Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters

Randy Appel, Pavel Trofimovich, Kazuya Saito, Talia Isaacs, and Stuart Webb

Waseda University | Concordia University | Birkbeck, University of London | University College London | Western University

This study analyzed the contribution of lexical factors to native-speaking raters' assessments of comprehensibility and nativeness in second language (L2) speech. Using transcribed samples to reduce non-lexical sources of bias, 10 naïve L1 English raters evaluated speech samples from 97 L2 English learners across two tasks (picture description and TOEFL integrated). Subsequently, the 194 transcripts were analyzed through statistical software (e.g., Coh-metrix, VocabProfile) for 29 variables spanning various lexical dimensions. For the picture description task, separation in lexical correlates of the two constructs was found, with distinct lexical measures tied to comprehensibility and nativeness. In the TOEFL integrated task, comprehensibility and nativeness were largely indistinguishable, with identical sets of lexical variables, covering dimensions of diversity and range. Findings are discussed in relation to the acquisition, assessment, and teaching of lexical properties in L2 speech.

Keywords: second language speech, vocabulary, comprehensibility, nativeness

Although diverse in their specific interests, studies of second language (L2) oral (speaking) performance have largely concentrated on global constructs of L2 speaking ability as evaluated by trained raters using prescribed grading rubrics from high-stakes assessments, such as the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), and the Chinese Ministry of Education's Test for English Majors (Crossley & McNamara, 2013; Crossley, Salsbury, & McNamara, 2015; Crossley, Salsbury, McNamara, & Jarvis, 2011; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lu, 2012). However,

https://doi.org/10.1075/itl.17026.app

ITL - International Journal of Applied Linguistics 170:1 (2019), pp. 24–52. issn 0019-0829 | e-issn 1783-1490

the goal of many L2 speakers, including instructed learners, is to prepare for life outside of the language classroom (Derwing & Munro, 2015). Therefore, it is important to target additional rater populations that may be more representative of the type of interlocutors L2 speakers are likely to encounter during real world interactions. For L2 speakers studying at English medium universities, one rater population that can provide a more accurate indication of the level of communicative success L2 speakers are likely to achieve while living and studying in target language communities is native-speaking (L1) English post-secondary students. Because these students frequently interact with L2 speakers in academic and social settings, they represent a key stakeholder group in understanding L2 speech. As a result, it is important to examine how L1 English post-secondary students perceive L2 speech and the specific factors that contribute most to their perceptions of successful communication.

Comprehensibility, defined as perceived ease of understanding, has emerged as a practical and reliable means of capturing raters' impressionistic judgements of L2 speech (e.g., Derwing & Munro, 2015). A focus on comprehensibility is consistent with the idea that what is relevant to communication is understandable speech, not necessarily nativelike or accent-free production (Derwing & Munro, 2015; Levis, 2005). However, the role of lexis in raters' assessments of comprehensibility, particularly as compared to nativeness[1] (i.e., L2 speech that adheres to native speaker norms), has remained largely unexplored.

The goal of the current study was to examine lexical correlates of comprehensibility and nativeness in L1 English raters' evaluations of L2 speech across two tasks (picture description, academic summary). This was achieved by targeting raters' judgements of these constructs using Likert-type scales in relation to a comprehensive set of 29 fine-grained lexical measures of speech content. To control for the influence of pronunciation, fluency, and prosody, and better understand the specific role lexis plays in judgements of spoken ability, raters in this study evaluated transcribed L2 English speech samples.

Research on L2 speaking

Early research on dimensions of L2 speech focused on the identification of specific subconstructs that contribute most to global evaluations of L2 speaking ability

[1] While previous research using audio recordings of L2 speech often refers to *accentedness* as a label for nativelike production, we adopt the term *nativeness* in the present study to reflect our focus on transcribed speech samples which are devoid of prosody, pronunciation, and fluency related factors.

(e.g., global proficiency), often using raters' holistic assessments of grammatical accuracy, lexical variation, and pronunciation, among others (e.g., Adams, 1980; McNamara, 1990). For example, Adams analyzed trained raters' global evaluations of L2 speaking ability in relation to holistic judgements of accent, comprehension, fluency, grammar, and vocabulary from oral interviews. For these raters, factors related to grammar and vocabulary were most closely associated with overall proficiency scores. Similarly, McNamara found that trained rater evaluations of grammar and expression were the strongest predictors of L2 speakers' overall communicative effectiveness. While these early studies revealed various subcomponents contributing to evaluations of L2 speaking ability by trained raters, the use of holistic judgements of grammar or lexis prevented researchers from making more nuanced conclusions about the specific lexical or morphosyntactic factors most relevant to these assessments.

More recently, computer-aided forms of analysis have allowed researchers to analyze L2 discourse in more fine-grained ways that are comparable across research studies. For example, Lu (2012) used computer-aided extraction of 26 lexical measures along three categories (lexical density, sophistication, and variation) to quantify trained English teachers' evaluations of 408 audio recordings from the Spoken English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005), showing that automated measures of lexical variation were moderately correlated with raters' assessments of L2 oral ability. Similarly, Crossley and McNamara (2013) used computer-aided extraction of measures targeting lexis, topic development, and delivery to model expert rater evaluations of 244 audio recordings of L2 English speech, showing that these measures (especially those related to word type counts and word frequency) accounted for 61% of variance in human judges' overall speaking proficiency scores (see also Crossley et al., 2015; Crossley, Salsbury, & McNamara, 2010; Ginther, Dimova, & Yang, 2010; Iwashita et al., 2008).

While this research has resulted in greater reliability and objectivity in the analysis of L2 speech, the emphasis on evaluations by expert/trained raters or experienced L2 teachers has largely remained (e.g., Crossley & McNamara, 2013; Crossley et al., 2011, 2015; Ginther et al., 2010; Iwashita et al., 2008; Lu, 2012). This focus on trained raters or experienced language teachers has left open the possibility that features being attended to simply result from the training process used to prepare raters, exposure to language teaching pedagogy and theory, or adherence to specific features listed in grading rubrics. In light of these limitations, it is necessary to look at additional rater populations – including naïve raters (i.e., raters with no specialized training in linguistics or language teaching) – to more fully understand how L2 speech is perceived by potential interlocutors in target language communities (Koizumi & In'nami, 2012). Particularly in the context of L2 speakers studying at English medium universities, it is important to target L1

English students' perceptions of L2 speech since it is these individuals who will frequently interact with L2 speakers in academic and social settings.

L2 comprehensibility versus nativeness

Comprehensibility, based on intuitive evaluations characteristic of the kinds of impressionistic judgements language users make about their daily experiences with language (Oppenheimer, 2008), has emerged as a useful construct in assessments of L2 speech. Highlighted as an important aspect of L2 ability and a central concern for L2 learners (Abercrombie, 1949; Derwing & Munro, 2015), comprehensibility is often discussed in contrast to the competing ideology of accent-free, nativelike L2 speech. Arguments in favor of a focus on comprehensibility over nativeness stem from a belief that even heavily accented speech can be considered highly comprehensible (Derwing & Munro, 2015), implying that understandable L2 output is ultimately more important for successful communication.

To date, scholars have primarily examined pronunciation, prosody, and fluency related dimensions of L2 speech that are associated with raters' assessments of comprehensibility and nativeness. With respect to comprehensibility, raters appear to attend to various aspects of L2 speech, including segmentals (Munro & Derwing, 2006), prosody (Kang et al., 2010), fluency (Derwing, Rossiter, Munro, & Thompson, 2004), and grammatical accuracy (Derwing, Rossiter, & Ehrensberger-Dow, 2002). In contrast, judgements of nativeness appear to be tied exclusively to segmental and prosodic accuracy (e.g., Saito, Trofimovich, & Isaacs, 2016; Munro, Derwing, & Burgess, 2010). However, the prevailing focus on pronunciation, fluency, and prosody aspects of L2 speech, in relation to comprehensibility and nativeness, has limited our understanding of the specific role lexis plays in rater evaluations of these constructs.

To control for the influence of nonlexical sources of bias in assessments of L2 speech, and better evaluate the specific role lexis plays in raters' judgements of spoken ability, researchers have recently begun targeting ratings of transcribed samples, as opposed to audio recordings (e.g., Saito, Webb, Trofimovich, & Isaacs, 2015; Crossley et al., 2014; Kyle & Crossley, 2015). For example, Saito et al. (2015) provided initial evidence for possible associations between lexical content of L2 speech and raters' comprehensibility judgements, using transcribed samples from 40 L1 French speakers of L2 English who completed a picture description task. Among 12 lexical measures tapping into various dimensions of L2 speech, four specific dimensions – lexical appropriateness, lexical fluency, lexical variation, and sense relations – were identified as relevant to raters' comprehensibility judgements. Conceptualized as a follow-up to this initial investigation, the current study extended these preliminary findings to a larger sample of L2 speakers (97 L2 speakers from varied language backgrounds and L2 proficiency levels) using two different tasks varying in complexity (picture description task and academic summary). More importantly, to the best of our knowledge, the current investigation was the first attempt not only to identify specific lexical contributions to raters' assessments of comprehensibility, but also to determine if comprehensibility can be distinguished from nativeness in terms of various measures of L2 lexis.

The current study

Given the need to better understand how raters perceive L2 speech and identify lexical characteristics of L2 comprehensibility (as distinct from nativeness), this study targeted naïve L1 English raters' impressionistic judgements of comprehensibility and nativeness in L2 English speech samples produced by 97 speakers from multiple language backgrounds. Following previous research (e.g., Saito et al., 2015; Crossley et al., 2014), speech samples were transcribed to remove all nonlexical sources of bias before being assessed. As a result, the current study was able to offer a lexis-based interpretation of the factors influencing rater perceptions of L2 English comprehensibility and nativenes. Because task type may influence ratings (Crowther, Trofimovich, Isaacs, & Saito, 2015; Kuiken & Vedder, 2014), this study focused on L2 speech produced across two different tasks to further explore the role of task type in assessments of linguistic ability. The study was guided by three questions:

1. Which lexical factors underlie holistic judgements of L2 English comprehensibility and nativeness as judged by naïve L1 English raters using transcribed speech samples?
2. Are the lexical correlates of comprehensibility and nativeness, as evaluated by naïve L1 English raters using transcribed samples, distinct? Put differently, can comprehensibility and nativelikeness be distinguished in terms of their lexical correlates?
3. Do the lexical correlates of L2 English comprehensibility and nativeness in transcribed speech samples vary according to task type?

Method

Speakers

The L2 participants were 97 speakers (20 female, 77 male) with a mean age of 24.2 years (*SD* = 3.14) from an unpublished corpus of L2 English speech (Trofimovich & Isaacs, 2012). The speakers were all international students in undergraduate (19) and graduate (78) programs at a large English-medium university in Canada. The language backgrounds represented in the corpus included Farsi (20), Hindi (11), Telugu (9), Chinese (10), Bengali (9), Punjabi (6), French (6), Spanish (5), Tamil (5), Arabic (4), Gujarati (3), Marathi (2), Urdu (2), Akan, Kannada, Russian, Malayalam, and Portuguese (1 each). The L2 speakers arrived in Canada at a mean age of 23.5 years (*SD* = 3.90) and participated in the study during their first term of university studies. Speakers reported having learned English for an average of 13.5 years (*SD* = 5.13) and estimated using English 10–100% of the time daily (*M* = 63%). All speakers had recently taken either the TOEFL iBT or IELTS. For the speaking component of each test, mean scores were 21.84 (*SD* = 3.19) for the TOEFL iBT and 6.63 (*SD* = 0.88) for the IELTS. The overall test scores were 90.58 (*SD* = 8.52) for the TOEFL iBT and 6.86 (*SD* = 0.68) for the IELTS. Self-reports indicated that speakers represented a range (3–9) of L2 speaking ability (*M* = 6.3), based on a 9-point scale (1 = *extremely poor*, 9 = *extremely proficient*).

Speaking tasks

All speakers performed two speaking tasks varying in cognitive demand to evaluate the potential impact this factor may have on perceptions of L2 speaking ability. The first task (picture description) involved a series of eight images depicting an encounter between a male and female traveler who realize they have accidentally exchanged suitcases after bumping into each other on a street corner (Derwing et al., 2004). After reviewing the images, speakers were given 30 seconds of planning time and 60 seconds to describe the series of events. The use of this task allowed for comparability of findings with previous studies that used this same task (e.g., Saito et al., 2015).

The second task (TOEFL integrated listening/speaking task), adapted from TOEFL preparation materials (Educational Testing Service, 2004), required speakers to listen to a brief academic lecture and read a short paragraph on the same topic before integrating this information into a response that demonstrated understanding of the subject. After listening to the lecture and reading the paragraph, speakers were allotted 30 seconds of preparation time, followed by 60 seconds to speak. As this study used an existing unpublished corpus of L2 speech samples, the data contained two task versions, featuring two topics (actor/observer effects, social influences on perception), with approximately half of the speakers assigned to each. However, since analyses showed no consistent differences in sample length or reported difficulty, all data across TOEFL integrated task versions were pooled together.

Using Robinson's (2005) task complexity framework, the picture description and TOEFL integrated tasks were evaluated for differences in complexity. As the picture description task contained less input (eliciting language constrained by the depicted objects, actions, and relationships), contained fewer elements, and did not call for any receptive skills (listening and reading) or reasoning, compared to the TOEFL integrated task, this task was considered the less cognitively demanding of the two tasks. In contrast, the TOEFL integrated task contained multiple elements (listening and reading components) and required reasoning to integrate these sources into a coherent response. Following Révész, Michel and Gilabert (2015), speakers' self-ratings were used to further substantiate claims of task differences in cognitive demand. Upon completion of audio recordings during initial data collection, all speakers were asked to estimate task difficulty using a 9-point scale (1 = *very easy*, 9 = *very difficult*), with the TOEFL integrated task ($M = 4.39$, $SD = 2.02$) rated as significantly more difficult than the picture description task ($M = 3.28$, $SD = 1.87$), $t\ (92) = 5.45$, $p < .001$, $r = .49$.

Materials

Following our previous work (e.g., Saito et al., 2015), speech samples were orthographically transcribed by a trained research assistant and verified by another trained coder. All nonlexical features that could potentially bias rater evaluations were excluded from the transcriptions, such as word stress, prosody, and pronunciation (i.e., *when* pronounced as *ven, that* pronounced as *zat*). In addition, filled pauses (e.g., ummm, ahhh) were also removed, as these were not believed represent lexical data that would aid the analysis.

To preserve as much of the original linguistic information as possible, and to keep in line with previous research targeting analyses of transcribed speech using similar automated measures (e.g., Crossley & McNamara, 2013; Kyle & Crossley, 2015), other dysfluency markers (e.g., false starts, word repetitions, repairs) were retained. Words that could not be transcribed due to audio quality issues or lack of understanding were indicated by /–/ (less than 1% of total words). Since punctuation is a feature of written English, and would have been based on the transcriber's subjective judgement, punctuation (including capitalization) was not included.

All transcripts were checked to ensure a minimum of 95 words per speaker. Although a cutoff of 100 words for certain lexical analyses (e.g., variation) is preferred (e.g., Koizumi & In'nami, 2012), a slightly lower minimum word count was implemented in hopes of capturing a wider range of linguistic abilities represented in the corpus, as lower level speakers often produce shorter samples than their higher level counterparts (e.g., Kormos & Dénes, 2004). In total, 7 samples in the picture description task and 6 in the TOEFL integrated task contained fewer than 100 words. The resulting corpora were comparable in total words (15,768 vs. 14,201) and mean sample length (in words) across the two tasks ($M=162$, *range* = 95–321 vs. $M=146$, *range* = 95–215). However, a paired samples t-test revealed that there was a statistically significant difference in length between the two corpora. Thus, raters evaluating the TOEFL integrated task had more lexical information to draw on when assessing these transcribed samples.

Raters

Raters included 10 untrained, L1 English speakers (6 female, 4 male), all enrolled in non-linguistics and non-education undergraduate programs at the same English-medium university. All raters ($M_{age}=21.3$ years, *range* = 19–24) learned English from childhood, with at least one native English-speaking parent, and reported no previous language teaching experience and no prior courses in applied linguistics or related fields. As students at a large university located in a multicultural urban setting with 16% of the student body comprising international students, the raters were likely familiar with L2 English speech by speakers from various language backgrounds.

Rating procedure

The 194 transcribed L2 English speech samples (97 from each task) were evaluated for comprehensibility and nativeness by L1 the English raters during two individual rating sessions of approximately 1.5 hours each. During each session, raters provided evaluations for one task type (picture description or TOEFL integrated), with half assessing the picture description task first and the remaining half rating the TOEFL integrated task first. Upon completing a short language background questionnaire, raters were given a brief explanation of the two rated constructs. Comprehensibility was defined as ease of understanding, with nativeness defined as how closely the language resembled that of a native speaker (training materials and sample on-screen interface are provided in Appendix A). After explaining the target constructs, raters were trained on the interface used to administer the task and record evaluations. Before beginning each session, raters were given three practice transcripts to test their understanding of the rating procedure and provide an opportunity to ask questions. These practice samples were taken from existing L2 speech samples that did not meet the required minimum word counts for inclusion but were modified to ensure a minimum of 95 words per sample. To eliminate the effect of topic/task familiarity on rater judgements, all raters were familiarized with the materials used to elicit responses in each task before beginning the corresponding rating session.

Transcribed speech samples were presented individually on screen without time limits in a unique random order for each rater. To allow raters the freedom to provide as accurate as possible ratings, below each transcript, there were two free-moving 1,000-point scales for assessing comprehensibility and nativeness, with the negative endpoint (corresponding to the rating of 0) labeled by a frowning face and the positive endpoint (corresponding to the rating of 1,000) labeled by a smiling face. Raters were informed that samples were taken from L2 English speakers with a variety of language backgrounds and linguistic abilities, and therefore encouraged to use the full range of each scale. To promote careful reading of each transcript before assigning a score, raters were only able to record their ratings after the text had remained on screen for at least 5 seconds.

Lexical analysis

To derive specific measures that could be explored in relation to rater evaluations of comprehensibility and nativeness, the 194 transcribed L2 English speech samples were analyzed for 16 lexical variables using Coh-metrix 3.0 (Grasesser, McNamara, Louwerse & Cai, 2004), a computational analysis tool that provides lexical indices spanning 11 main categories. Metrics from 7 of these categories were selected for the present study: descriptive statistics, lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, and word information. An additional 13 lexical measures came from VocabProfile (Cobb, 2016), which allows for the analysis of texts based on word frequencies drawn from several large-scale corpora. In total, 29 lexical measures were computed for each task (see Table 1 for the full list of measures, described in greater detail in the next two sections).

As this study was largely exploratory in nature and targeted ratings of transcribed speech, which contained dysfluency markers, a wide range of lexical measures were included to better understand which specific indices would be best suited to this type of data. Although automated lexical measures have been successfully applied in previous studies targeting similarly transcribed speech samples (e.g., Crossley & McNamara, 2013; Kyle & Crossley, 2015), the inclusion of false starts, word repetitions, and repairs may negatively impact the accuracy of several indices. For example, automatically extracted indices related to lexical sophistication (see Table 1) are likely to be affected by the inclusion of dysfluency markers, as these elements would be counted as separate words, thereby altering frequency counts and ultimate scores awarded for each of these measures. In addition, since coh-metrix, the main tool used for automatic extraction of relevant indices in this study, was trained on written data, precision of the text parser is likely to suffer when analyzing transcribed speech. As a result, several metrics, such as modifiers per noun phrase, may also be negatively impacted by the analysis of samples that retain the previously mentioned dysfluency markers.

By using several potentially overlapping measures, we aimed to mitigate these limitations by taking a multi-angled approach to the analysis of our data and test the feasibility of each freely available metric as a predictor of naïve L1 English raters' assessment of L2 English spoken ability using transcribed samples. As all measures were automatically extracted using publicly available resources, these metrics also enabled the removal of rater bias in value assignments for each lexical category and allow for easy replication using alternative corpora.

Coh-metrix measures

Coh-metrix, which adheres to theoretical frameworks that view understanding as functioning on various levels (Graesser, McNamara, & Kulikowich, 2011), uses characteristics of individual words, sentences, and discourse level connections to evaluate text at multiple levels of analysis (McNamara, Graesser, McCarthy, & Cai, 2014). The 16 lexical measures calculated through Coh-metrix spanned seven categories.

- *Mean number of syllables* (descriptive statistics category) was used as a measure of the average word length within each transcribed sample. As word length is one potential indicator of readability, this measure likely reflected some aspects of understanding and processing ease. Because the data involved transcribed speech, mean number of syllables, rather than character length, was considered a more appropriate estimate of word length.
- *Measure of Textual Lexical Diversity* (MTLD, lexical diversity category) was used as an adjusted measure of lexical diversity. In general, higher lexical variation is indicative of less lexical overlap and more unique words (Jarvis & Daller, 2013). Because greater lexical variation is interpreted by raters as a sign of increased linguistic ability (e.g., Lu, 2012), MTLD was considered to contribute in similar ways to raters' L2 speech judgements.

- *Causal connectives* (e.g., because, therefore), *logical connectives* (e.g., and, or), and *additive connectives* (e.g., furthermore, moreover) were derived for each transcribed speech sample (connectives category). Connectives are an important aspect of cohesion that help bind discourse together, making it easier to process and understand (e.g., Crossley, Yang, & McNamara, 2014). All connectives were recorded as incidence scores (averaged to occurrences per 1,000 words).
- *Causal verb frequency* (e.g., hit, move), combined incidence score for *causal verbs and causal particles* (e.g., hit, move, because, in order to), and *verb overlap* (situation model category) were computed as measures of causality in each speech sample. Based on research from cognitive science and discourse processing, situation model refers to mental representations present within a text that go beyond the surface level of word-by-word comprehension (McNamara et al., 2014). In this sense, situation model refers to the rater's mental representations of the meaning conveyed by each sample. As measures from this category are seen as indicators of level of understanding, they were computed for their potential associations with comprehensibility and nativeness. As with the connectives category, incidence scores for causal verbs and combined incidence scores for causal verbs and particles were based on averaged occurrences per 1,000 words. For the verb overlap measure, Coh-metrix uses WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) to classify verb categories, calculating it on a point-based system where, if two verbs are found to be synonyms, they are awarded a score of 1; otherwise, a score of 0 is given.
- *Average number of modifiers per noun phrase* (syntactic complexity category) and two measures of verb form incidence, relative *frequency of -ing verb forms*[2] (syntactic pattern density category) and relative frequency of *infinitives* (verbs in unmarked form, such as be, have), served as measures of lexical concentration for various parts of speech. Because increased levels of syntactic complexity are associated with greater processing difficulty (Perfetti, Landi, & Oakhill, 2005), these measures were used as potential indicators of comprehensibility and nativeness.
- *Word frequency, word age of acquisition, word familiarity, word concreteness*, and *word meaningfulness* (word information category) were calculated for each sample. Word frequency counts indicate the frequency with which each word appearing in a sample occurs within the English language in general,

[2] Although labeled as "gerunds" in Coh-metrix, this measure may be more accurately referred to as an index of *-ing* forms since the Coh-metrix text parser is unable to accurately distinguish between gerunds and participles sharing the same form (McNamara et al., 2014)

using the 17.9 million word CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995) and may help differentiate discourse produced by users of varying linguistic abilities (Crossley et al., 2014). To arrive at a score for each sample, Coh-metrix uses the logarithm of word frequency for all identified words. Any words that are absent from the CELEX corpus receive a score of 0 and are therefore not included in the resulting score. The remaining measures in this category were computed based on information from the Medical Research Council Psycholinguistics Database (Wilson, 1988), which contains human ratings for more than 150,000 words on 26 psychological properties. Word age of acquisition is an averaged indication of the age at which words are acquired by native speakers. With more complex words being acquired later than simple words, raters may interpret this measure as a sign of linguistic maturity. Word familiarity is an estimate of the average level of familiarity for all words present within each sample to adult English users and may be effective in discriminating between different L2 English proficiency levels (Salsbury, Crossley, & McNamara, 2011). Word concreteness is a representation of the average level of concreteness, or non-abstractness for all words in a sample. As L2 learners tend to acquire concrete words at earlier stages of linguistic development (Crossley, Salsbury, & McNamara, 2009), raters may interpret increased word concreteness as an indication of lower linguistic ability. Lastly, word meaningfulness provides an estimate of the level of association among content words in each sample. Less meaningful words (e.g., cosine, squib) involve fewer associations, while more meaningful words (e.g., quick, quiet) evoke associations with a wider range of other words. As learners' linguistic ability improves, they begin to use less meaningful words with fewer associations (Salsbury et al., 2011), suggesting that this might impact raters' judgements of comprehensibility and nativeness. For each of these measures, Coh-metrix averages psycholinguistic ratings for all content words in each sample.

VocabProfile measures

VocabProfile and TexLex Compare, available through Compleat Lexical Tutor (Cobb, 2016), are online tools used to compare lexis in user-supplied discourse samples with various preestablished word lists, and to calculate the degree of lexical overlap between two user supplied samples.

– Four separate word lists from the VocabProfile database were used to calculate additional lexical sophistication measures: Browne, Culligan and Phillips' (2013) New General Service List (NGSL), Browne, Culligan and Phillips'

(2013) New Academic Word List (NAWL), Neufeld and Billuroglu's (2005) Billuroglu-Neufeld List (BNL), and Nation's (2012) British National Corpus and the Corpus of Contemporary American English word lists (BNC/COCA). These four word lists were targeted on the assumption that they might provide unique coverage statistics due to their disparate nature. With the exception of the NAWL, which provides a single lexical measure based on the entire 965-word list, all remaining lexical sophistication measures were divided into separate variables indicating coverage for specific frequency bands. For example, BNL sophistication was divided into measures representing coverage of the three most frequent bands (*BNL_1, BNL_2, BNL_3*), as well as the percentage of words appearing in each transcribed speech sample that did not appear in the BNL (*BNL_Off*). All overlap measures were calculated as percentage of token overlap. As previous research (e.g., Crossley, Cobb, & McNamara, 2013) has indicated that more lexically proficient L2 users produce less frequent words, the working assumption was that greater incidence of words from the later frequency bands of each word list would be associated with greater lexical proficiency, and potentially increased comprehensibility and nativeness.

Table 1. Summary of lexical measures

Category	Analysis tool	Measure
Lexical Diversity	Coh-Metrix	MTLD
Connectives	Coh-Metrix	Causal connectives Logical connectives Additive connectives
Situation Model	Coh-Metrix	Causal verb frequency Causal verbs and causal particles Verb overlap
Syntactic Complexity	Coh-Metrix	Average number of modifiers per noun phrase
Syntactic Pattern Density	Coh-Metrix	Frequency of -ing verbs Frequency of infinitives
Lexical Sophistication	Coh-Metrix	Word Length Word frequency Word age of acquisition Word familiarity Word concreteness Word meaningfulness
	VocabProfile	NAWL NGSL_1, NGSL_2, NGSL_3, NGSL_Off BNL_0, BNL_1, BNL_2, BNL_Off BNC/COCA_1, BNC/COCA_2, BNC/COCA_3, BNC/COCA_Off

Identification of lexical factors underlying holistic judgements of comprehensibility and nativeness followed a two-step process. First, Pearson correlation coefficients were calculated between all lexical measures and the two target constructs, separately for each task. Second, those lexical measures with the strongest correlations (identified in step one) were used as predictor variables in stepwise multiple regressions to uncover which of these measures accounted for significant proportions of variance in target construct scores.

Results

Comprehensibility and nativeness ratings

Interrater reliability for the 10 raters (Cronbach's alpha) met or exceeded the preestablished benchmark of .70–.80 (Larson-Hall, 2010) in the picture description task for both comprehensibility ($a=.79$) and nativeness ($a=.77$), and in the TOEFL integrated task for both comprehensibility ($a=.81$) and nativeness ($a=.81$). Therefore, the 10 individual ratings for comprehensibility and nativeness in each task were averaged to attain a single mean score for each construct in each sample. Cronbach's alpha was selected as a measure of reliability in the present study since it is one of the most widely applied measures of reliability (Field, 2009), is considered appropriate for studies involving multiple judges assessing the same construct(s) (Larson-Hall, 2010), and has a recognized interpretation scale. Correlations between comprehensibility and nativeness ratings were high for both the picture description ($r=.93$) and TOEFL integrated ($r=.94$) tasks.

The strong correlation between ratings of comprehensibility and nativeness suggests that raters may have had difficulty clearly differentiating these two constructs, at least as they were presented in this study using orthographic transcriptions, although previous research using audio recordings of L2 speech has consistently suggested that they are, in fact, two separate yet partially overlapping constructs (Trofimovich & Isaacs, 2012; Munro & Derwing, 2009), However, the fact that raters provided overall lower ratings for nativeness than comprehensibility, in the picture description task (445 vs. 541), t $(96)=21.02$, $p<.001$, $r=.91$, and in the TOEFL integrated task (440 vs. 536), t $(96)=20.78$, $p<.001$, $r=.90$ indicates that nativeness was evaluated more strictly than comprehensibility, which is consistent with previous research using audio recordings. Based on these findings, it would seem that, although comprehensibility and nativeness are highly correlated, comprehensibility can be viewed as a more easily attainable goal than nativeness (at least as judged by L1 English raters using transcribed speech samples). No significant differences, in terms of rater perceptions of L2 spoken ability, were found across the two tasks.

Lexical variables and rated constructs

As a first step in exploring the role of lexical variables in L1 English raters' assessments of comprehensibility and nativeness, Pearson correlation coefficients were computed using all 29 lexical measures and the two target constructs, separately for each task. Before running these correlations, all data was reviewed to ensure relatively normal distributions, as well as a lack of missing data and outliers. Tables 2 and 3 summarize all significant relationships for the two tasks ($a = .05$); a full list of all correlation coefficients for each task can be found in Appendix B.

Table 2. Significant correlations for the picture description task

Category	Lexical variable	Comprehensibility	Nativeness
Situation Model	Verb overlap	−.27**	−.24*
Syntactic Pattern Density	-ing verbs	−.26*	−.19
	Infinitives	.22*	.27**
Lexical Sophistication	Word Length	.31*	.30**
	Word frequency	−.28**	−.25*
	Word age of acquisition	−.22*	−.20
	Word concreteness	−.24*	−.21*
	Word meaningfulness	.23*	.19
	NAWL	.24*	.28**
	BNL_0	−.23*	−.18
Lexical Diversity	MTLD	.31**	.32**

Note.
* $p < .05$. ** $p < .05$, two-tailed.

For the picture description task (Table 2), 11 lexical variables spanning situation model, syntactic pattern density, lexical sophistication, and diversity were found to be significantly correlated with comprehensibility. Seven lexical variables covering these same categories were also significantly correlated with nativeness in this task. While all seven variables correlated with nativeness were also significantly associated with comprehensibility, four variables uniquely correlated with comprehensibility helped to distinguish these two constructs (-ing verbs, word age of acquisition, word meaningfulness, BNL_0).

In the TOEFL integrated task (Table 3), there was a wider range of variables associated with comprehensibility and nativeness. This was largely due to the

Table 3. Significant correlations for the TOEFL integrated task

Category	Lexical variable	Comprehensibility	Nativeness
Connectives	Additive connectives	−.25 *	−.21 *
Situation Model	Causal verbs	.17	.22 *
	Causal verbs & particles	.23 *	.18
Syntactic Complexity	Modifiers per noun phrase	−.25 *	−.22 *
Lexical Sophistication	Word Length	.29 **	.33 **
	Word frequency	−.22 *	−.20
	Word age of acquisition	.27 **	.27 **
	Word familiarity	−.27 **	−.30 **
	NGSL_3	.31 **	.34 **
	NGSL_Off	−.23 *	−.24 *
	BNL_1	−.17	−.21 *
	BNL_2	.23 *	.27 **
	BNL_Off	−.33 **	−.35 **
	BNC/COCA_Off	−.34 **	−.36 **
	BNC/COCA _2	.21 *	.23 *
	BNC/COCA _3	.26 **	.28 **
Lexical Diversity	MTLD	.49 **	.44 **

Note.
* $p<.05$. ** $p<.01$, two-tailed.

increased number of lexical sophistication measures that could be linked to rater assessments of each construct. In fact, while only two lexical sophistication measures from VocabProfile showed significant correlations with either construct in the picture description task, eight such measures were significantly associated in the TOEFL integrated task. For comprehensibility, 15 measures were significantly correlated with this construct. Similarly, for nativeness, 15 (largely identical) variables showed significant correlations. The lexical variables distinguishing between the two constructs in this task included causal verbs, causal verbs and particles, word frequency, and BNL_1.

Lexical predictors of comprehensibility and nativeness

To better understand the relationship between the two constructs and the lexical variables associated with them (see Tables 2 and 3), stepwise multiple regressions were conducted, with comprehensibility and nativeness used as criterion factors. However, given that multiple lexical variables had significant associations with each construct, a more restrictive inclusion criterion for predictor variables was implemented. First, a minimum correlation coefficient of .25 ($p < .05$) was set, as this is considered the benchmark for small associations in L2 research (Plonsky & Oswald, 2014). Second, the relationship between the identified predictor variables was checked for multicollinearity. As in previous research (e.g., Crossley et al., 2011), a typical collinearity threshold of .70 was set. Using this threshold, two predictor variables in the TOEFL integrated task were found to be highly correlated: NGSL_Off and BNC/COCA_Off ($r = .80$, $p < .01$). Therefore, only BNC/COCA_Off was included in the subsequent regressions, as this variable held the stronger correlation with both comprehensibility and nativeness. As the maximum correlation between all other predictors was .53, collinearity was not considered a problem. To ensure comparability, all predictors meeting the inclusion criteria in a specific task were included in the regressions for both constructs. For the picture description task, seven variables covering the categories of lexical sophistication (word length, word frequency, NAWL), situation model (verb overlap), syntactic pattern density (-ing verbs, infinitives), and diversity (MTLD) met the inclusion criteria and were therefore used in the multiple regressions for this task. For the TOEFL integrated task, 10 variables, spanning the categories of lexical sophistication (word length, word age of acquisition, word familiarity, NGSL_3, BNL_2, BNC/COCA_Off, BNC/COCA_3), connectives (additive connectives), syntactic complexity (modifiers per noun phrase), and diversity (MTLD) met the criteria. Results of multiple regression analyses are summarized in Tables 4 and 5.

Table 4. Results of the multiple regression for the picture description task

Criterion variable	Predictors	R^2	ΔR^2	B	95% CI	t	p
Comprehensibility	Word length (sophistication)	0.10	0.10	0.446	[0.051, 0.841]	2.24	.027
	MTLD (diversity)	0.15	0.05	0.003	[0.001, 0.005]	2.48	.015
	-ing verbs (density)	0.20	0.05	−0.005	[−0.010, −0.001]	−2.43	.017
Nativeness	MTLD (diversity)	0.10	0.10	0.003	[0.001, 0.005]	3.09	.003
	NAWL (sophistication)	0.18	0.08	0.033	[0.009, 0.057]	2.70	.008
	Infinitives (density)	0.21	0.03	0.001	[0.000, 0.003]	2.02	.047

In the picture description task (Table 4), comprehensibility was mainly predicted by dimensions of lexical sophistication (word length) and syntactic pattern density (-ing verbs), with 15% of total variance explained. An additional 5% of variance was accounted for by lexical diversity (MTLD). Beta values showed that, while both word length and MTLD were positively associated with comprehensibility, -ing verbs represented a negative association. Nativeness was primarily predicted by a combination of lexical diversity (MTLD) and lexical sophistication (NAWL), with 18% of total variance explained. Syntactic pattern density (infinitives) accounted for an additional 3% of total variance. In contrast, for the TOEFL integrated task (Table 5), both comprehensibility and nativeness were predicted by the same four variables covering categories of lexical diversity (MTLD) and sophistication (NGSL_3, BNC/COCA_Off, BNL_2), with similar individual contributions and a comparable total variance explained (39% for comprehensibility, 41% for nativeness).

Table 5. Results of the multiple regression for the TOEFL integrated task

Criterion variable	Predictors	R^2	ΔR^2	B	95% CI	t	p
Comprehensibility	MTLD (diversity)	0.24	0.24	0.005	[0.003, 0.007]	4.40	.000
	NGSL_3 (sophistication)	0.32	0.08	0.026	[0.012, 0.040]	3.70	.000
	BNC/COCA_Off (sophistication)	0.36	0.04	−0.013	[−0.024, −0.002]	−2.36	.020
	BNL_2 (sophistication)	0.39	0.04	0.008	[0.001, 0.014]	2.29	.024
Nativeness	MTLD (diversity)	0.19	0.19	0.004	[0.002, 0.006]	3.58	.001
	NGSL_3 (sophistication)	0.30	0.11	0.029	[0.015, 0.042]	4.30	.000
	BNC/COCA_Off (sophistication)	0.36	0.06	−0.016	[−0.026, −0.005]	−2.96	.004
	BNL_2 (sophistication)	0.41	0.06	0.009	[0.003, 0.015]	2.89	.005

Discussion

The current study explored the relationship between lexical measures of L2 speech and ratings of L2 comprehensibility and nativeness across two tasks. In contrast to previous research which has targeted trained raters' evaluations of L2 speech (e.g., Crossley et al., 2010; Lu, 2012), this study examined assessments by naïve L1 English post-secondary students, using transcribed samples as opposed to audio recordings to target the specific contribution of lexis to judgements of comprehensibility and nativeness.

Lexical correlates of speech ratings

In response to the first research question, which asked which lexical factors underlie holistic judgements of L2 English comprehensibility and nativeness, a measure of lexical diversity (MTLD) was identified as the sole common significant predictor of scores for both constructs in each task. Therefore, lexical diversity can be seen as an important element in naïve L1 undergraduate raters' assessments of L2 comprehensibility and nativeness (at least in English when evaluated using transcribed samples), regardless of task type. As L2 speakers' linguistic ability advances, they are able to use a greater proportion of different word types in their oral productions, thereby improving comprehensibility and nativeness. Naïve raters (the present study) and trained assessors (e.g., Lu, 2012) likely recognize greater lexical diversity as a sign of improved linguistic ability, associating it with enhanced comprehensibility and nativeness. While lexical diversity (measured by MTLD) was a significant predictor of raters' assessments of comprehensibility and nativeness for each task, all remaining lexical variables identified through multiple regressions were specific to individual tasks, constructs, or both.

Comprehensibility versus nativeness

For the second research question, which assessed level of independence between comprehensibility and nativeness, analyses revealed important task differences. For the picture description task (see Table 4), there appeared to be a separation in lexical correlates of the two constructs, with comprehensibility tied to measures of word length, MTLD, and *-ing* verb forms, whereas nativeness was associated with MTLD, NAWL, and frequency of infinitives. However, in the TOEFL integrated task (Table 5), raters' assessments of comprehensibility and nativeness were largely indistinguishable, with identical sets of lexical variables, covering dimensions of diversity (MTLD) and sophistication (BNC/COCA_Off, BNL_2), accounting for comparable amounts of variance in each construct. Potential reasons for the indistinguishability of target constructs in the TOEFL task are discussed in *Task Effects* (below).

For the picture description task, beyond the common measure of lexical diversity (MTLD), significant predictors of comprehensibility included lexical sophistication (word length) and pattern density (-ing verbs) measures, accounting for

15% of total variance. The significance of word length as an indicator of comprehensibility likely relates to higher ranked samples containing longer, more complex vocabulary representative of later stages of linguistic development. Post hoc analyses support this conclusion, with negative correlations identified between average word length and frequency bands for the most commonly occurring words in the NGSL ($r=-.28$, $p<.01$), BNL ($r=-.37$, $p<.01$), and BNC/COCA ($r=-.37$, $p<.01$). Thus, for L2 speakers to be judged as more comprehensible, they must move beyond a reliance on the most frequently occurring vocabulary and begin using rarer items that are generally characterized by longer words acquired in later stages of L2 development.

The negative correlation between the frequency of *-ing* verbs and comprehensibility highlighted in the multiple regressions also points to a nontrivial association. With the picture description task requiring speakers to create a narrative describing a series of events, repeated use of *-ing* forms may have decreased comprehensibility by obscuring ordering and suggesting overlapping continuous actions. For example, in the sample text below, which received a high *-ing* verb score, sequencing is primarily achieved through fronting information regarding the picture being referred to, while the actual language suggests ongoing events.

– …the first picture is showing a corner of two roads… two people are coming from two different directions those people are intersecting the third picture is showing the collision so they are holding their head he is putting his spectacles on… (Participant 97)

In this sample, each image is described in correct order; however, the repeated use of *-ing* verbs suggests a series of overlapping actions. Thus, repeated use of *-ing* forms may have decreased comprehensibility by obscuring event sequencing. In contrast, the sample text listed below, which elicited a low *-ing* verb score, makes use of more varied tense and aspect forms to describe the same series of images.

– two people are walking the street… they contact each other suddenly… they didn't see each other from the other side after that when they want to continue their way… when they arrive home… when they open their bags… (Participant 119)

Here, the speaker employs very few *-ing* forms, essentially limiting their use to the initial description of the first image, relying on other tense and aspect forms for remaining events in the narrative. This increased use of infinitives, which was also a significant predictor of nativeness in this task, indicates preference for specific (uninflected) verb forms can influence perceptions of comprehensibility and nativeness by L1 raters.

For the TOEFL integrated task, lexical differences between comprehensibility and nativeness were less clear-cut. In fact, the same four variables (MTLD, NGSL_3, BNL_2, BNC/COCA_Off), covering dimensions of lexical diversity and sophistication, emerged as significant predictors in the multiple regressions for both constructs, with comparable total amounts of variance explained. These variables point to the same common underlying factor – notably, rich and varied lexis. With NGSL_3 and BNL_2 representing indices of lexical use beyond the most frequent bands in each word list, these measures indicate the importance of diverse and sophisticated lexis to assessments of L2 ability. To appropriately implement these measures, it is important to look not only at coverage of the most frequent bands from each list, but also off-list measures, as BNC/COCA_Off was found to hold a significant negative correlation with scores of comprehensibility and nativeness in the TOEFL integrated task.

However, caution should be taken when using off-list measures in assessments of linguistic ability as their value is, at least partially, task dependent. For example, while both the TOEFL integrated task and picture description task contained a small percentage of off-list words (1.76% and 5.95%, respectively), BNC/COCA_Off was only significantly associated with target construct scores in the TOEFL integrated task (Table 3). Based on manual reviews of transcribed samples, this discrepancy seems to result from differences in the lexical categories captured by off-list measures in each task. For the TOEFL integrated task, off-list words were exclusively related to the presence of false starts (e.g., *wha* as a false start for what), lexical inventions (e.g., *huriness*), and grammatical mistakes (e.g., *feeled* instead of felt). In contrast, for the picture description task, moderately frequent use of proper nouns (used to label the imagined city in which the events were taking place or the main characters involved in the narrative) contributed to off-list measures but was unrelated to perceptions of linguistic ability. Thus, the value of BNC_COCA_Off as a metric of speaking performance in the TOEFL task likely reflected its ability to identify lexical errors that would negatively impact raters' assessments of comprehensibility and nativeness.

Task effects

In response to the third research question, which targeted possible task effects in the identified lexical correlates of comprehensibility and nativeness, two observations could be made. The first observation concerned the total variance accounted for by each of the regression models and the related importance of lexical sophistication measures in these models. For the picture description task, 20–21% of the total variance in target construct scores was explained by the predictor variables. For the TOEFL integrated task, the total variance explained by each model was substantially higher, with 39% and 41% of comprehensibility and nativeness scores accounted for by the predictors. The most likely cause of this discrepancy is differences in task complexity. According to the Cognition Hypothesis (Robinson, 2005), cognitively more demanding tasks, compared to simpler tasks, result in more elaborate language with richer and more complex vocabulary and grammar (e.g., Robinson, 2001). Thus, cognitively demanding tasks likely call for the use of more varied and sophisticated vocabulary, increasing the likelihood for communication difficulties to arise, at least with respect to the use of vocabulary, which may have led to greater total variance being explained by dimensions of lexical sophistication and diversity in the TOEFL integrated task. Conversely, for the picture description task, which arguably elicited simpler language constrained by the visual input, there may not have been as many lexical measures relevant to raters' assessments of comprehensibility and nativeness, which would account for the lower amount of total variance explained and fewer significant associations, particularly with lexical sophistication measures. In close relation to this point, it should also be noted that average length of response varied between the two tasks, with the TOEFL integrated task eliciting, on average, longer responses. This feature may have contributed to the differences in variance accounted for by providing raters with more linguistic data on which to base their assessments, thereby leading to a greater number of associated lexical indices being identified.

The second observation relates to differences in degree of construct independence discovered in the two tasks. While multiple regressions for the picture description task revealed a clear separation between lexical correlates of comprehensibly and nativeness, results from the TOEFL integrated task indicated greater overlap, and therefore increased difficulty in distinguishing between these two constructs. One possible reason for this finding is register-based differences. As with level of complexity, there is a clear difference in terms of the register required to respond to each task. The picture description task, which used a series of illustrated images depicting a nonacademic scenario to stimulate a narrative describing an event sequence, is unlikely to be found in any content-based postsecondary program. Conversely, the TOEFL integrated task, which aims to replicate a common event in a content-based academic program (i.e., reading and listening to academic discourse before displaying a coherent understanding of the material) requires a higher level of academic English, and speakers' use of this academic register might obscure lexical differences contributing to the comprehensibility-nativeness distinction.

Implications

In terms of implications for theory, findings of this study question the scope of distinction between comprehensibility and nativeness as partially overlapping yet independent constructs of L2 speech (at least when analyzed using automated lexical measures of transcribed speech). That the two target constructs were largely indistinguishable in a cognitively more complex task eliciting academic language suggests that this distinction is likely task-specific, in the sense that the linguistic dimensions relevant to each construct vary with the linguistic and cognitive demands of a given speaking task. In addition, because the majority of prior evidence for the independence of these constructs came from analyses of audio recordings of L2 speech (e.g., Saito et al., 2016), conclusions drawn from these studies may have been overtly influenced by pronunciation, prosody, and fluency related factors (e.g., speech rate, segmental errors). Thus, the distinction between comprehensibility and nativeness appears to be sensitive to the mode in which L2 speech is evaluated, because the use of transcribed samples in this study largely eliminated potential speech- and fluency-related factors. What emerges, then, is a complex relationship between linguistic correlates of comprehensibility versus nativeness, one that must be situated within task and register differences and identified in relation to both spoken and written features of discourse.

In terms practical implications, findings suggest that, regardless of the desired goal (comprehensibility or nativeness), L2 speakers would benefit from a focus on increasing depth and breadth of vocabulary knowledge, as lexical diversity (MTLD) was revealed as an important variable in L1 English raters' assessments of both constructs in each task. Although literature on this topic has defined breadth and depth in various ways (see Read, 2004), we follow Anderson & Freebody (1981) in using *breadth* to refer to the number of different words the user has at least some knowledge of, and *depth* as a representation of the level of knowledge the user has for each of these words. Thus, our findings suggest that it is important for L2 speakers to acquire both a wide range of vocabulary, and knowledge of important meaning characteristics that will enable them to effectively distinguish between related words in order to appropriately apply the correct term given the specific context in which they are operating.

This combination of breadth and depth ultimately allows for increased specificity, thereby leading to more effective communication. For example, in the picture description task, the items being held by the two central characters can be referred to in several ways (e.g., stuff, things, bags). However, use of the specific term, *suitcase*, provides a more precise meaning that likely improves understanding of the described situation. While L2 speakers may know several words that refer to the concept of '*items that can be carried*' (vocabulary breadth), they must combine this with depth of knowledge to select appropriate vocabulary that will aid understanding.

Although fluency related factors (e.g., speed of delivery, pausing, prosody) were largely eliminated from the present study in favor of a lexical focus, fluency is also likely play an important role in how L2 speakers are perceived. Thus, future studies should aim to explore the role of fluency in rater evaluations of spoken ability. Particularly in timed tasks, such as those used in the present research, speed of delivery is likely a contributing factor that influences perceptions of spoken ability and further contributes to comprehensibility and nativeness. However, in order to fully assess the role of fluency in relation to comprehensibility and nativeness, it will likely be necessary to use audio recordings as opposed to transcribed samples.

Another unique contribution of the current dataset is in demonstrating that the importance of lexical diversity in evaluations of L2 speech is relevant to learners from multiple language backgrounds, as opposed to those from particular languages, such as French (Saito et al., 2015), Chinese (Lu, 2012), and Japanese (Saito et al., 2016). Thus, a focus on improving lexical diversity is suitable for instruction in linguistically diverse L2 classrooms (e.g., English for Academic Purposes courses at English-medium universities). Perhaps the most important implication is that, at least when it comes to more academically oriented tasks, it may not be necessary for teachers and learners to adopt an exclusive focus on either comprehensibility or nativeness, at least when viewed from a lexical perspective using transcribed speech samples, as these constructs were largely overlapping in the TOEFL integrated task (when assessed in relation to a collection of automated lexical measures). Therefore, teachers and learners can focus on both goals simultaneously when working within the academic register.

Limitations and future research

As this study was largely exploratory, there are several limitations that should be addressed in future research. In relation to task differences in assessments of comprehensibility and nativeness, findings were based on the analysis of only two tasks. With the TOEFL integrated task being more cognitively demanding, featuring a different register, and generally including responses of greater length than those from the picture description task, it remains to be seen which of these factors (complexity, register, or length of response), and to which extent, resulted in the observed task effects. Therefore, future research should attempt to delineate the specific role each of these elements may play in assessments comprehensibility and nativeness. Additionally, while efforts were made to ensure careful reading of all samples (5-second delay before scores could be assigned), we could not gauge the percentage of text raters deemed necessary to complete before assigning a score. Thus, in future research it may prove useful to use eye-tracking devices to better understand the reading patterns of raters.

Despite efforts to analyze speech performances from users with a range of L2 speaking proficiencies, this study may not have captured the full range of desired L2 ability levels. Because the L2 speakers were mainly graduate students with moderate to advanced levels of L2 speaking proficiency (as assessed through the speaking component of the TOEFL iBT and IELTS), future research should target a wider range educational levels and language proficiencies, ideally controlling for potentially influencing factors, such as age of acquisition (Moyer, 2013), length of residence (Derwing & Munro, 2013), and aptitude (Granena, 2014).

Because the L1 English raters were undergraduate students, they may have lacked sufficient familiarity with academic English to fully distinguish between perceptions of comprehensibility and nativeness in this register. With their exposure to academic English relatively limited, the concept of academic English – in terms of its comprehensibility, nativeness, or both – may not yet have fully developed. This is supported by the high degree of correlation between comprehensibility and nativeness. As a result, future research on assessments of comprehensibility and nativeness in L2 academic discourse may wish to target more academically experienced raters (i.e., postgraduate students). Furthermore, to more fully explore this issue, it may prove necessary to have raters assess a third construct (e.g., proficiency) that could be used as a separate marker of linguistic ability in determining construct independence.

Lastly, our reliance on automated extraction of freely available lexical measures may also be interpreted as a potential limitation due to some of the algorithms used. As the Coh-metrix text parser can result in questionable word class assignments (for example, see Note 2 on gerunds versus -ing word forms), caution should be taken when interpreting results based on these measures. This limitation may be especially important in light of the type of data analyzed in this study, as the use of transcribed samples (including false starts, word repetitions, and repairs) could further limit the accuracy of the text-parser, and thereby the indices being used. Thus, the conclusions drawn from this study should be applied with caution and additional measures must continue to be explored. For example, while this study only targeted single word-lexical measures, future studies should aim to explore multi-word measures (e.g., Wray, 2002) as a way of better understanding important linguistic features related to evaluations of L2 English spoken ability.

Conclusion

Three main conclusions can be drawn from this study. First, as a substantial portion of variance in L1 English raters' assessments of comprehensibility and nativeness was attributed to lexical (and associated grammatical) features, variables targeting these aspects should be regarded as an important indicator of L2 spoken ability, at least when viewed from a lexical perspective using automated measures of L2 speech. Second, since each task was linked to a separate set of significant predictors (beyond the common measure of lexical diversity), it can be concluded that task type and register hold a strong influence on which linguistic features are significantly associated with assessments of comprehensibility and nativeness. Finally, construct independence for comprehensibility and nativeness appears to depend on task type (complex/academic vs. simple/nonacademic), the mode in which spoken discourse is evaluated (spoken vs. written), and also likely on raters' experience with the target language domain (academic vs. nonacademic). Given these findings, researchers might wish to refine, and perhaps even leave aside, the debate about comprehensibility and nativeness in academic tasks, at least when viewed from a lexical point of view in relation to automatically extracted metrics, as these constructs, at least in this dataset, were seen as largely indistinguishable in raters' assessments.

References

Abercrombie, D. (1949). Teaching pronunciation. *ELT Journal*, 3, 113–122. https://doi.org/10.1093/elt/III.5.113

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Gutherie (Ed.) *Comprehension and teaching: Researching reviews* (pp. 77–117). Newark, DE. International Reading Association.

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

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Retrieved from <http://www.newgeneralservicelist.org>

Cobb, T. (2016). *Compleat Lexical Tutor* [computer program]. <http://www.lextutor.ca> (15 January 2016).

Crossley, S. A., Cobb, T., & McNamara, D. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41, 965–981. https://doi.org/10.1016/j.system.2013.08.002

Crossley, S. A., & 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. https://doi.org/10.1111/j.1467-9922.2009.00508.x

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. https://doi.org/10.1111/j.1467-9922.2010.00568.x

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

Crossley, S. A., 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. https://doi.org/10.5054/tq.2010.244019

Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26, 92–113.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, 99, 80–95. https://doi.org/10.1111/modl.12185

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.42

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. https://doi.org/10.1111/lang.12000

Derwing, T. M., Rossiter, M. J., & Ehrensberger-Dow, M. (2002). They speaked and wrote real good: Judgements of non-native and native grammar. *Language Awareness*, 11, 84–99. https://doi.org/10.1080/09658410208667048

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). L2 fluency: Judgements on different tasks. *Language Learning*, 54, 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Educational Testing Services (2004). Independent Speaking Scoring Rubrics. Retrieved from <http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf>

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27, 379–399. https://doi.org/10.1177/0265532210364407

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher*, 40, 223–234. https://doi.org/10.3102/0013189X11413260

Granena, G. (2014). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, 63, 665–703. https://doi.org/10.1111/lang.12018

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29, 24–49. https://doi.org/10.1093/applin/amm017

Jarvis, S., & Daller, M. (2013). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.47

Kang, O., Rubin, D., Pickering, L. (2010). Suprasegmental measures of accentedness and judgements of English language learner proficiency in oral English. *The Modern Language Journal*, 94, 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

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. https://doi.org/10.1016/j.system.2012.10.012

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32, 145–164. https://doi.org/10.1016/j.system.2004.01.001

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, 31, 279–284. https://doi.org/10.1177/0265532214526179

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and applications. *TESOL Quarterly*, 49, 757–786. https://doi.org/10.1002/tesq.194

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39, 369–377. https://doi.org/10.2307/3588485

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Review*, 96, 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7, 52–75. https://doi.org/10.1177/026553229000700105

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 235–244. https://doi.org/10.1093/ijl/3.4.235

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511794407

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34, 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *The Canadian Modern Language Review*, 66, 181–202. https://doi.org/10.3138/cmlr.66.2.181

Munro, M. J., & Derwing, T. M., Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, 52, 626–637. https://doi.org/10.1016/j.specom.2010.02.013

Nation, I. S. P. (2012). The BNC/COCA word family lists (17 September 2012). Unpublished paper. Available at <www.victoria.ac.nz/lals/about/staff/paul-nation>

Neufeld, S., & Billuroğlu, A. (2005). In search of the critical lexical mass: How 'general' is the GSL? How 'academic' is the AWL? Available at <http://www.academia.edu/download/2951027/8985atj340yff5z.pdf>

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, 12, 237–241. https://doi.org/10.1016/j.tics.2008.02.014

Perfetti, C. A., Landi, N., & Oakhill, J. The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford: Blackwell.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effects sizes in L2 research. *Language Learning*, 64, 878–912. https://doi.org/10.1111/lang.12079

Read, J. (2004). Plumbing the depths. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in second language* (pp. 209–227). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.10.15rea

Révész, A., Michel, M., & Gilabert, R. (2015). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgements. *Studies in Second Language Acquisition*, 38(4), 703–737. https://doi.org/10.1017/S0272263115000339

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43, 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Saito, K., Trofimovich, P., Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37, 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2015). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition*, 38, 677–701. https://doi.org/10.1017/S0272263115000297

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27, 343–360. https://doi.org/10.1177/0267658310395851

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15, 905–916. https://doi.org/10.1017/S1366728912000168

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners*. Beijing: Foreign Language Teaching and Research Press.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary. *Behavioural Research Methods, Instruments and Computers*, 20, 6–11. https://doi.org/10.3758/BF03202594

Wray, A. (2002). *Formulaic language and the lexicon*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Address for correspondence

Randy Appel
Waseda University
1-104 Totsukamachi, Shinjuku-ku
Tokyo, 169-8050
Japan
r_appel@outlook.com