *a case in point*), the actual state of things (e.g., *that's the case*), situation (e.g., *mine is a sad case*), a small container (e.g., *a jewel case*), and a pair or couple (e.g., *a case of pistols*). However, *sidewalk* has few senses, limited to the meaning of a paved area at the side of a street in North American English. As their L2 proficiency increases, learners tend to acquire more polysemous words with the potential for multiple sense relations and ambiguity (Schmitt, 1998). Initially, learners likely focus on the core sense of a polysemous word and then gradually shift their attention toward the peripheral senses (Verspoor & Lowie, 2003). Learners ultimately come to solidify their lexical knowledge of polysemous words by using different sense relations more frequently, appropriately, and fluently (Crossley, Salsbury, & McNamara, 2010).

RESULTS

Comprehensibility

As in previous speech research (Derwing & Munro, 2009), the raters showed high interrater consistency in their comprehensibility judgments (Cronbach's alpha = .95), suggesting that native-speaking raters shared a notion of what constitutes comprehensible L2 output even though they only read transcripts and received little instruction on how to assess comprehensibility. The 10 raters' comprehensibility scores, which were deemed sufficiently consistent, were then averaged to derive a single mean score per speaker. The 40 speakers' comprehensibility scores ranged between 80 and 970 on a 1,000-point scale (M = 604, SD = 202) and were normally distributed according to a one-sample Kolmogorov-Smirnov test (p = .200).

Lexical Variables

Pearson correlational analyses were carried out first to determine which of the 12 lexical variables were correlated with comprehensibility scores (Bonferroni corrected a = .004). As shown in Table 1, 8 of the 12 lexical measures spanning all six targeted categories (appropriateness, fluency, variability, sophistication, abstractness, and sense relations) were significantly associated with comprehensibility. Their statistical power was relatively strong for the significant lexical correlates of L2 comprehensibility (.8–1).

A stepwise multiple regression analysis was conducted next to determine the extent to which the eight significant associations (lemma and morphology errors, filler ratio, MTLD, familiarity, imageability,

Table 1. Pearson correlations and statistical power between comprehensibility rating and 12 lexical variables

Lexical variable	Comprehensibility rating	*p* value	Statistical power
Appropriateness (lemma)	–.80*	< .001	1.00
Appropriateness (morphology)	–.48*	.002	.94
Fluency (text length)	.13	.432	.20
Fluency (filler ratio)	–.76*	< .001	1.00
Variation (MTLD)	.72*	< .001	1.00
Sophistication (frequency)	–.32	.041	.65
Sophistication (familiarity)	–.53*	< .001	.97
Abstractness (hypernymy)	–.13	.442	.20
Abstractness (concreteness)	–.27	.090	.52
Abstractness (imageability)	–.52*	.001	.97
Abstractness (meaningfulness)	–.57*	< .001	.99
Sense relations (polysemy)	.55*	< .001	.98

* $p < .004$ (Bonferroni corrected).

meaningfulness, polysemy) predicted L2 comprehensibility scores. Comprehensibility scores served as the dependent variable, with the eight lexical variables used as predictors (see Table 2). The regression model, which included three variables, accounted for 90.1% of the variance in comprehensibility, $F(3, 36) = 52.01$, $p < .001$, with no evidence of strong collinearity in the model (VIF < 1.87). Lemma errors (appropriateness) alone accounted for a substantial proportion of variance (63%), whereas filler ratio (fluency) and MTLD (variability) made additional contributions, explaining 11% and 6% of the variance, respectively.

Lexical Variables at Different Comprehensibility Levels

The final analysis examined which lexical variables distinguished beginner, intermediate, and advanced levels of L2 comprehensibility.

Table 2. Results of multiple regression analysis using lexical variables as predictors of l2 comprehensibility

Predicted variable	Predictor variables	Adjusted R^2	R^2 change	*F*	*p*
Comprehensibility	Lemma errors	.63	.63	68.27	< .001
	Filler ratio	.74	.11	56.29	< .001
	MTLD	.80	.06	52.01	< .001

Note. The variables entered into the regression equation included lemma and morphology errors, filler ratio, MTLD, familiarity, imageability, meaningfulness, and polysemy.

For this analysis, the 40 speakers were first divided into three nonoverlapping groups (beginner, intermediate, advanced) based on their comprehensibility scores, shown in Table 3. Then, the scores for the eight lexical variables with significant associations with comprehensibility (see Table 1) were submitted to one-way ANOVAs, with speaker group used as a between-subjects variable and Bonferroni post hoc tests carried out to explore between-group comparisons.

Table 4 summarizes the significant F ratios along with significant between-group differences. The measure of textual lexical diversity (lexical variation) distinguished between all three levels of comprehensibility such that the advanced comprehensibility group produced narratives with more lexical variation than the intermediate group (p = .006), which in turn had more lexical variation than the beginner group (p = .033). Lemma errors (appropriateness), filler ratio (fluency), as well as imageability and meaningfulness (abstractness) distinguished between beginner and intermediate levels of comprehensibility. Compared to the beginner comprehensibility group, the intermediate group produced oral narratives that included fewer lexical errors (p < .001) and fillers (p = .003) and contained a greater number of abstract words in terms of their imageability (p = .049) and their links to similar words (p = .044). Morphology errors (appropriateness) differentiated between the intermediate and advanced levels of comprehensibility, with the advanced comprehensibility group producing fewer morphology errors than the intermediate group (p = .021). Finally, word familiarity (sophistication) and polysemy (sense relations) distinguished beginner from advanced levels of comprehensibility such that the advanced comprehensibility group produced oral narratives featuring less familiar words (p = .012) and words with more polysemous senses (p = .018) compared to the beginner group.

DISCUSSION

The current study examined the relationship between lexical variables and human ratings of L2 comprehensibility (one dimension of L2 speaking proficiency) using oral picture narratives produced by French

Table 3. Summary of comprehensibility scores for three speaker groups

Group	*M*	*SD*	*Range*
Beginner (n = 13)	312	129	80–480
Intermediate (n = 14)	638	78	520–750
Advanced (n = 13)	860	62	770–970

Note. Based on comprehensibility rating (0 = *hard to understand*, 1,000 = *easy to understand*).

Table 4. Summary of group differences for beginner, intermediate, and advanced levels of comprehensibility

	ANOVA results			
Lexical variable	$F(2, 37)$	p	η_p^2	Significant group differences
Appropriateness (lemma)	21.94	< .001	.54	Beginner < Intermediate
Appropriateness (morphology)	6.92	.003	.27	Intermediate < Advanced
Fluency (filler ratio)	11.94	< .001	.39	Beginner < Intermediate
Variation (MTLD)	17.31	< .001	.48	Beginner < Intermediate < Advanced
Sophistication (familiarity)	4.88	.013	.20	Beginner < Advanced
Abstractness (imageability)	7.46	.002	.28	Beginner < Intermediate
Abstractness (meaningfulness)	8.73	< .001	.32	Beginner < Intermediate
Sense relations (polysemy)	4.29	.021	.18	Beginner < Advanced

speakers of L2 English. The study targeted a comprehensive set of 12 lexical measures (lemma and morphology errors, text length, filler ratio, MTLD, familiarity, frequency, hypernymy, concreteness, imageability, meaningfulness, and polysemy) in an attempt to identify lexical correlates of L2 speaking without using preexisting scoring rubrics or detailed descriptions of what constitutes speaking performance. Native-speaking raters indeed demonstrated a shared understanding of what constitutes L2 comprehensibility, as shown by high internal consistency of ratings, even though raters received minimal instruction about comprehensibility and evaluated oral narratives through reading, not listening.

In response to the first research question, which targeted lexical characteristics of L2 speech linked to comprehensibility, results overall suggested that lexical factors contribute to rater-based judgments of comprehensibility in multiple ways. Of the 12 targeted lexical measures, 8 were significantly associated with comprehensibility (lemma and morphology errors, filler ratio, MTLD, familiarity, imageability, meaningfulness, polysemy). These findings are in line with previous results showing that L2 speakers' lexical usage is tied to measures of speaking (e.g., Crossley et al., 2014; Crossley et al., 2011). These associations tap into different domains of lexical knowledge such as variation (Crossley et al., 2011; Koizumi & In'nami, 2012), appropriateness (Iwashita et al., 2008), fluency (Iwashita et al., 2008; Lu, 2012), sophistication (Lu, 2012), abstractness (Crossley et al., 2011), and sense relations (Crossley et al., 2009).

In terms of the relative contribution of the lexical variables to raters' overall judgment, L2 comprehensibility was mainly predicted by the appropriateness factor (lemma errors, 63%) as well as by lexical fluency (filler ratio, 11%). Additionally, some of the remaining variance, beyond what had already been explained by lemma errors and filler ratio, was related to variation (MTLD, 6%). Therefore, in assigning comprehensibility scores, raters seemed to rely primarily on the extent to which L2 speakers can select conceptually and contextually appropriate words while also taking into account the degree to which they can produce them fluently (i.e., without undue pauses and hesitations) and to which these words represent a diverse lexical set. These results are consistent with Crossley and colleagues' (2014) earlier finding that lexical appropriateness (operationalized as collocation accuracy) plays a significant role in native-speaking raters' holistic judgments of lexical proficiency in L2 speech (84% of the variance explained), with contributions of variation and sophistication being less pronounced (3–5% of the variance explained). This convergence in findings is especially interesting in light of methodological differences between the current study (focusing on untrained raters' intuitive judgments) and Crossley and colleagues' (2014) research (targeting trained raters' assessments based on specific rubrics).

Therefore, regardless of rater training procedures and rating materials used, native-speaking raters appear to assess lexical qualities of L2 speech by attending to appropriate and fluent uses of words as a primary cue and to lexical diversity and sophistication as a secondary cue. This implies that improved L2 oral ability is most strongly linked to speakers' accurate and perhaps fluent use of L2 words and (to a lesser extent) their use of lexically diverse and sophisticated vocabulary.

With respect to the second research question, which asked how lexical correlates of comprehensibility vary as a function of speakers' comprehensibility level, results showed that different lexical variables related to comprehensibility in distinct ways and that the relative weight of lexical factors varied according to speakers' comprehensibility level (beginner, intermediate, advanced). Variation (MTLD) significantly differentiated the three comprehensibility groups, suggesting that each stage of L2 comprehensibility reflects the extent to which L2 speakers can use a wide variety of words without much repetition. Yet, which types of words speakers choose (in terms of abstractness, sophistication, and sense relations) and how they use these words (in terms of fluency and lemma and morphology accuracy) may be specific to each comprehensibility level. Thus, lexical appropriateness (lemma errors), variation (MTLD), fluency (filler ratio), and abstractness (imageablity, meaningfulness) were crucial for distinguishing beginner from intermediate comprehensibility levels. When it came to advanced comprehensibility, raters seemed to attend not only to variation (MTLD) but also to morphological appropriateness (morphology errors), sophistication (familiarity), and sense relations (polysemy).

Although cross-sectional data cannot be unambiguously regarded as evidence of development, there is both a theoretical (e.g., Gass & Mackey, 2006; Long, 1996) and an empirical (e.g., Derwing & Munro, 2013; Saito, 2015) basis for arguing that adult SLA processes take place on a continuum of comprehensibility that is largely determined through learners' input and interaction opportunities with native and nonnative speakers. Given that the current dataset consisted of L2 learners with a wide range of proficiency levels (beginner to advanced), examining lexical features at different proficiency levels provides some evidence for how adult L2 learners can enhance the comprehensibility of their speech (low → mid → high) over time.

Because lexical variables were associated with the ratings for low-to-mid L2 comprehensibility learners, the beginner phase of L2 comprehensibility seems to be associated with fluent use of varied and appropriate vocabulary (Nation & Webb, 2011). In essence, the learning process appears to constitute a gradual transition from basic patterns of L2 vocabulary use (e.g., relatively fluent and accurate use of concrete words) to more sophisticated vocabulary usage (e.g., morphologically

accurate use of complex, less familiar, polysemous words). Given that even beginner-level learners with limited L2 knowledge show little difficulty acquiring words that elicit clear mental images and have strong associations with other words (Ellis & Beaton, 1993), the lexical profile of intermediate-level comprehensibility can be characterized by how much learners have control over words that are less imageable and those that feature fewer lexical associations (van Zeeland & Schmitt, 2013).

To reach higher levels of comprehensibility, however, L2 speakers may need to develop knowledge of less familiar and more polysemous words. With respect to word familiarity, this finding supports research showing that L2 learners may begin to understand less familiar words after accumulating a certain amount of L2 experience (e.g., greater than a year of residence in a L2-speaking country) and may become ready to use less familiar words in their output (Salsbury et al., 2011; Schmitt & Meara, 1997). With respect to polysemous words (i.e., complex words with multiple meanings) as a marker of high-level comprehensibility, these findings are in line with research showing that L2 learners first master the core meaning of a polysemous word and only then shift their attention to its peripheral senses (Verspoor & Lowie, 2003). Such semantic learning likely occurs only after learners have spent some time in L2-speaking contexts (e.g., longer than 4 months), which implies that the development of semantic associations between individual senses of polysemous words is a gradual process requiring extensive linguistic input and experience (Crossley et al., 2010; Schmitt, 1998). Additionally, from a psycholinguistic perspective, establishing robust knowledge of polysemous words may also be seen as an important milestone. This is because storing multiple meanings in a single lexical entry (rather than as separate entries), at least in theory, allows L2 users to efficiently manage cognitive resources involved in accessing and using words (Verspoor & Lowie, 2003).

With respect to morphological accuracy, speakers at the advanced level of L2 comprehensibility seemed to use proper morphology in a highly consistent manner with verbs, nouns, derivations, articles, and possessive determiners in their spontaneous speech. Supporting prior evidence of the important role of morphosyntactic accuracy for L2 comprehensibility (Varonis & Gass, 1982), this finding further implies that this factor may be particularly crucial at advanced levels of L2 speech learning. In essence, learners' attention to morphosyntactic form appears to be an important component of developing a targetlike L2 speaking ability, likely as a result of conversational practice in communicatively oriented classrooms and/or immersion experience in a L2-speaking environment (e.g., Lyster, 2007). However, morphosyntactic errors are usually less likely to impact L2 comprehensibility compared to pronunciation errors (Mackey, Gass, & McDonough, 2000), and

morphological markers may also be rendered perceptually nonsalient in L2 speech (Goldschneider & DeKeyser, 2001). Therefore, to achieve higher level comprehensibility, L2 speakers may need to be pushed (through either explicit teaching or extensive input and output practice) to notice and incorporate morphosyntactically accurate language forms in their speech so that their L2 production can be understood by interlocutors both accurately and efficiently (Jiang, 2007). Needless to say, more longitudinal studies are warranted to better understand how L2 learners enhance their comprehensibility over a prolonged period of time (cf. Derwing & Munro, 2013). Such future studies will, as a result, shed some light on the developmental sequence of L2 lexical proficiency as a function of the quantity and quality of input, as suggested in the present study.

One well-researched variable that did not turn out to be a significant predictor for L2 comprehensibility was word frequency. Previous research has shown that L2 learners need to increase their vocabulary size beyond the first 2,000 word families in order to understand everyday spoken discourse (e.g., van Zeeland & Schmitt, 2013) and other speech genres such as TV shows and movies (e.g., Webb & Rodgers, 2009). Yet, the link between lexical frequency and L2 comprehensibility was relatively weak in this study. One possible interpretation of this finding could be that lexical sophistication and word frequency would most likely be related to ratings of language proficiency rather than comprehensibility. In other words, perceived comprehensibility may be a construct that is essentially different from L2 speaking proficiency and lexical knowledge. Indeed, previous research has provided some evidence that native-speaking raters pay attention to lexical sophistication (frequency), especially when explicitly trained to judge L2 speaking and lexical proficiency as described in the rating rubrics of the ACTFL (Crossley et al., 2011) and TOEFL iBT (Crossley & McNamara, 2013) tests.

Another reason for the weak predictive power of lexical frequency could be ascribed to the nature of the task (i.e., describing a picture sequence). Essentially, the picture task was not designed to elicit a sufficiently wide range of infrequent lexical items. As such, even advanced L2 speakers could have completed the task successfully by using a restricted word frequency range, and native-speaking raters may not have been sufficiently sensitive to the ratio of frequent to infrequent words, focusing instead on whether speakers used frequent words appropriately and fluently to convey their intended message during the task. Importantly, the nature of the task may have also influenced several other variables such as those related to abstractness, morphological accuracy, and fluency. The picture task used here is a concrete, straightforward story that does not allow for much in the way of highly abstract lexical usage. In addition, L2 learners typically do not have much

difficulty with morphological accuracy while describing cartoon pictures depicting a relatively simple story line (Tavakoli & Foster, 2010). Finally, although text length was not a significant predictor for L2 comprehensibility, the results could be specific to the picture task, whereby advanced L2 learners can narrate the same story clearly and concisely (shorter texts do not necessarily indicate a lack of fluency). Therefore, to provide a more nuanced picture of the contribution of lexical factors (including lexical frequency, abstractness, morphological accuracy, and text length) to L2 speaking performance, future research needs to target different speaking tasks, especially cognitively complex ones, as they may reveal both strengths and weaknesses of L2 speakers' lexical and morphological knowledge (see Crowther, Trofimovich, Isaacs, & Saito, 2015; Hulstijn, Schoonen, De Jong, Steinel, & Florijn, 2012).

IMPLICATIONS AND CONCLUSIONS

The findings of this study, which examined how lexical factors influence native-speaking raters' intuitive assessments of L2 comprehensibility, have several implications for teaching practice. First, with the goal of attaining comprehensible L2 performance, learners may need to be encouraged to expand their lexical repertoires beyond highly imageable, meaningful, and familiar words. Although L2 vocabulary development likely starts from the learning of core meanings of these easier words, learners should gradually shift their attention toward acquiring less familiar words and words with multiple senses. Because comprehensibility appears to be related to fluent use of words in context, it is crucial that learners also experience relevant lexical items not only through explicit instruction and language-focused activities but also through practice in communicative tasks that help learners establish form-meaning mappings using intensive exposure to meaningful L2 input and interaction (Gatbonton & Segalowitz, 2005).

As comprehensibility is linked to multiple linguistic domains including phonology, fluency, lexicon, grammar, and discourse structure (e.g., Isaacs & Trofimovich, 2012; Kang et al., 2010; Munro & Derwing, 1999; Saito et al., 2015), pronunciation and fluency training should also be introduced in the context of vocabulary teaching. For instance, Field (2005) recommended focusing on word stress as a part of vocabulary teaching, arguing that "the responsibility for presenting [lexical stress] falls as much on the vocabulary teacher as on the pronunciation teacher, and the oral practice of new items should include attention to their stress pattern" (p. 420). Thus, to help learners acquire L2 comprehensibility efficiently and successfully, future research is needed to evaluate the pedagogical effectiveness of teaching methods that target the

learning of not only new word lemmas (core and peripheral meanings) but also their lexemes (orthographic, segmental, and suprasegmental forms) in a complimentary fashion.

Received 7 December 2014
Accepted 4 March 2015
Final Version Received 14 April 2015

NOTES

1. Although previous studies have noted that expert native speakers with professional L2 assessment experience (e.g., experienced English as a second/foreign language teachers) demonstrate different perceptions of comprehensibility compared to novice raters (Isaacs & Thomson, 2013), this rater variable was not systematically controlled in the current study. We discuss the role of L2 assessment experience in evaluating lexical correlates of L2 comprehensibility elsewhere (Saito, Trofimovich, Isaacs, & Webb, in press).

2. All of the examples were retrieved from the dataset in the study.

REFERENCES

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*, 425–438.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX*. Philadelphia, PA: Linguistic Data Consortium.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 159–175.

Cook, V. (Ed.). (2002). *Portraits of the L2 user*. Clevedon, UK: Multilingual Matters.

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

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*, 307–334.

Crossley, S. A., Salsbury, T., & 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. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amt056

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

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

De Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning*, *46*, 529–555.

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. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 476–490.

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

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

Ellis, N., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary acquisition. *Language Learning*, *43*, 559–617.
Færch, C., & Kasper, G. (1984). Two ways of defining communication strategies. *Language Learning*, *34*, 45–63.
Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399–423.
Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language. *Applied Linguistics*, *21*, 354–375.
Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA Review*, *19*, 3–17.
Gatbonton, E., & Segalowitz, N. (2005). Rethinking the communicative approach: A focus on accuracy and fluency. *Canadian Modern Language Review*, *61*, 325–353.
Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50.
Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–223.
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., & 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.
Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 29–49.
Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, UK: Oxford University Press.
Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language acquisition. *Language Learning*, *57*, 1–33.
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.
Koizumi, R. (2012). Vocabulary and speaking. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Oxford, UK: Wiley Blackwell.
Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*, 554–564.
Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.
Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417.
Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 367–377.
Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition. Vol. 2: Second language acquisition* (pp. 413–468). New York, NY: Academic Press.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Review*, *96*, 190–208.
Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam, the Netherlands: Benjamins.
Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.
MATLAB (Version 8.1) [Computer software]. (2013). Natick, MA: The Mathworks Inc.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.
Munro, M. J., & Derwing, T. M. (1999). Foreign accent comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310.
Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*, 520–531.
Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.
Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York, NY: Academic Press.
Patkowski, M. (1980). The sensitive period for the acquisition of syntax in a second language. *Language Learning*, *30*, 449–472.
Piske, T., MacKay, I., & Flege, J. (2001). Factors affecting degree of foreign accents in an L2: A review. *Journal of Phonetics*, *29*, 191–215.
Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.
Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, *65*(3).
Saito, K., Trofimovich, P., & Isaacs, T. (2015). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*. Advance online publication. doi: 10.1017/S0142716414000502
Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (in press). 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*. Bristol, UK: Multilingual Matters.
Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.
Schmitt, N. (1998). Tracking the incremental acquisition of a second language vocabulary: A longitudinal study. *Language Learning*, *48*, 281–317.
Schmitt, N. (2008). State of the art: Instructed second language vocabulary acquisition. *Language Teaching Research*, *12*, 329–363.
Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.
Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*, 598–605.
Tavakoli, P., & Foster, P. (2010). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*, 439–473.
Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30.
van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*, 609–624.
Varonis, E. M., & Gass, S. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition*, *4*, 114–136.
Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*, 547–586.
Webb, S., & Rodgers, M. P. H. (2009). The vocabulary demands of television programs. *Language Learning*, *59*, 335–366.
Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, Version 2. *Behavior Research Methods, Instruments and Computers*, *20*, 6–11.
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.
Yao, Z., Saito, K., Trofimvich, P., & Isaacs, T. (2013). Z-Lab [Computer software]. Retrieved from https://github.com/ZeshanYao/Z-Lab (Accessed on August 1, 2013).
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.
Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research*, *23*, 123–153.

APPENDIX

TRAINING MATERIALS AND ONSCREEN LABELS FOR COMPREHENSIBILITY JUDGMENT

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