Bilingualism: Language and Cognition 19 (3), 2016, 597–609 © Cambridge University Press 2015 doi:10.1017/S1366728915000255

Lexical correlates of comprehensibility versus accentedness in second language speech*

KAZUYA SAITO
Birkbeck, University of London
STUART WEBB
University of Western Ontario
PAVEL TROFIMOVICH
Concordia University
TALIA ISAACS
University of Bristol

(Received: August 10, 2014; final revision received: April 27, 2015; accepted: April 30, 2015; first published online 17 June 2015)

The current project investigated the extent to which several lexical aspects of second language (L2) speech – appropriateness, fluency, variation, sophistication, abstractness, sense relations – interact to influence native speakers' judgements of comprehensibility (ease of understanding) and accentedness (linguistic nativelikeness). Extemporaneous speech elicited from 40 French speakers of English with varied L2 proficiency levels was first evaluated by 10 native-speaking raters for comprehensibility and accentedness. Subsequently, the dataset was transcribed and analyzed for 12 lexical factors. Various lexical properties of L2 speech were found to be associated with L2 comprehensibility, and especially lexical accuracy (lemma appropriateness) and complexity (polysemy), indicating that these lexical variables are associated with successful L2 communication. In contrast, native speakers' accent judgements seemed to be linked to surface-level details of lexical content (abstractness) and form (variation, morphological accuracy) rather than to its conceptual and contextual details (e.g., lemma appropriateness, polysemy).

Keywords: Second language speech, comprehensibility, accentedness, vocabulary

Many second language (L2) researchers and practitioners agree that it is crucial to set realistic goals for L2 speakers (e.g., prioritizing understanding over nativelikeness), to enable them to communicate successfully with other native and non-native speakers (Derwing & Munro, 2009). Whereas recent research (e.g., Kang, Rubin & Pickering, 2010) has begun to examine how phonological aspects of language (e.g., segmentals, word stress, intonation) contribute to comprehensibility (listeners' perception of how easy it is for them to understand L2 speech) and accentedness (listeners' perception of the degree to which L2 speech is influenced by speakers' native language and/or is coloured by other non-native features), studies focusing on lexical correlates of L2 comprehensibility and accentedness are few in number. The goal of the current study was therefore to examine various lexical variables – appropriateness, fluency, variation, sophistication, abstractness, and sense relations – in the extemporaneous speech of 40 French speakers of English with a wide range of proficiency levels, and to identify potentially different contributions of these variables to listener judgements of comprehensibility and accentedness.

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

Comprehensibility versus accentedness

Improving L2 speaking proficiency has become an important challenge for L2 users, in terms of helping them not only succeed in L2 communication but also achieve their career-related goals. When it comes to linguistic skills, researchers have noted that L2 users tend to view native-speaker ability as an ideal teaching and learning goal (e.g., Derwing, 2003; Tokumoto & Shibata, 2011). However, previous L2 research has convincingly shown that few adult L2 learners attain nativelike linguistic abilities, even if they start L2 learning at an early age, and that a perceptible foreign accent is considered a largely unavoidable characteristic of L2 speech (e.g., Flege, Munro & MacKay, 1995). Thus, learning goals aimed towards successful communication in real-world settings need to be realistic. For example, improving comprehensibility may be an achievable L2 learning target while reducing or eliminating foreign accent may not (Derwing & Munro, 2009). This is based on the assumption that comprehensibility and accentedness are two interrelated yet independent constructs and that not all linguistic errors related to

Address for correspondence:
Kazuya Saito, Birkbeck, University of London, The Department of Applied Linguistics and Communication, 30 Russell Square, London, WC1B 5DT, UK
k.saito@bbk.ac.uk

accent may equally hinder comprehensibility (e.g., Hahn, 2004).

Distinguishing between these two global constructs of L2 speech – accentedness versus comprehensibility – is crucial in terms of both pedagogical relevance and theory building. From a practitioner's perspective, the importance of designing an optimal syllabus targeting L2 speaking skills has been strongly emphasized (Derwing, 2007), with a focus on those linguistic features that have the strongest potential to influence interlocutors' understanding. Some linguistic features that have been pinpointed as playing a role in transmitting the message include word stress (Levis, 2005), sentence stress (Hahn, 2004), speech and articulation rate (Kang et al., 2010), and some segmental contrasts (Jenkins, 2000; Munro & Derwing, 2006). Although L2 speakers' desire to sound nativelike cannot be fully ignored, they do not need to pursue accent-free mastery of the target language in order to succeed in L2 communication, perform well in most jobs, or integrate into society (Derwing & Munro, 2009; Jenkins, 2000; Levis, 2005). Foote, Holtby and Derwing's (2011) survey showed that teachers, particularly in the Canadian setting (the context of the current study), indeed increasingly acknowledge making students comfortably comprehensible to their listeners as the *rightful* goal of instruction.

From a theoretical perspective, comprehensibility may reflect L2 speakers' conversational experience better than linguistic nativelikeness. According to the interaction hypothesis (e.g., Long, 1996; Gass & Mackey, 2006; Pica, 1994), L2 speakers constantly make conscious or intuitive efforts to repair and improve their non-target production as they negotiate for meaning when faced with communication breakdowns with interlocutors in real-time aural-oral communication. If some linguistic features in L2 speech are more likely than others to cause communication breakdowns and thus trigger negotiation for meaning (Mackey, Gass & McDonough, 2000), then the learning value of L2 conversational interaction will be greatest for those linguistic features that are tied to comprehensibility rather than those that only contribute to the perception of accent (Saito, 2015). For example, in Derwing and Munro's (2013) longitudinal investigation, late-starting L2 speakers continued to improve over seven years of immersion when their oral proficiency was assessed through ratings of comprehensibility, but not accentedness.

Taken together, examining linguistic profiles of comprehensibility and accentedness is assumed to reveal a list of crucial linguistic features that L2 users need to prioritize in study, depending on whether they seek successful communication or perceived nativelikeness. This line of L2 research can thus clarify how L2 users improve their oral skills as their experience increases via selectively focusing on particular aspects of pronunciation, vocabulary, and grammar that are directly related to comprehensibility, but not necessarily to accentedness. Building on previous studies investigating phonological influences on comprehensibility and accentedness (e.g., Hahn, 2004), the current study is a first attempt to scrutinize the relationship between lexical characteristics of L2 speech and comprehensibility and accentedness.

Lexical Measures of L2 Speech Production

In the field of L2 learning, word knowledge is fundamental not only to theoretical views of speaking, such as psycholinguistic models of speech production (De Bot, 1996; Kormos, 2006), but also to practical aspects of teaching and assessment of L2 learners' speaking ability (Schmitt, 2008). However, "empirical studies on vocabulary and speaking proficiency are limited in scope" (Koizumi, 2012, p. 1), with research largely focusing on the percentage of words L2 speakers need to know to achieve various levels of comprehension of oral texts (e.g., van Zeeland & Schmitt, 2012) or the number of word families characterizing several genres of spoken discourse such as daily conversations (Adolphs & Schmitt, 2003), movies (Webb & Rodgers, 2009a), and TV programs (Webb & Rodgers, 2009b). This research has shown that knowledge of about 3,000–4,000 of the most frequent word families is sufficient for L2 speakers to reach a threshold of successful comprehension of spoken texts (Nation & Webb, 2011).

Though few in number, several studies have recently begun to analyze various lexical variables in L2 speech production, investigating how these variables interact to influence listeners' holistic judgements of L2 speaking (Crossley, Salsbury & McNamara, 2014; Crossley, Salsbury, McNamara & Jarvis, 2011; Iwashita, Brown, McNamara & O'Hagan, 2008; Lu, 2012). This research has chiefly focused on lexical profiles of L2 speech targeting several domains of word knowledge: (a) appropriateness (i.e., how appropriately words are chosen and used); (b) fluency (how many words are produced per minute or in total); (c) variation (how many different words are produced in total); (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).

Iwashita et al. (2008) examined lexical fluency and variation in L2 learners' TOEFL iBT speaking test performance, showing that both factors (i.e., token and type frequency) equally predicted five different levels of learners' overall speaking proficiency (beginner to advanced). These results contributed to the validation of the TOEFL iBT independent and integrated speaking scales (see also Brown, Iwashita & McNamara, 2005).

In Lu's (2012) study, large-scale speech data consisting of oral narratives from Chinese learners of English were transcribed and then computationally analyzed using 25 lexical fluency, variation, and sophistication measures. Learners' overall speaking proficiency (ranging from low to excellent) was mainly predicted by lexical variation (e.g., type-token ratio) and to a lesser degree by lexical fluency (e.g., text length, speech rate), but not by any lexical sophistication factors (e.g., ratio of infrequent words). Finally, Crossley et al. (2011, 2014) examined the relationship between computational analyses of lexical variables and L2 speaking proficiency judged by trained native speakers through the rating of written transcripts. Computational analyses involved the coding of transcripts for several lexical abstractness, sophistication, and sense relation factors using the Coh-Metrix software (McNamara, Graesser, McCarthy & Cai, 2014). Rater-based proficiency ratings were found to be related to several lexical variables, such as frequency, familiarity, imageability, concreteness, hypernymy (Crossley et al., 2011), and collocation appropriateness (Crossley et al., 2014).

To sum up, the findings of the aforementioned L2 vocabulary studies have shown that human ratings of L2 speech take into account various aspects of L2 lexical usage, ranging from lexical quality (appropriateness), quantity (fluency, variation), and type (sophistication, abstractness) to sense relations (polysemy). At the same time, native-speaking judges in these previous studies received specific rater training based on L2 proficiency scale descriptors adopted from existing tests such as TOEFL iBT (Iwashita et al., 2008), Test for English Majors (Lu, 2012), or ACTFL oral proficiency guidelines (Crossley et al., 2011, 2014). Because judges (who in some cases were trained assessors highly familiar with specific scales) were explicitly asked to apply existing test descriptors to make holistic judgements of L2 lexical proficiency, the results could have been influenced by the use of pre-existing definitions of L2 lexical proficiency contained in test descriptors (Koizumi, 2012). The current study extended this line of L2 vocabulary and speaking research by examining in depth the relationship between lexical properties of L2 speech and untrained native listeners' INTUITIVE judgements of L2 comprehensibility and accentedness. Our chief goal was to examine the extent to which previous findings based on trained raters (e.g., Crossley et al., 2011, 2014) can be generalized to untrained raters' intuitions and to determine the extent to which lexical properties of L2 speech relate to native-speaking raters' intuitive judgements of comprehensibility and accentedness.

Motivation for the Current Study

The two types of listener-based intuitive judgements of L2 speech – comprehensibility and accentedness – are partially overlapping but essentially different constructs (Derwing & Munro, 2009). While comprehensibility focuses on listeners' ease of understanding and describes a realistic goal of using the L2 for successful communication, accentedness characterizes the ideal goal of sounding like a native speaker. Trofimovich and Isaacs (2012) focused specifically on L2 comprehensibility and accentedness, investigating which linguistic aspects of L2 speech were associated with comprehensibility and which were uniquely linked to accentedness. Trofimovich and Isaacs analyzed picture narratives elicited from French speakers of L2 English for several linguistic measures spanning the domains of phonology (e.g., segmentals, suprasegmentals), lexis (e.g., appropriateness, fluency, variation), grammar (e.g., appropriateness, accuracy), and fluency (e.g., speech and articulation rate). While comprehensibility was associated with several linguistic categories (pronunciation, lexis, grammar, fluency), accentedness was mainly linked to pronunciation.

One limitation of this research was that it focused on relatively short speech samples (i.e., the first 30s of each recording). While these were adequate for thorough phonological coding and listener-based speech rating (Derwing & Munro, 1997; Hopp & Schmid, 2013; Munro & Mann, 2005), they were too short for robust and reliable lexical analyses, which require longer samples: for instance, those in excess of 100 words (Koizumi & In'nami, 2012). In fact, most previous research on the lexical content of L2 speech involves oral texts that are 3–5 min in length and consist of 100–200 words (e.g., Crossley et al., 2011, 2014; Lu, 2012). Thus, due to the limited length of oral texts, previous research by Trofimovich and Isaacs (2012) included only three lexical measures (lexical error rate, type and token frequency).

The primary aim of the present study was to investigate the effects of lexical variables on perceived comprehensibility and accentedness of L2 speech by targeting a more extensive set of lexical measures applied to oral texts longer than a few seconds in length and by isolating lexis-specific influences on L2 comprehensibility and accentedness through controlling effects of phonological variables. In this study, extemporaneous speech from L2 speakers of English with varied speaking proficiency levels was first rated globally for comprehensibility and accentedness, and then evaluated for several phonological variables (e.g., segmentals, syllable structure, word stress, intonation). L2 speakers' oral productions were subsequently transcribed and submitted to detailed lexical analyses, which involved 12 measures tapping into various domains of appropriateness (lemma, morphology), fluency (text length, speech rate), variation (breadth), sophistication (familiarity, frequency), abstractness (hypernymy, concreteness, imageability, meaningfulness), and sense relations (polysemy). The chief goal was to explore

the relationship between lexical variables and listeners' judgements of L2 comprehensibility and accentedness, while statistically controlling the effect of phonological factors.

Method

Speaking Task

Talkers

The L2 participants included the same 40 native speakers of French that participated in an earlier study (Trofimovich & Isaacs, 2012). The speakers (13 males, 27 females), who were on average 35.6 years old (28–61), were born and raised in monolingual French homes in Quebec, Canada. Apart from two early bilinguals, the speakers began L2 learning in elementary school at a mean age of 9.3 years, receiving up to three hours per week of subsequent L2 instruction. The speakers varied in their self-rated daily English use (range: 0–70%) and in their self-reported L2 proficiency in speaking, listening, reading, and writing (range: 1–9, where 1 = "extremely poor", 9 = "extremely proficient").

To ascertain that the speakers indeed represented variable levels of L2 speaking proficiency, several measures were derived from a 440-word read-aloud task recorded by all speakers. The recordings were subsequently evaluated by 10 native-speaking judges for accuracy of English /ð/ (as in "weather"), which represents a difficult segmental target for French speakers but that does not likely interfere with listener understanding relative to other phonemic substitution errors (Munro & Derwing, 2006; 0 = "does not sound like good English /ð/", 1 = "sounds like good English /ð/"). The recordings were also evaluated by five native-speaking listeners, who assessed speakers' nativelikeness (1 = "heavily accented," 9 = "not accented at all"). In addition, a measure of articulation rate (syllables per second) was computed, defined as the total number of syllables (including repetitions, hesitations) over the total duration of the sample. Individual speaker scores were wide-ranging for all three measures (7–99% correct for /ð/ accuracy, 1.8- 9.0 for accent, 0.4–3.4 for articulation rate), indicating that the speakers represented different ability levels, from beginner to advanced.

Material

Following previous L2 speech research (Derwing & Munro, 2009), extemporaneous L2 speech was elicited via a picture description task. Speakers described eight images depicting a story about two strangers bumping into each other on a busy street corner and inadvertently switching their suitcases, which were identical in appearance. The recorded speech samples were matched for peak amplitude, with initial dysfluencies (e.g., false starts, pausing) removed. Compared to the 30s speech samples used in an earlier study (Trofimovich & Isaacs, 2012), the speech samples used here included the entire picture descriptions so that lexical influences on L2 comprehensibility and accentedness could be investigated. In terms of text length, all samples exceeded the 100-word threshold set by Koizumi and In'nami (2012) for robust lexical diversity analysis, except for two samples (75 and 81 words). These samples came from low-level speakers who had difficulty producing a sustained narrative due to their limited linguistic ability. They were included in the final analysis because our aim was to carry out lexical profiling for speakers from a range of L2 speaking abilities, including beginner-level speakers. The two samples were of sufficient length (75s and 190s), compared to the remaining samples, but featured less lexical content, which is consistent with the idea that the two speakers spent a comparable amount of time on the task but produced less linguistic content overall. As a result, the target sample length varied between 55 and 351s, with a mean of 146s. The mean length of the final dataset was 209.2 words (range: 75–485 words).

Global Analysis

Raters

The L2 speakers' picture narratives were evaluated for two global dimensions of speech (comprehensibility, accentedness) by 10 listeners, native speakers of Canadian English from Montreal with a mean age of 23.4 years. The listeners were all born and raised in English-speaking homes in Canada, with at least one parent a native English speaker. All listeners estimated using English over 90% of the time in their daily lives but, as residents of Montreal (a bilingual French–English city), reported high familiarity with French-accented English. No listener reported any hearing problems. Based on previous research (e.g., Derwing & Munro, 2009), comprehensibility was defined as the perceived effort of how easy or difficult it is for listeners to understand an L2 speaker, while accentedness concerned listeners' perception of how different an L2 speaker's accent sounded from that of the native-speaker community.

Procedure

Speech rating was administered by a trained research assistant through the MATLAB software. The listeners used two free-moving 1000-point sliders on a computer screen to judge comprehensibility and accentedness in each speech sample. When the slider was placed at the leftmost (negative) end of the continuum, labeled with a frowning face ("hard to understand" or "heavily accented"), the rating was recorded as "0". If it was placed at the rightmost (positive) end of the continuum,

labeled with a smiley face ("easy to understand" or "not accented at all"), the rating was recorded as "1000". Apart from endpoint descriptors, the sliders included no marked intervals or labels (see Appendix for training scripts and onscreen labels). Each sample was played once, in line with previous research on L2 speech rating (e.g., Derwing & Munro, 2009), and listeners were allowed to make their judgements only after listening to each sample in its entirety. They were told that the speech samples represented a wide range of L2 speaking proficiency levels and were encouraged to use the entire scale. After judging three additional samples in a practice task and clarifying any remaining questions with the research assistant, listeners proceeded to evaluate the 40 speech samples in a unique randomized order. To minimize listener fatigue effects, the rating was divided into two one-hour sessions.

Phonological Analysis

Among the phonology categories analyzed in the earlier study targeting the same L2 speakers (see Trofimovich & Isaacs, 2012), four measures were directly related to knowledge of the spoken form of L2 words (Nation, 2001). These measures were re-used in this study to control for phonological influences on listener judgements of comprehensibility and accentedness.

(1) Segmental error ratio, defined as the total number of phonemic substitutions (e.g., "hit" spoken with /i/ in place of /ɪ/), 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., "WO-man" spoken as "wo-MAN"), over the total number of polysyllabic words produced.

(4) Intonation accuracy ratio, defined as the number of correct pitch patterns (rising, falling, or level tones) at the end of phrases (syntactic boundaries), over the total number of instances where pitch patterns were expected (e.g., "In a big city [level tone] Robert and Jane bumped into each other [falling tone]"). Typical intonation errors included the use of wrong intonation patterns (rising instead of falling contours and vice versa) and/or the lack of adequate and varied intonation (monotonous speech).

One trained coder with extensive linguistic training and experience received a detailed description for each measure, and then analyzed the entire dataset. Since all phonological measures (intonation, in particular) were subject to the influence of discourse context and individual speaker variability, the coder used not only speech software, i.e., *Praat* (Boersma & Weenink, 2012), but also relied on her intuition as a native speaker of English for the purpose of reliable error judgements. Subsequently, another trained coder checked 40% of the speech samples. Intraclass correlations, computed to determine coding agreement, revealed high consistency values exceeding .90 for all measures.

These four measures were statistically combined to derive a single pronunciation score for each L2 speaker using a Principal Component Analysis (PCA) with varimax rotation. The factorability of the entire dataset was examined and validated via the Bartlett's test of sphericity, $\chi^2 = 29.348$, $p < .001$, and the Kaiser-Meyer-Oklin measure of sampling adequacy (.665), which both exceeded required thresholds indicating excellent factorability of the correlation matrix (Hutcheson & Sofroniou, 1999). Creating a composite pronunciation factor, expressed as the phonology (pronunciation) PCA factor score combining the four initial phonology categories, thus allowed us to use it as a covariate in subsequent statistical analyses (partial correlations) in order to identify lexis-specific contributions to listener judgements of comprehensibility and accentedness.

Lexical Analysis

L2 speakers' picture narratives were first transcribed, and fillers (e.g., uh, ah, oh, umm) were eliminated from the original transcripts. The analyses of lexical appropriateness and speech rate (see below) were then carried out by trained coders. The remaining analyses of lexical fluency, variation, sophistication, abstractness, and sense relations were computed through the Coh-Metrix software, a computational tool which yields a total of 108 linguistic and discourse measures for a given text (McNamara et al., 2014). All of the first language (L1) intrusions specific to French (e.g., *malette* [for suitcase], *ah mon Dieu les temps en plus*) were counted as lemma errors in analyses of lexical appropriateness, but removed from texts for the remaining computational analyses.

Lexical appropriateness

Based on previous literature (e.g., Yuan & Ellis, 2003), two subcategories of lexical appropriateness were created.

(1) Lemma appropriateness, defined as the number of inaccurate and inappropriate words used (e.g.,

"high houses" instead of "high buildings"), which included L1 substitutions (e.g., "valise" [suitcase] for "suitcase"), over the total number of words. The majority of errors in this category involved inappropriate word choices, with an average of 8.8 errors per speaker ($SD = 6.6$). L1 substitutions were fewer in number, accounting for a mean of 2.5 errors per speaker ($SD = 5.7$).

(2) Morphological appropriateness, defined as the number of morphological errors related to verbs (i.e., tense, aspect, modality and subject-verb agreement), nouns (i.e., plural usage related to countable and uncountable nouns), derivations (i.e., wrong derivational forms, such as "confused" instead of "confuse"), articles (i.e., article usage in terms of definite, indefinite and null articles), and possessive determiners (i.e., use of "his" instead of "her") over the total number of words.

The 40 transcripts were first coded by a trained coder. Another trained coder then re-coded 25% of the transcripts (10/40) for each lexical appropriateness measure. Intraclass correlations were high for lemma ($r = .97$) and morphological ($r = .88$) appropriateness.

Fluency

Fluency was operationalized as the total number of words in a text, and was measured using two subcategories of raw and normalized fluency.

(3) Text length, defined as the total number of words in each transcript and computed through the Coh-Metrix software. Text length appears to be a good predictor of L2 speaking proficiency (Iwashita et al., 2008; Lu, 2012) and L2 writing proficiency (Engber, 1995; Johnson, Mercado & Acevedo, 2012).

(4) Speech rate, defined as the number of words produced per minute of speaking time and computed manually by dividing the total number of words in each sample by its duration in minutes (Lu, 2012). Speech rate has frequently been used as an index of fluency in various domains of SLA research, such as L2 pronunciation (e.g., Derwing et al., 2004) and vocabulary (e.g., Johnson et al., 2012; for a review, see Norris & Ortega, 2009). Speech rate has also been shown to strongly predict perceived fluency judgements (e.g., Cucchiarini, Strik & Boves, 2002).

Lexical variation

Lexical variation refers to "the range and variety of vocabulary deployed in a text by either a speaker or a writer" (McCarthy & Jarvis, 2007, p. 459).

(5) Lexical variation operationalized as a Measure of Textual Lexical Diversity (MTLD), following McCarthy and Jarvis (2010), and calculated through the Coh-Metrix software. Although lexical variation has traditionally been measured by computing the number of unique words in a sample (i.e., type-token ratio), such measures are highly dependent on text length, with longer texts associated with lower variation values. Lexical variation measures thus need to be mathematically transformed using such measures as MTLD, which, according to Koizumi and In'nami (2012), can be considered as an appropriate index of lexical variation, especially for oral texts consisting of 100–200 words.

Lexical sophistication

Lexical sophistication usually refers to "the proportion of relatively unusual or advanced words" in a text (Read, 2000, p. 203) and is typically measured based on the ratio of frequent to infrequent words (Laufer & Nation, 1995). This category was defined using subjective (familiarity) and objective (frequency) measures.

(6) Lexical familiarity, defined as the average familiarity rating for all content words in each transcript and computed through the Coh-Matrix software using 7-point rated word familiarity norms (1 = "word never seen", 7 = "word seen every day") from the MRC psycholinguistics database (Wilson, 1988). Subjective word familiarity, which captures how commonly a word is experienced, is believed to play a role in L2 vocabulary development because learners increase their vocabulary size through incidental and intentional encounters with words in real-life L2 experiences (Schmitt & Meara, 1997). For example, native speakers report more familiarity with words like "student", "city", and "book" than words like "figure", "fool", and "husband." Although learners are unlikely to show developmental change in this domain soon after L2 immersion, they use fewer familiar words as their proficiency improves (Salsbury, Crossley & McNamara, 2011).

(7) Lexical frequency, defined as the average frequency of all words in each transcript and calculated through the Coh-Matrix software using word frequency norms from the MRC psycholinguistic database (Wilson, 1988). More experienced learners, compared to less experienced ones, tend to use a greater proportion of words from lower-frequency bands, and objective word frequency helps differentiate written and oral texts produced by learners of varying ability levels (Crossley et al., 2011, 2014; Laufer & Nation, 1995).

Lexical abstractness

Following computational modeling of lexical development (Crossley, Salsbury, & McNamara, 2009,

2010; Crossley et al., 2011, 2014; Salsbury et al., 2011), four subcategories of lexical abstractness were targeted, namely, (a) hypernymy, (b) concreteness, (c) imageability, and (d) meaningfulness. All measures, which previously have been found to be related to interlanguage development, were designed to estimate the degree of abstractness of word meanings from various perspectives.

(8) Hypernymy, defined as the average number of subordinate and superordinate words for all nouns and verbs in each transcript and computed through the Coh-Matrix software using the WordNet lexical database of English (Fellbaum, 1998). This measure was used to capture hierarchical connections between superordinate (general) and subordinate (specific) lexical items which facilitate efficient processing and generalization of word knowledge. For example, words like "building" (superordinate terms) are less specific and more abstract than words like "library" and "hotel" (subordinate terms). Therefore, a lower hypernymy value suggests an overall use of more abstract words while a higher value characterizes an overall use of specific words. With an increasing amount of experience, learners typically produce words that are less specific and more abstract, which contributes to listeners' perceptions of their overall lexical proficiency (Crossley et al., 2009).

(9) Concreteness, defined as the average word concreteness value for all content words in a transcript and calculated through the Coh-Metrix software using concreteness ratings from the MRC psycholinguistics database (Wilson, 1988). This measure refers to the degree of concreteness of word meanings. Words referring to an object, material, or person that people can experience in the real world (e.g., "house", "car", "people") demonstrate relatively high concreteness values in relation to those words that are less concrete (e.g., "life", "problem"). L2 learners tend to learn concrete words at earlier stages of lexical development, and with greater ease, compared to abstract words (Crossley et al., 2009; Ellis & Beaton, 1993).

(10) Imageability, defined as the average word imageability value for all content words in a transcript and computed through the Coh-Metrix software using imageability ratings from the MRC psycholinguistics database (Wilson, 1988). This measure was used to capture the ease with which words elicit mental images of meanings. For example, some words (e.g., "sun", "mouth", "horse") denote more imageable meanings than others (e.g., "soul", "fault", "death"). Learners may have less difficulty learning more imageable words, compared to less imageable ones, because they can more easily experience and analyze these words (Ellis & Beaton, 1993). As their proficiency increases, learners also use less imageable words that do not evoke mental pictures, with utterances becoming more abstract and less context dependent (Salsbury et al., 2011).

(11) Meaningfulness, defined as the average number of word associations for all content words in a transcript and computed through the Coh-Metrix software using meaningfulness ratings from the MRC psycholinguistics database (Wilson, 1988). This measure estimated the degree of association of lexical items with other words. While more meaningful words (e.g., "food", "people") evoke many other related words, less meaningful words (e.g., "chance", "fault", "soul") involve limited links. As learners' proficiency improves, they increase the number of word associations (Zareva, 2007) and start using less meaningful words with fewer word associations (Salsbury et al., 2011).

Sense relations

The final lexical variable focused on semantic complexity and was used to estimate the number of individual senses of lexical items.

(12) Polysemy, defined as the average number of word senses (core meanings) for all content words in each transcript and computed through the Coh-Matrix software using the WordNet lexical database of English (Fellbaum, 1998). 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). In contrast, "wallet" has fewer senses that are limited to the meaning of a small, flat case used to hold things. Initially, learners likely focus on the core sense of a polysemous lexical item and then gradually shift their attention towards peripheral senses (Verspoor & Lowie, 2003).

Results

Global Analysis

In line with previous L2 speech research (e.g., Derwing & Munro, 2009), the 10 listeners demonstrated relatively high intraclass correlations for comprehensibility (r = .82) and accentedness (r = .86), suggesting that they were consistent in rating both dimensions of L2 speaking proficiency. Because these scores were sufficiently consistent, they were averaged across the listeners to derive a single mean comprehensibility and accentedness score for each L2 speaker. Table 1 summarizes descriptive

Table 1. *Descriptive Statistics for L2 Comprehensibility and Accentedness Ratings on a 1000-Point Scale*

Speaking dimension	*Mean*	*SD*	*Range*
Comprehensibility	690	210	240–1000
Accentedness	521	225	140–1000

Note. 1000 point scale (1 = heavily accented, difficult to understand, 1000 = not accented at all, easy to understand)

statistics for comprehensibility and accentedness ratings. According to a paired-samples *t*-test, the L2 speakers were rated lower in accentedness than in comprehensibility, $t(39) = -11.17$, $p < .001$, $d = .77$, with both sets of ratings being strongly associated, $r(38) = .91$, $p < .001$.

Lexical correlates of comprehensibility versus accentedness

Partial correlation analyses were first conducted to examine how the 12 target lexical variables (spanning the dimensions of lexical appropriateness, fluency, variation, sophistication, abstractness, and sense relations) related to L2 comprehensibility and accentedness ratings, with the pronunciation variable partialled out. These analyses showed that both comprehensibility and accentedness had strong positive associations with the fluency and variation measures (speech rate, MTLD) and with several abstractness measures (hypernymy, imageability, meaningfulness). In addition, whereas accentedness was moderately correlated with morphological appropriateness, comprehensibility was strongly associated with both lexical appropriateness measures (lemma and morphological) as well as with the sense relations (polysemy) measure (see Table 2).

To determine the unique contribution of the pronunciation variable to L2 comprehensibility and accentedness, partial correlation analyses were conducted between the raters' comprehensibility/accentedness scores and the pronunciation variable, with six relevant lexical variables (lemma errors for appropriateness, speech rate for fluency, MTLD for variation, frequency for sophistication, hypernymy for abstractness, polysemy for sense relation) partialled out. The results showed that, when controlling for lexical factors, the pronunciation variable was associated with both comprehensibility, $r(32) = -.41$, $p = .02$, and accentedness, $r(32) = -.40$, $p = .02$, to a similar extent.

To summarize, while the raters certainly relied on phonological information in their assessments (accounting for about 16% of shared variance), lexical factors nevertheless made an independent contribution to rater judgements of L2 speech, accounting for 14–50% of shared variance for comprehensibility (appropriateness, fluency, variation, abstractness, sense relations), and 10–22% for accentedness (morphological accuracy, fluency, variation, abstractness).

Table 2. *Partial Correlations between the Lexical Variables and Mean Comprehensibility and Accentedness Ratings*[a]

Lexical variable	Comprehensibility	Accentedness
Appropriateness[b]		
Lemma	−.68∗	−.28
Morphology	−.45∗	−.33∗
Fluency		
Text length	.12	−.08
Speech rate	.71∗	.43∗
Variation		
MTLD	.40∗	.41∗
Sophistication[c]		
Familiarity	−.27	−.28
Frequency	−.15	−.10
Abstractness[d]		
Hypernymy	−.37∗	−.33∗
Concreteness	−.30	−.23
Imageability	−.48∗	−.34∗
Meaningfulness	−.55∗	−.47∗
Sense relations		
Polysemy	.50∗	.21

Note. ∗$p < .05$; [a]A composite pronunciation score was partialled out from each correlation. [b]Since the appropriateness measures draw on the number of lemma and morphology errors, the correlations are negative. [c]Lower sophistication values indicate the use of more frequent and familiar words. [d]Lower abstractness values indicate the use of the use of more abstract and less specific words.

Discussion

Building on previous research investigating the relationship between pronunciation, fluency, vocabulary, and L2 oral ability, the current study examined to what extent lexical aspects of L2 speech relate to native speaker ratings of how easily L2 speech is understood (comprehensibility) versus how nativelike L2 speech sounds (accentedness). This study featured a comprehensive set of 12 lexical measures, including lexical appropriateness (Iwashita et al., 2008), fluency and variation (Lu, 2012), abstractness and sense relations (Crossley et al., 2010), while controlling for the effects of phonology (pronunciation). Unlike previous studies which targeted trained raters using exiting scale descriptors (e.g., Crossley et al., 2010; Lu, 2012), the current study examined L2 oral skills

through native-speaking listeners' intuitive judgements of comprehensibility (a realistic goal for using L2 as a successful non-native speaker) and accentedness (an ideal goal for speaking like a native speaker, at least for some L2 users).

First, we found that lexical properties of L2 speech explained considerable variance in native speakers' global judgements, even after the pronunciation factor (segmental, syllable structure, word stress, and intonation accuracy) was partialled out. These findings, which are based on untrained raters' scalar ratings, are in line with previous literature showing significant associations between L2 speech assessment by trained raters and the lexical domains of appropriateness (Iwashita et al., 2008), fluency (Lu, 2012), variation (Koizumi & In'nami, 2012), hypernymy (Crossley et al., 2009), concreteness, imageability, meaningfulness (Salsbury et al., 2011), and polysemy (Crossley et al., 2010). These findings imply that native speakers consider various aspects of L2 lexical information in their assessments of L2 speech, regardless of the presence of explicit rater training or the use of assessment rubrics associated with particular proficiency tests (e.g., TOEFL, ACTFL).

The native raters in this study differed in their weighting of lexical factors when evaluating comprehensibility versus accentedness. According to correlation analyses, comprehensibility was strongly associated with lexical appropriateness and fluency (r = .6-.7), and moderately associated with lexical variation, abstractness, and sense relations (r = .4-.5). When evaluating comprehensibility (with 14–50% of its variance related to vocabulary usage), therefore, listeners likely react not only to how often L2 users choose advanced vocabulary items (e.g., more abstract words with multiple senses) but also to how they use them in a contextually and conceptually appropriate manner. With respect to accentedness (with 10–22% of its variance linked to vocabulary), listeners' judgements were equally accounted for by aspects of variation (MTLD), fluency (speech rate), and abstractness (hypernymy, concreteness, imageability, meaningfulness) (r = .4–.5). Unlike comprehensibility, however, accent judgements were significantly correlated with morphological accuracy (r = .3), but not with lemma appropriateness or polysemy (p > .05).

Several explanations are possible to answer why the two lexical factors – lemma appropriateness and polysemy – distinguished between comprehensibility and accentedness. Our findings can be discussed in relation to previous studies on rater behaviour in assessment of comprehensibility and accentedness. To arrive at the overall MEANING of an utterance for comprehensibility judgement, native-speaking raters try to collect as much linguistic information as possible from L2 speech, which would include information about how accurately words are used (lemma appropriateness) and how semantically complex word meanings are (polysemy), with more comprehensible L2 speech linked to semantically accurate and rich content of utterances (see Munro & Derwing, 1995). In contrast, accent ratings can be invariably fast, effortless, and intuitive, arguably because raters likely focus on the FORM (i.e., morphological accuracy) rather than the meaning aspects of language (see Munro, Derwing & Burgess, 2010).

In the psycholinguistic literature, it has also been shown that native speakers generally recognize polysemous words with related multiple meanings faster and more easily than unambiguous words with few senses (Borowsky & Masson, 1996; Piercey & Joordens, 2000; Rodd, Gaskell & Marslen-Wilson, 2004). Listeners can rely on numerous semantic and syntactic entries stored in their mental lexicon to recognize a polysemous word and can also use these semantic and syntactic networks to inhibit any other competing lexical candidates (Klepousniotou & Baum, 2007). Based on psycholinguistic models of word recognition (e.g., Rodd et al., 2004), it is reasonable to assume that non-native speakers' frequent use of polysemous words in L2 speech likely helps native-speaking listeners activate rich conceptual networks, thereby making the task of understanding non-native speech less effortful (which should be reflected in comprehensibility ratings). However, polysemy may not be as relevant for accent judgements, arguably because listeners may not require rich, complex conceptual information in L2 speech to arrive at a nativelikeness judgement (Munro et al., 2010).

Overall, our findings are in line with prior speech research showing that L2 speakers who have reached a minimum pronunciation level required for successful L2 communication can be highly comprehensible, while still being fairly accented due to problems at the segmental and/or suprasegmental levels (Jenkins, 2000; Kang et al., 2010). Taken as a group, the L2 speakers in this study were rated more positively on the rubric of comprehensibility rather than accent. Our study extended this prior research by demonstrating how native listeners' perception of comprehensibility versus accentedness differs based on the lexical content of L2 speech. On the one hand, judgements of comprehensibility appear to guide listeners to process all available linguistic information in L2 speech, including lexical detail, from various perspectives (appropriateness, fluency, variation, abstractness, sense relations). In this regard, comprehensibility can serve as a good index for assessing the extent to which L2 speakers reach a threshold in terms of each lexical domain of L2 speech production with a view of successful communication, and therefore may reflect the development of L2 lexis (Crossley et al., 2009,

2010, 2011, 2014; Derwing & Munro, 2013; Saito, 2015).

On the other hand, accent judgement mostly requires listeners to attend to surface-level details of lexical content (abstractness) and form (variation, morphology), without processing much word meaning, particularly with respect to conceptual and contextual appropriateness or complexity of L2 word use (i.e., lemma appropriateness, polysemy). Since accentedness does not appear to tap into a wide range of lexical constructs in L2 production, this rubric may not be sensitive enough for capturing detailed lexical profiles of L2 oral ability. Recent speech research has shown that perceived accents are strongly tied to those linguistic features which are extremely difficult for even advanced L2 learners to master, such as segmentals (Saito, Trofimovich & Isaacs, 2015). To this end, more demanding lexical measures designed to assess linguistic nativelikeness, such as the use of proverbs and idiomatic expressions in L2 speech (Abrahamsson & Hyltenstam, 2009), may need to be included to understand the full scope of lexical influences on listener-based accent judgement in advanced users' speech.

Limitations

The present study took an exploratory approach towards defining two listeners-based L2 speaking constructs (comprehensibility vs. accentedness) and measuring various domains of L2 lexical production (appropriateness, fluency, variation, sophistication, abstractness, sense relations). Thus, several limitations need to be acknowledged for the purpose of providing a comprehensive picture of the role of vocabulary in L2 speaking. First, our findings were solely based on a single task (i.e., picture narratives). Although the sophistication variable (i.e., frequency, familiarity) was not a significant predictor for L2 comprehensibility or accent, it remains unclear whether and to what degree the nature of the task (describing a cartoon) succeeded in eliciting a sufficiently wide range of infrequent and unfamiliar lexical items. Thus, the generalizability of the study's results need to be tested in the context of various speaking tasks, such as more argumentative, formal, and complex tasks which induce speakers to demonstrate more varied and sophisticated L2 vocabulary use (see Hulstijn, Schoonen, de Jong, Steinel & Florijn, 2012), as well as other integrated task types which require speakers to synthesize various sources of information in their oral responses, such as integrated TOEFL iBT tasks (ETS, 2005).

Next, despite our efforts to recruit L1 French speakers with varying levels of L2 English proficiency (i.e., from total beginners to simultaneous bilinguals), the generalizability of the results may be limited, and should be tested with larger samples and wider proficiency ranges. To ensure a good balance of beginner, intermediate, and advanced L2 users, participants in future studies need to be carefully screened for or matched on several variables, such as age of acquisition (Flege et al., 1995), length of residence (Derwing & Munro, 2013), and aptitude (Granena, 2014), or recruited in reference to an L2 proficiency benchmark (e.g., the Common European Framework of Reference Levels; Council of Europe, 2001) to ensure greater variability in proficiency levels. Furthermore, these findings also need to be replicated with speakers from various L1 backgrounds acquiring different L2s (Crowther, Trofimovich, Saito & Isaacs, 2014). This point is particularly important due to a significant overlap between English and French in vocabulary (e.g., over 30% of English words are Latinate in origin), especially for less frequent, more abstract, and more academic words. Thus, future research needs to examine how the results specific to native French speakers of English compare to findings for speakers of English from non-Romance L1s, that is, language users who may not readily benefit from L1 transfer (Granger, 1993).

Another limitation of the current study relates to the listener factor. The present findings are specific to 10 native-speaking listeners who had experience with French-accented English.[1] Thus, to ensure the generalizability of findings to more varied listener populations, future research should target larger cohorts of listeners with specific backgrounds. For instance, it would be interesting to examine whether and to what degree lexical correlates of L2 comprehensibility and accentedness differ for native versus non-native raters (i.e., Munro, Derwing & Morton, 2006), with and without familiarity with the targeted L2 speech (i.e., Winke, Gass & Myford, 2013), and with varying degrees of linguistic and pedagogical experience (i.e., Isaacs & Thomson, 2013).

Finally, it is important to point out that although we used the Coh-Metrix software to compute measures of lexical abstractness and sophistication, with the view of ensuring comparability of our dataset with previous findings (e.g., Crossley et al., 2009, 2010, 2011, 2014), the validity of Coh-Metrix as a research tool must be evaluated in future research (for further discussion of the applicability of Coh-Metrix to L2 vocabulary research, see Salsbury et al., 2011). For instance, Coh-Metrix uses the MRC psycholinguistic database, which includes ratings for only a subset of words. It is thus possible that the familiarity, hypernymy, concreteness, imageability, and meaningfulness variables used in this study were

[1] Notably, the lemma appropriateness measure, which featured the strongest association with L2 comprehensibility, included both the ratio of incorrect word choices as well as L1 substitutions. This suggests that even if some francophone speakers used French words, most listeners would have been familiar with them. Future studies need to clarify and, if possible, control for the role of listener familiarity with speakers' languages in listener-based assessments of comprehensibility and accentedness (Winke et al., 2013).

calculated for only a portion of words in the texts we analyzed. Therefore, future studies should develop or refine methodological tools that target specific sets of lexical variables, consistent with the relevant theoretical principles and instructional contexts.

Conclusion

The current study demonstrated differential effects of L2 lexical usage on native speakers' judgements of L2 speech from the perspective of comprehensibility and accentedness. Three broad conclusions can be drawn from the findings of this study. First, a substantial proportion of variance in native speakers' judgements of L2 speech was explained by the lexical usage of tokens, even after the pronunciation factor was statistically controlled. Second, variables tied to appropriate and fluent use of words, as well as those linked to more abstract and less specific words with multiple senses were predictive of comprehensibility. Third, variables related to fluency, variation, and abstractness were associated with accentedness. The results suggest that various lexical aspects of L2 speech are associated with L2 comprehensibility, implying that using not only abstract, but also polysemous words in an appropriate and fluent manner is fundamental to successful L2 communication. In contrast, native speakers' accent judgements may largely concern only surface-level details of lexical content in L2 speech, rather than conceptual and contextual appropriateness or complexity of L2 word use.

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of acquisition and nativelikeness in a second language – listener perception vs. linguistic scrutiny. *Language Learning, 59,* 249–306.

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

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer*.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 63–85.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks. Monograph Series MS-29. Princeton, NJ: Educational Testing Service.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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*.

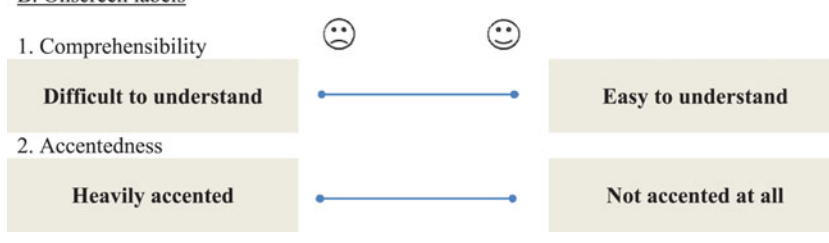
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45,* 182–193.

Appendix

A. Training script

Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.
Accentedness	This refers to how much a speaker's speech is influenced by his/her native language and/or is coloured by other non-native features.

B. Onscreen labels



Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2014). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*. Published online 28 October 2014. doi:10.1002/tesq.203

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111,* 2862–2873.

De Bot, K. (1996). The psycholinguistics of the Output Hypothesis. *Language learning, 46,* 529–555.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42,* 476–490.

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review, 59,* 545–564.

Derwing, T. M. (2007). Curriculum issues in teaching pronunciation to second language learners. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 347–369). Amsterdam: John Benjamins.

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

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.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4,* 139–155.

ETS. (2005). *TOEFL iBT tips: How to prepare for the next generation TOEFL test and communicate with confidence*. Princeton, NJ: Author.

Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

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

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

Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA review, 19,* 3–17.

Granena, G. (2014). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63,* 665–703.

Granger, S. (1993). Cognates: an aid or a barrier to successful L2 vocabulary development?. *ITL. Institut voor Togepaste Linguistik*, (99–100), 43–56.

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

Hopp, H., & Schmid, M. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics, 34,* 361–394.

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.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage.

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

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: Oxford University Press.

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing, 21,* 264–282.

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

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics, 20,* 1–24.

Koizumi, R. (2012). Vocabulary and speaking. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Oxford: 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.

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: Second language acquisition* (pp. 413–468). New York: Academic Press.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal, 96,* 190–208.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition, 22,* 471–497.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

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.

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

Munro, M. J., Derwing, T. M., & Burgess, C. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication, 52,* 626–637.

Munro, M., & Derwing, T. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38,* 289–306.

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

Munro, M., & Mann, V. (2005). Age of Immersion as a predictor of foreign accent. *Applied Psycholinguistics, 26,* 311–341.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30,* 555–578.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning, 44,* 493–527.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition, 28,* 657–666.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28,* 89–104.

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*. Advance online publication. doi: 10.1111/lang.12120

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Developing a process-oriented model for linguistic influences on comprehensibility and accentedness in second language speech production. *Applied Psycholinguistics*. Advance online publication. doi: 10.1017/S0142716414000502

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.

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

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

van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34,* 457–479.

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language Learning, 53,* 547–586.

Webb, S., & Rodgers, M. P. H. (2009a). The vocabulary demands of television programs. *Language Learning, 59,* 335–366.

Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics, 30,* 407–427.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary. *Behavioural 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.

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.