MLJ The Modern Language Journal

The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives

XIAOFEI LU
The Pennsylvania State University
Department of Applied Linguistics
301 Sparks Building
University Park, PA 16802
Email: XXL13@psu.edu

This study was an examination of the relationship of lexical richness to the quality of English as a second language (ESL) learners' oral narratives. A computational system was designed to automate the measurement of 3 dimensions of lexical richness, that is, lexical density, sophistication, and variation, using 25 different metrics proposed in the language acquisition literature. This system was used to analyze large-scale data from the Spoken English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005) together with the vocd utility of the Computerized Language Analysis programs (MacWhinney, 2000), which offers an additional measure of lexical variation, the D measure (Malvern, Richards, Chipere, & Durán, 2004; McKee, Malvern, & Richards, 2000). This comprehensive analysis allowed us to identify measures that correlate strongly with the raters' judgments of the quality of ESL learners' oral narratives, as well as to understand the relationships among these measures. This research provides ESL teachers and researchers with a robust tool for assessing the lexical richness of ESL language samples and insights into how lexical richness measures may be effectively used as indices of the quality of ESL learners' speaking task performance.

LEXICAL RICHNESS IS MANIFEST IN SECOND language (L2) use in terms of the sophistication and range of an L2 learner's productive vocabulary (Wolfe-Quintero, Inagaki, & Kim, 1998). It has been recognized as an important construct in L2 teaching and research, as it is directly related to the learner's ability to communicate effectively in both spoken and written form. A large variety of lexical richness measures have been proposed in the language acquisition literature. Many L2 development studies have examined the extent to which these measures, along with measures of accuracy, fluency, and grammatical complexity, can be used as reliable and valid indices of the learner's developmental level or overall proficiency in an L2 (see, e.g., Larsen-Freeman, 1978; Wolfe-Quintero et al., 1998; Zareva, Schwanenflugel, & Nikolova, 2005).

The L2 testing and assessment literature, in contrast, has focused primarily on evaluating the relationship between lexical richness and the quality of learners' writing or speaking task performance (Laufer & Nation, 1995; Vermeer, 2000; Yu, 2010). This trend is not surprising, as the rating scales of various language proficiency tests have explicitly stated a link between the two. For example, the rating scale of the Test for English Majors Band 4 (TEM–4), administered by the Chinese Ministry of Education to second-year English majors in 4-year colleges in China, specifies that for the speaking tasks, a test taker needs to demonstrate appropriate usage of a large range of vocabulary to be rated "excellent." Similarly, the rating scale of the Michigan English Language Assessment Battery (MELAB) also specifies that for an "excellent speaker," "idiomatic, general, and specific vocabulary range is extensive" (English Language Institute, 2001).

The Modern Language Journal, 96, ii, (2012)
DOI: 10.1111/j.1540-4781.2011.01232.x
0026-7902/11/190–208 $1.50/0

This study focused on the relationship of lexical richness to the quality of English as a second language (ESL) learners' oral narratives. A computational system was designed to automate the measurement of three dimensions of lexical richness—that is, lexical density, sophistication, and variation—using 25 different metrics proposed in the language acquisition literature. This system was used to analyze large-scale TEM–4 spoken test data from the Spoken English Corpus of Chinese Learners (SECCL; Wen, Wang, & Liang, 2005) together with the vocd utility of the Computerized Language Analysis (CLAN) programs (MacWhinney, 2000), which offers an additional measure of lexical variation, the D measure (Malvern, Richards, Chipere, & Durán, 2004; McKee, Malvern, & Richards, 2000). In this comprehensive analysis, we assessed the effect of each of the 26 measures on the raters' judgments of the quality of test takers' oral narratives, as well as their relationships with each other.

As Yu (2010) has pointed out, this type of research is needed to establish empirically whether and to what extent the link claimed in the rating scales of many language proficiency tests between lexical richness and the scores assigned to test takers' task performance exist in actual test performance data. The current study complements existing literature in this area in important ways. The unavailability of computational tools to automate some of the measures proposed and the labor-intensiveness of manual analysis have limited the number of measures that could be examined and the size of the sample that could be analyzed in many previous studies. This poses a major challenge for comparing the performance of the wide array of measures examined in different studies, as well as for understanding how these measures relate to each other, as there is considerable inconsistency and variability among previous studies in terms of choice and definition of measures, language task used, sample size, corpus length, and so on (Wolfe-Quintero et al., 1998).

With the capacity to automate the computation of lexical richness using 26 different measures, this study was free of such inconsistency and variability and led to valuable insights into how these measures compare with and relate to each other as indices of the quality of test takers' speaking task performance, which would have been hard to obtain through smaller scale studies or research syntheses. It is clear that linguistic features other than lexical richness, such as accuracy, fluency, grammar, intonation, pronunciation, and sociolinguistics, all come into play in determining the quality of learners' speaking task performance, and it is important to understand the relative contribution of these factors in the raters' global score assignments (see, e.g., De Jong & van Ginkel, 1992; Higgs & Clifford, 1982; Iwashita, Brown, McNamara, & O'Hagan, 2008). However, our hope was that in-depth knowledge gained about the individual factor, lexical richness, would contribute to understanding the full picture.

LEXICAL RICHNESS MEASURES

Following Read (2000), we conceptualize lexical richness as a multidimensional feature of a learner's language use that consists of the following four interrelated components: lexical density, lexical sophistication, lexical variation, and number of errors in vocabulary use. As mentioned earlier, we acknowledge the relationship between lexical errors, or accuracy in general, and task performance quality (see, e.g., Engber, 1995; Wolfe-Quintero et al., 1998), but we did not focus on lexical errors in this research. In this section, we provide a brief overview of the lexical richness measures that were evaluated in this study. We also review empirical studies of the relationship of lexical richness to general language proficiency or the quality of learners' writing and speaking task performance. Researchers have used a large number of different measures primarily because they have attempted to approach lexical richness from many different perspectives in the search for valid and reliable indices of learners' language proficiency or task performance quality. In some cases, researchers have also examined measures that apply to one or more particular lexical categories instead of to the total vocabulary, for example, measures of verb sophistication or noun variation, out of interest in understanding learners' lexical behavior with respect to these specific categories and the relationship of such behavior to their language proficiency or task performance quality.

Lexical Density

Lexical density, originally coined by Ure (1971), refers to the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text. Spoken texts reportedly have a lower lexical density than written texts (Halliday, 1985; Ure, 1971), and lexical density in learners' oral productions may be affected by such factors as plannedness and degree of interactiveness (O'Loughlin, 1995; Ure, 1971). Both Linnarud (1986) and Engber (1995) uncovered nonsignificant correlations between lexical density and holistic ratings of L2 writing. However, the relationship of lexical density to the quality of L2 learners' speaking task performance has not been extensively examined.

Although lexical words generally refer to open-class, as opposed to closed-class grammatical words, there is notable variability in the specific definitions in previous studies. For example, for lexical adverbs, O'Loughlin (1995) counted all adverbs of time, manner, and place, whereas Engber (1995) counted "adverbs with an adjectival base, especially those with an *–ly* suffix" (p. 145). Some studies failed to provide any explicit definition of lexical words (e.g., Ure, 1971). For this study, we defined lexical words as nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, "be," and "have"), and adverbs with an adjectival base, including those that can function as both an adjective and adverb (e.g., "fast") and those formed by attaching the *–ly* suffix to an adjectival root (e.g., "particularly"). Lexical density was the ratio of the number of lexical words (N_{lex}) to the number of words (N), as shown in Table 1.

Lexical Sophistication

Lexical sophistication, also known as lexical rareness, measures "the proportion of relatively unusual or advanced words in the learner's text" (Read, 2000, p. 203). Table 1 contains five different measures of lexical sophistication proposed in previous studies. Linnarud (1986) and Hyltenstam (1988) both calculated lexical sophistication as the ratio of the number of sophisticated lexical words (N_{slex}) to the total number of lexical words (N_{lex}) in a text. Linnarud defined sophisticated lexical words as English words introduced at Grade 9 or later in the Swedish educational system and noted significant differences between compositions written by native English speakers and Swedish learners of English. Hyltenstam defined sophisticated lexical words as words beyond the 7,000 most frequent Swedish words and found no significant difference between native and near-native Swedish writers.

Laufer (1994) and Laufer and Nation (1995) proposed the Lexical Frequency Profile (LFP) that looks at the proportion of word types in a text covered by the list of the first 1,000 most frequent words, the list of the second 1,000 most frequent words, the university word list (Xue & Nation, 1984), and none of these lists, respectively. Considering words among the first 2,000 most frequent as the "basic 2,000" words and those that are not the "beyond 2,000" words, Laufer (1994) reported significant differences in the proportion of these types of words between pre- and post-compositions written by students in two university classes. Laufer and Nation (1995) claimed that LFP is a valid and reliable measure of lexical richness, but Meara (2005) challenged this claim by showing that it may not be able to "pick up modest changes in vocabulary size" (p. 32) (see Laufer, 2005). Although one can examine the details of LFP in different ways, the model also provides a measure of lexical sophistication, computed as the ratio of the number of sophisticated word types (T_s; i.e., the "beyond 2,000" words) to the total number of word types in a text (Wolfe-Quintero et al., 1998).

Harley and King (1989) proposed a verb sophistication measure, which was computed as the ratio of the number of sophisticated verb types (T_{sverb}) to the total number of verbs (N_{verb}) in a text. They defined sophisticated verbs as verbs not on the list of 20 or 200 most frequent French verbs in two analyses, respectively, and detected significant differences between native and L2 writers in both cases. Wolfe-Quintero et al. (1998) recommended using Chaudron and Parker's (1990) squared version of this measure or an adaptation of Carroll's (1964) corrected type–token ration to reduce the sample size effect (see discussion on lexical variation below).

As is evident from this discussion, in addition to differences in the specific measure proposed, there is also considerable variability in how sophisticated words are defined across previous studies. This variability may have partially contributed to the mixed results reported on the relationship between lexical sophistication and language proficiency. In the present study, we examined all five measures of lexical sophistication discussed previously and listed in Table 1. We counted words, lexical words, and verbs as sophisticated if they were not on the list of the 2,000 most frequent words generated from the British National Corpus (BNC; Leech, Rayson, & Wilson, 2001). In counting word types, we considered different inflections of the same lemma (e.g., "go," "goes," "going," "went," and "gone") as one type.

Lexical Variation

Lexical variation, also labeled lexical diversity (see, e.g., Malvern et al., 2004; Yu, 2010) or lexical range (Crystal, 1982), refers to the range of a learner's vocabulary as displayed in his or her language use. Malvern et al. (2004) and Wolfe-Quintero et al. (1998) provided a thorough discussion of the many different measures of lexical variation that exist.

TABLE 1
Measures of Lexical Density and Sophistication

Measure	Code	Formula	Examples
Lexical Density	LD	N_{lex}/N	Engber (1995)
Lexical Sophistication-I	LS1	N_{slex}/N_{lex}	Linnarud (1986), Hyltenstam (1988)
Lexical Sophistication-II	LS2	T_s/T	Laufer (1994)
Verb Sophistication-I	VS1	T_{sverb}/N_{verb}	Harley & King (1989)
Corrected VS1	CVS1	$T_{sverb}/\sqrt{2N_{verb}}$	Wolfe-Quintero et al. (1998)
Verb Sophistication-II	VS2	T^2_{sverb}/N_{verb}	Chaudron & Parker (1990)

An intuitively straightforward measure of lexical variation is the number of different words (NDW) used in a language sample, which has proved to be a potentially useful measure of child language development (Klee, 1992; Miller, 1991). However, NDW is dependent on the length of the language sample, and some form of standardization may be desired in comparing samples of different lengths. A simple method is to truncate all samples to a set length that is no longer than the shortest sample (see, e.g., Thordardottir & Ellis Weismer, 2001). Malvern et al. (2004) discussed two other standardization procedures. In both cases, one randomly selects a set of subsamples of the same size from the sample, and averages the NDW for these subsamples to approximate the expected value of NDW. In one procedure, each subsample consists of a standard number of words randomly selected from the sample. In the other, each subsample contains a standard number of consecutive words from the sample with a random starting point. They argued that truncation is wasteful of data and showed that different standardization procedures might change the overall results in comparing the NDW of different samples.

Type–token ratio (TTR), that is, the ratio of the number of word types (T) to the number of words (N) in a text (Templin, 1957), is widely used in both first language (L1) and L2 acquisition studies. However, this measure has been criticized for its sensitivity to sample size, as the ratio tends to decrease as the size of the sample increases (Arnaud, 1992; Hess, Sefton, & Landry, 1986; Richards, 1987). Mean segmental TTR (MSTTR) (Johnson, 1944) constitutes one way to improve TTR, which is computed by dividing a sample into successive segments of a given length and then calculating the average TTR of all segments. Although MSTTR reduces the sample size problem, Malvern et al. (2004) pointed out that it is not without its disadvantages; for example, samples do not always divide into standard-sized segments, resulting in some waste of data.

Other transformations of TTR include Corrected TTR (CTTR; Carroll, 1964), Root TTR (RTTR; Guiraud, 1960), Bilogarithmic TTR (LogTTR; Herdan, 1964), and the Uber Index (Dugast, 1979), as listed in Table 1. Vermeer (2000) examined TTR and these four transformations using spontaneous speech data of L1 and L2 children learning Dutch. He reported that the validity and reliability of these measures in his data were unsatisfactory and recommended examining measures of lexical sophistication. However, Daller, Van Hout, and Treffers-Daller (2003) examined TTR and RTTR in conjunction with Laufer and Nation's (1995) notion of LFP as measures of lexical variation of spontaneous oral productions of college-level Turkish–German bilinguals, and reported significant correlations between lexical variation based on advanced lexical items and the participants' language proficiency.

Malvern et al. (2004) and McKee et al. (2000) presented a new transformation of TTR, the D measure, which they showed to be "a robust measure of lexical diversity which is not a function of sample size in the way TTR and its simple transformations are" (Malvern et al., 2004, p. 60). A full discussion of the rather complex mathematical derivation of the D measure is beyond the scope of this report. The reader can see Malvern et al. (2004) and McKee et al. (2000) for details. Briefly, however, the model produces a family of ideal curves of TTR against tokens (N), represented by the mathematical expression in Figure 1. Given a language sample, one finds the ideal curve that best matches the actual curve of TTR against tokens (N) for the sample. The value of the parameter D for this best fit curve is considered the index of lexical variation for the sample.

Malvern et al. (2004) demonstrated the methodological advantage of D over other measures of lexical variation by analyzing texts from various domains. Yu (2010) also reported that data from the MELAB showed significant positive

FIGURE 1
Equation for the D Measure

$$TTR = \frac{D}{N}\left[(1 + 2\frac{N}{D})^{\frac{1}{2}} - 1\right]$$

correlations between the D measure and test takers' general proficiency, as well as the quality of their writing and speaking task performance. Jarvis (2002), however, reported mixed results for the relationship of the D measure to holistic ratings of English-as-a-foreign-language written narratives. He determined that although the overall correlation was consistently significant, groups with the highest D means exhibited the lowest correlation between the D measure and the holistic ratings.

In addition to its use to assess the variation of total vocabulary, TTR and some of its transformations have also been used to evaluate the variation of specific classes of words. Several studies examined lexical word variation, that is, the ratio of the number of lexical word types to the total number of lexical words in a text (Casanave, 1994; Engber, 1995; Hyltenstam, 1988; Linnarud, 1986). Results pertaining to its relationship to language proficiency have been mixed. For example, Engber (1995) reported a significant relationship to holistic ratings of ESL compositions by intermediate and advanced learners, but Linnarud (1986) detected no significant relationship between holistic scores and this measure for advanced learners.

Harley and King (1989) examined a verb variation measure, computed as the ratio of the number of verb types to the total number of verbs in a text, and uncovered significant differences between native and L2 French writers in timed writing tasks. As was the case for the verb sophistication measure, Wolfe-Quintero et al. (1998) again recommended using Chaudron and Parker's (1990) squared version of this measure or an adaptation of Carroll's (1964) CTTR to reduce the sample size effect.

McClure (1991) examined five ratios with the same denominator—the number of lexical words—and the following five numerators, respectively: number of verb types, noun types, adjective types, adverb types, and modifier (both adjective and adverb) types. She compared Spanish–English bilingual and English monolingual 4th- and 9th-grade students and discovered significant differences for her noun, adverb, and modifier variation measures, weak differences for the adjective variation measure, but no difference for the verb variation measure.

As the previous discussion shows, several different conceptualizations of lexical variation exist. Although some measures have obvious problems and have been criticized more than others, there is no consensus among researchers concerning a single best measure. Results for most of the measures that have been examined in more than one study tend to be mixed, and due to the variability in research design and measure definition, it is not easy to compare the results. In this study, we examined all of the 20 measures of lexical variation discussed previously and summarized in Table 2 to determine how they compare with and relate to each other as indices of the quality of ESL learners' oral narratives. For the NDW-based measures and MSTTR, we used 50 as the standard number, as the shortest sample in our dataset was 105 words long.

METHOD

Research Questions

This study focused on the relationship of lexical richness to the raters' judgments of the quality of oral narratives produced by Chinese learners of English taking TEM–4. Specifically, we addressed the following four main research questions.

1. How does lexical density relate to the raters' judgments of the quality of test takers' oral narratives?
2. How do the different measures of lexical sophistication compare with and relate to each other as indices of the quality of test takers' oral narratives?
3. How do the different measures of lexical variation compare with and relate to each other as indices of the quality of test takers' oral narratives?
4. How do lexical density, lexical sophistication, and lexical variation compare with and relate to each other as indices of the quality of test takers' oral narratives?

Data

The data used in this study were selected from the Spoken English Corpus of Chinese Learners (Wen et al., 2005) and included transcriptions of 408 test takers' oral productions in the TEM–4 spoken test administered nationwide to second-year English majors in 4-year colleges in China[1]

TABLE 2
Measures of Lexical Variation

Measure	Code	Formula
Number of Different Words	NDW	T
NDW (first 50 words)	NDW–50	Tin the first 50 words of sample
NDW (expected random 50)	NDW–ER50	Mean T of 10 random 50-word samples
NDW (expected sequence 50)	NDW–ES50	Mean T of 10 random 50-word sequences
Type–Token Ratio	TTR	T/N
Mean Segmental TTR (50)	MSTTR–50	Mean TTR of all 50-word segments
Corrected TTR	CTTR	$T/\sqrt{2N}$
Root TTR	RTTR	$T/\sqrt{N}$
Bilogarithmic TTR	LogTTR	$LogT/LogN$
Uber Index	Uber	$Log^2N/Log(N/T)$
D Measure	D	Based on D in Equation (1)
Lexical Word Variation	LV	T_{lex}/N_{lex}
Verb Variation-I	VV1	T_{verb}/N_{verb}
Squared VV1	SVV1	T_{verb}^2/N_{verb}
Corrected VV1	CVV1	$T_{verb}/\sqrt{2N_{verb}}$
Verb Variation-II	VV2	T_{verb}/N_{lex}
Noun Variation	NV	T_{noun}/N_{lex}
Adjective Variation	AdjV	T_{adj}/N_{lex}
Adverb Variation	AdvV	T_{adv}/N_{lex}
Modifier Variation	ModV	$(T_{adj} + T_{adv})/N_{lex}$

during the period of 1996 to 2002. The test consisted of three tasks: retelling a story, talking about a given topic, and role-playing. For the first task, test takers' vocabulary use was likely to be strongly influenced by the vocabulary in the story, as they heard the story twice, could take notes, and retold the story right away. For the third task, the corpus unfortunately did not contain the information needed for distinguishing utterances produced by a test taker and his or her partner. Therefore, only the data for the second task were used in this study. In this task, a test taker was asked to read a topic, prepare for 3 minutes, and then produce a 3-minute oral narrative.

Test takers' performances were tape-recorded and were rated as follows. The tapes were randomly divided into groups of 32 to 35 and rated group by group. The pool of raters consisted of English teachers selected from 4-year colleges across the country, and raters were trained to rate the tapes on a 100-point scale following the rating scale provided in Table 3. As the test was administered to second-year English majors in 4-year colleges nationwide, the pool of raters was also fairly large. In each group, all tapes were independently rated by the same two raters and were subjected to re-evaluation by a quality control committee if significant interrater discrepancy existed. The quality control committee then averaged the two scores for each tape, re-evaluated some of the tapes scored among the highest and lowest, and ranked the tapes within the group. These rankings were used to determine the final ratings (excellent, good, pass, or fail) of the test takers. Wen et al. (2005) did not report interrater reliability for the scores, but noted that this grouping strategy had proven to ensure reliability of the final ratings and scoring efficiency. The corpus contained the test takers' rankings within their groups, but not their actual scores or final ratings.

The set of files that transcribe the oral narratives produced by test takers for task two were retrieved from the corpus. Each file contained a header for the following information: mode (spoken), test (TEM–4), school level (second year in college), test year (1996 through 2002), group (a number identifying the group the test taker is in), task type (task two), gender, and rank (the test taker's rank within his or her group). However, rank information appeared in only 408 of the 1,148 files in the corpus, and our final dataset consisted of these files only. As Table 4 shows, these files represented 12 groups in the years 1999 to 2002, with 3 groups from each year and 32 to 35 participants in each group. The length of the samples in these files ranged from 105 to 476 words. The topics used in these years were as follows (Wen et al., 2005):

1. 1999: Describe one of your experiences in which you had a burning desire to learn something.

TABLE 3
Rating Scale for Oral Narratives in TEM–4 Spoken Test

Rating	Narrative	Pronunciation and Intonation	Grammar and Vocabulary
Excellent (90–100)	Excellent relevance and structure; rich content; no unnecessary pauses.	Accurate and clear pronunciation; natural intonation.	Few grammatical errors; appropriate word usage and large vocabulary range.
Good (80–89)	Good relevance and structure; good content; a few unnecessary pauses that do not hinder communication.	Accurate and clear pronunciation; good intonation.	Some minor grammatical errors; good word usage and vocabulary range.
Pass (60–79)	Inadequate and partially irrelevant content; some unnecessary pauses that may hinder communication.	Fair accuracy and clarity in pronunciation; fair intonation.	Some major grammatical errors that do not hinder communication; fair word usage and vocabulary range.
Fail (0–59)	Poor structure; bare and irrelevant content; excessive unnecessary pauses that hinder communication.	Poor accuracy and clarity in pronunciation; poor intonation.	Excessive grammatical errors that hinder communication; excessive word usage errors and small vocabulary range.

TABLE 4
Summary of Dataset

Group[2]	Year	Group ID	Female	Male	Total
1	1999	21	26	7	33
2	1999	44	23	10	33
3	1999	66	23	10	33
4	2000	11	27	7	34
5	2000	29	25	7	32
6	2000	65	30	4	34
7	2001	01	28	7	35
8	2001	08	26	9	35
9	2001	30	32	3	35
10	2002	01	25	9	34
11	2002	31	29	6	35
12	2002	50	29	6	35
Total	-	-	323	85	408

2. 2000: Describe the most unforgettable birthday party you've ever had.

3. 2001: Describe a teacher of yours whom you found unusual.

4. 2002: Describe an embarrassing situation in which you got very angry.

Automating Lexical Richness Measurement

All text files in the dataset were preprocessed using a python script before being analyzed. The script removed the following from each file: the header; fillers, including "ah," "eh," "er," "mm," "oh," and "um"; and extra elliptical marks—only one set of elliptical marks was retained if two or more consecutive sets were used to mark disfluency pauses. In the original text files, a word form containing an obvious pronunciation error (e.g., "need" mispronounced as "leed") or a grammatical error involving tense, aspect, or number was enclosed in a pair of angle brackets and preceded by the correct form provided by the transcriber (e.g., "the sun has risen <rised>"). These tagged erroneous word forms were also removed. Errors involving inappropriate word usage, however, were not tagged in the original text files and were not corrected. The cleaned text files were saved separately.

The D measure was calculated using the vocd utility of CLAN. A cleaned text file was first converted to the CHAT[3] (Codes for the Human Analyses of Transcripts) format (MacWhinney, 2000) using the textin utility in CLAN, and then subjected to three times of vocd analyses. The average of the three analyses served as the D value of the sample.

A computational system, Lexical Complexity Analyzer,[4] was designed to compute the other 25 measures. The system took a cleaned text file as input and outputted a numeric score for each of the 25 measures. This process entailed the following steps. A cleaned text file was first part-of-speech (POS) tagged using the Stanford tagger (Toutanova, Klein, Manning, & Singer, 2003), which assigns every token in the language sample a label that indicates its part-of-speech category, for example, adjective, adverb, and so on. The POS-tagged sample was then lemmatized using MORPHA (Minnen, Carroll, & Pearce, 2001), a robust morphological analyzer for English that returns the lemma and inflection of a word, given the word form and its part of speech. Next, the lemmatized sample was processed by a python script, which computed the values of the 25 measures. To this end, the script counted the number of types and tokens of the word classes involved in computing one or more of the measures, including:

1. the number of types of words (T), sophisticated words (T_s), lexical words (T_{lex}), sophisticated lexical words (T_{slex}), verbs (T_{verb}), sophisticated verbs (T_{sverb}), nouns (T_{noun}), adjectives (T_{adj}), and adverbs (T_{adv}); and
2. the number of tokens of words (N), lexical words (N_{lex}), sophisticated lexical words (N_{slex}), and verbs (N_{verb}).

Punctuation did not count as words. Since the sample was lemmatized, we were able to treat different inflections of the same lemma (e.g., “go,” “goes,” “going,” “went,” and “gone”) as the same type. As discussed previously, a word was considered a lexical word if it was a noun, verb (excluding modal verb, auxiliary verb, “be,” and “have”), adjective, or lexical adverb. An adverb was considered a lexical adverb if it also appeared as an adjective in the BNC word list (Leech et al., 2001), or if it consisted of an adjectival root (i.e., an adjective on the BNC word list) and the *–ly* suffix. We considered a word, lexical word, or verb sophisticated if it was not among the 2,000 most frequent words in the BNC word list. The type and token counts were used to compute directly 22 of the 25 measures other than NDW–50, NDW–ER50, and NDW–ES50, using the formulas listed in Tables 1 and 2. For NDW–50, the number of word types in the first 50 words of the sample was counted. For NDW–ER50, the script generated 10 independent random subsamples, each of which consisted of 50 words randomly selected from the sample, and averaged the number of word types in the 10 subsamples.[5] For NDW–ES50, the script generated 10 independent subsamples, each of which consisted of a consecutive sequence of 50 words from the sample with a random starting point, and averaged the number of word types in the 10 subsamples.

TABLE 5
Cut Points for Classifying Test Takers Based on Rankings Within Group

Group Size	Level A	Level B	Level C	Level D
32 (1 Group)	1–8	9–16	17–24	25–32
33 (3 Groups)	1–8	9–16	17–24	25–33
34 (3 Groups)	1–8	9–16	17–25	26–34
35 (5 Groups)	1–8	9–17	18–26	27–35
Total Samples	96	102	103	107

Analysis

As the test takers’ actual scores were not available, Spearman’s rank order correlation (rho) was calculated for each lexical richness measure and test takers’ rankings for each of the 12 groups. A fixed-effect meta-analysis[6] (see, e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001) was then conducted using Comprehensive Meta-Analysis software to combine the results of the 12 individual groups and derive the average correlation. Specifically, the average correlation was calculated as a weighted mean of the correlations in the 12 groups, where the weight for each group was equal to the inverse variance (i.e., $1/\sigma^2$, where σ^2 denotes variance) of the group’s correlation.[7] To assess student performance variation in the lack of the test takers’ final ratings, we adopted a secondary way to analyze the data, in which all test takers were proportionally divided into four levels based on their rankings, following the scheme summarized in Table 5. For example, for groups in which $n = 33$, students ranked among 1–8, 9–16, 17–24, and 25–33 were included in Levels A, B, C, and D, respectively. With the assumption that these levels represent different proficiency levels in a similar way as the final ratings presumably do, this allowed us to examine whether significant differences existed in the mean values of the different lexical richness measures among the four levels. For assessing the relationships among the lexical richness measures, all the data in the 12 groups were pooled and Pearson’s correlations were computed. Following Wolfe-Quintero et al. (1998), we interpreted correlation coefficients as high ($r \geq .650$), moderate ($.450 \leq r < .650$), and weak ($.250 \leq r < .450$). In addition, we interpreted coefficients lower than .250 as very low ($.100 \leq r < .250$) and trivial ($r < .100$).

TABLE 6
Descriptive Statistics of All Measures Over the Entire Dataset

Measure	Minimum	Maximum	Mean	*SD*
Words	105	476	295.480	67.603
Words/ Minute	35	158.667	98.493	22.534
LD	.320	.510	.414	.033
LS1	.070	.460	.227	.070
LS2	.150	.410	.262	.043
VS1	.000	.260	.074	.051
CVS1	.000	1.190	.330	.218
VS2	.000	2.810	.312	.388
NDW	56	205	119.830	22.530
NDW–50	25	46	34.880	3.638
NDW–ER50	30.200	41.600	36.829	2.033
NDW–ES50	23.600	41.200	34.228	2.851
TTR	.250	.600	.414	.058
MSTTR	.480	.803	.686	.052
CTTR	3.300	6.960	4.942	.567
RTTR	4.660	9.840	6.989	.803
LogTTR	.770	.900	.843	.021
Uber	15.400	45.390	26.152	4.634
D	23.940	101.380	55.561	13.389
LV	.310	.840	.573	.080
VV1	.320	.920	.578	.116
SVV1	3.230	28.260	13.415	4.322
CVV1	1.270	3.760	2.556	.419
VV2	.090	.370	.193	.038
NV	.300	.900	.590	.096
AdjV	.030	.230	.108	.035
AdvV	.000	.120	.042	.018
ModV	.060	.300	.150	.038

RESULTS AND DISCUSSION

Descriptive Statistics and the Effect of Fluency

Table 6 shows the descriptive statistics of sample length (in terms of words), number of words spoken per minute (words/minute), and each of the 26 measures examined in this study. For each measure, the table contains its minimum and maximum values, as well as its mean and standard deviation across all samples in our dataset. For example, the length of the samples ranged from 105 to 476 words, with a mean of 295.480 words and a standard deviation of 67.603 words; the LS1 value of the samples ranged from .070 to .460, with a mean of .262 and a standard deviation of .043.

Table 7 summarizes the effects of sample length, words/minute, and the lexical density and sophistication measures. Sample length and words/minute both correlated with test takers' rankings significantly in 9 of the 12 groups, and the combined correlations were significant ($\rho = .437$, $p < .001$). Table 8 shows the means of sample length, words/minute, and the lexical density and sophistication measures among the four levels. A one-way ANOVA indicated significant differences in mean sample length and words/minute among the four levels (F(3,404) = 28.335, $p < .001$). In either case, the means for Levels B and C were close to each other, but both were significantly lower than the mean for Level A and significantly higher than the mean for Level D. This finding indicates that narratives ranked higher were generally longer and produced at a faster speaking rate than those ranked lower. Although we did not focus on the effect of fluency and its relationship to lexical richness in this study, we noted in passing that these results revealed a significant effect of fluency, as measured by words spoken per minute, for oral proficiency.

Research Question 1

The results yielded no evidence of a significant correlation between lexical density and test takers' rankings, except for Group 8 ($\rho = .371$, $p = .028$). The combined correlation from the meta-analysis was nonsignificant ($\rho = .011$, $p = .836$). A one-way ANOVA showed no significant difference in the mean values of lexical density among the samples at the four levels ($F(3,404) = .896$, $p = .443$). These results are similar to the ones reported in previous L2 writing studies (Engber, 1995; Linnarud, 1986) and suggest that the proportion of lexical words in an oral narrative does not appear to relate to its quality.

Research Question 2

For the two general lexical sophistication measures, LS1 and LS2, none of the 12 groups showed a significant correlation between them and test takers' rankings. Results of the meta-analysis also revealed nonsignificant combined correlations for both LS1 ($\rho = .011$, $p = .836$) and LS2 ($\rho = .048$, $p < .355$). A one-way ANOVA yielded no significant difference in the mean values of LS1 ($F(3,404) = .681$, $p = .564$) among the four levels. Between-level differences in the mean values of LS2 ($F(3,404) = 2.736$, $p = .043$) were not significant after the Bonferroni correction. These results differ from the ones reported in previous L2 writing studies (e.g., Laufer, 1994; Linnarud, 1986), and suggest a difference in the role lexical sophistication plays in spoken and written proficiency.

TABLE 7
Effects of Lexical Density and Sophistication Measures

	Individual Groups*			Meta-Analysis**				
Measure	$p < .01$	$p < .05$	$p < .10$	Average rho	95% CI Lower	95% CI Upper	Fisher's Z	p-value
Words	6	3	0	.437	.351	.516	9.038	.000
Words/Minute	6	3	0	.437	.351	.516	9.038	.000
LD	0	1	0	.011	−.091	.112	.206	.836
LS1	0	0	0	.048	−.054	.148	.925	.355
LS2	0	0	0	.050	−.052	.150	.961	.336
VS1	0	0	2	.133	.032	.231	2.583	.010
CVS1	0	1	2	.166	.066	.263	3.225	.001
VS2	0	1	2	.165	.064	.262	3.205	.001

*Individual Groups = Number of groups in which the p-value of the rho (two-tailed) between the measure and test takers' rankings is in the following three ranges respectively: $p < .01$, $.01 \leq p < .05$, $.05 \leq p < .10$.
**The alpha value for the analysis was adjusted to 0.05/28, or 0.0018, by the Bonferroni correction for multiple tests, as 28 tests were done (including those reported in Table 10).

TABLE 8
Comparison of Level Means of Lexical Density and Sophistication Measures

	Level Mean				ANOVA*	
Measure	A ($n = 96$)	B ($n = 102$)	C ($n = 103$)	D ($n = 107$)	F	Sig.
Words	336.160	295.950	297.760	256.340	28.335	.000
Words/ Minute	112.052	98.650	99.252	85.446	28.335	.000
LD	.417	.415	.409	.414	.896	.443
LS1	.227	.235	.221	.225	.681	.564
LS2	.261	.272	.256	.260	2.736	.043
VS1	.072	.086	.067	.073	2.629	.050
CVS1	.343	.383	.299	.297	3.722	.042
VS2	.314	.401	.274	.262	2.760	.042

*The alpha value for the analysis of variance (ANOVA) was adjusted to 0.05/28, or 0.0018, by the Bonferroni correction for multiple tests, as 28 tests were done (including those reported in Table 11).

The three verb sophistication measures had stronger relationships to test takers' rankings. Among these, VS1 was the most problematic. None of the groups showed a significant correlation. The combined correlation ($\rho = .133$, $p = .010$) was not significant after the Bonferroni correction, nor were the differences in the mean values of VS1 among the four levels ($F(3{,}404) = 2.629$, $p = .050$). The two transformed verb sophistication measures performed better than VS1, with a significant combined correlation for both CVS1 ($\rho = .166$, $p < .001$) and VS2 ($\rho = .165$, $p < .001$).

These results offer some support for Wolfe-Quintero et al.'s (1998) suspicion that the transformations of VS1 have a stronger effect for proficiency, as the effect of the sample size is reduced. Nevertheless, we noted that the combined correlations between the two transformed measures and test takers' rankings, although statistically significant, were both low. In addition, a one-way ANOVA indicated nonsignificant between-level differences in the mean values of CVS1 ($F(3{,}404) = 3.722$, $p = .012$) and VS2 ($F(3{,}404) = 2.760$, $p = .042$) after the Bonferroni correction.

Table 9 summarizes the correlations among the lexical sophistication measures. There were high correlations among the three verb sophistication measures. The strongest correlation was found between CVS1 and VS1 ($r = .966$, $p < .001$), followed by that between CVS1 and VS2 ($r = .935$, $p < .001$). This is probably not surprising, as CVS1 is a transformation of VS1 and is mathematically similar to VS2. LS1 and LS2 were moderately correlated ($r = .637$, $p < .001$)

TABLE 9
Correlations Among Lexical Sophistication Measures

	LS1	LS2	VS1	CVS1	VS2
LS1	1.000	.637**	.456**	.414**	.381**
LS2	.637**	1.000	.391**	.382**	.350**
VS1	.456**	.391**	1.000	.966**	.909**
CVS1	.414**	.382**	.966**	1.000	.935**
VS2	.381**	.350**	.909**	.935**	1.000

**Correlation is significant at the .01 level (2-tailed).

but generally only weakly correlated with the three verb sophistication measures. However, LS1 showed slightly stronger correlations with the three verb sophistication measures (r = .381–.456, $p < .001$) than LS2 (r = .350–.391, $p < .001$). This was the case perhaps because the denominators in LS1 and the verb sophistication measures were all token counts, whereas that in LS2 was a type count.

Research Question 3

Table 10 summarizes the effects of the lexical variation measures, and Table 11 compares the mean values of the lexical variation measures among the four levels. A number of lexical variation measures demonstrated significant correlations with test takers' rankings.

We first examined the four measures that gauge the number of different words in the sample. Among these, NDW showed the strongest relationship to test takers' rankings: 11 of the 12 groups showed significant correlations between the two, and the combined correlation (ρ = .526, $p < .001$), while moderate, was significant. The mean values of NDW decreased linearly across the four levels, and a one-way ANOVA revealed highly significant between-level differences ($F(3,404) = 41.675$, $p < .001$). These results indicated that although NDW depended on the length of the sample, it was a useful predictor of the quality of timed oral narratives. NDW–50 showed a much weaker effect, with a significant correlation for only one group. The combined correlation was significant but low (ρ = .176, $p < .001$). The between-level differences in the mean values of NDW–50 ($F(3,404) = 4.757$, $p = .003$) were not significant after the Bonferroni correction. The other two standardized measures, NDW–ER50 and NDW–ES50, both correlated significantly with test takers' rankings in five groups, and both had weak, but significant combined correlations (ρ = .282/.319, $p < .001$). Between-level differences in the mean values of NDW–ER50 ($F(3,404) = 11.719$, $p < .001$) and NDW–ES50 ($F(3,404) = 13.122$, $p < .001$) were both significant, with the means decreasing linearly from Level A to Level D. The superiority of these two standardization procedures to NDW–50 supports Malvern et al.'s (2004) argument that truncating samples at an arbitrary point is not a desirable method for standardization, as it can lead to a significant loss of data.

Table 12 shows the correlations among the four NDW-based measures and sample length (number of words). NDW and sample length were strongly correlated ($r = .808$, $p < .001$), as were NDW–ER50 and NDW–ES50 ($r = .731$, $p < .001$). NDW was also moderately correlated with NDW–ER50 ($r = .579$, $p < .001$) and NDW–ES50 ($r = .478$, $p < .001$). Correlations between other pairs of measures were all weak to trivial.

We now turn to the seven measures based on TTR applied to the total vocabulary. The original TTR had nonsignificant correlations with test takers' rankings in all 12 groups, and the combined correlation ($\rho = –.038$, $p = .459$) was nonsignificant, as well. The mean values of TTR did not change linearly across the four levels. Most of the transformations of TTR showed a stronger effect than the original TTR. Among these, CTTR and RTTR had the strongest relationship to test takers' rankings, with significant correlations appearing in nine groups as well as weak, significant combined correlations (ρ = .430/.429, $p < .001$). The mean values for these measures decreased linearly across the four levels, and a one-way ANOVA revealed significant between-level differences in these means ($F(3,404)$ = 24.108/24.097, $p < .001$). MSTTR–50 also showed significant correlations in four groups and a weak, significant combined correlation ($\rho = .332$, $p < .001$). A one-way ANOVA revealed significant between-level differences in the mean values of MSTTR–50 ($F(3,404) = 14.899$, $p < .001$), which decreased linearly from Level A to Level C and remained at the same level at Levels C and D. Uber showed significant correlations in three groups. The combined correlation was low, but significant ($\rho = .204$, $p < .001$). A one-way ANOVA indicated significant between-level differences in the mean values for Uber ($F(3,404) = 6.499$, $p < .001$), with the means exhibiting the same pattern as MSTTR–50 in decreasing linearly from Level A to Level C and remaining at about the same level at Levels C and D. The only transformation of TTR that had no

TABLE 10
Effects of Lexical Variation Measures

	Individual Groups*			Meta-Analysis**				
Measure	$p < .01$	$p < .05$	$p < .10$	Average rho	95% CI Lower	95% CI Upper	Fisher's Z	p-value
NDW	10	1	0	.526	.449	.596	11.288	.000
NDW–50	1	0	2	.176	.077	.272	3.446	.001
NDW–ER50	2	3	1	.282	.186	.373	5.590	.000
NDW–ES50	3	2	3	.319	.225	.407	6.371	.000
TTR	0	0	0	–.038	–.139	.063	–.741	.459
MSTTR–50	2	2	3	.332	.239	.419	6.655	.000
CTTR	7	2	2	.430	.343	.509	8.863	.000
RTTR	7	2	2	.429	.343	.508	8.853	.000
LogTTR	0	0	0	.082	–.019	.182	1.586	.113
Uber	0	3	2	.204	.105	.299	3.999	.000
D	4	3	1	.352	.260	.437	7.088	.000
LV	1	0	0	.063	–.039	.163	1.214	.225
VV1	0	1	0	.047	–.054	.148	.909	.363
SVV1	2	3	1	.291	.196	.381	5.786	.000
CVV1	2	3	1	.290	.195	.380	5.762	.000
VV2	0	0	0	–.091	–.190	.011	–1.752	.080
NV	0	1	0	.015	–.087	.116	.282	.778
AdjV	0	1	0	.074	–.028	.174	1.422	.155
AdvV	0	1	3	.203	.104	.298	3.966	.000
ModV	0	0	3	.173	.073	.269	3.369	.001

*Individual Groups = Number of groups in which the p-value of the rho (two-tailed) between the measure and test takers' rankings is in the following three ranges respectively: $p < .01$, $.01 \leq p < .05$, $.05 \leq p < .10$.
**The alpha value for the analysis was adjusted to 0.05/28, or 0.0018, by the Bonferroni correction for multiple tests, as 28 tests were done (including those reported in Table 7).

significant effect was LogTTR, which had nonsignificant correlations in all 12 groups, as well as a nonsignificant combined correlation ($\rho = .082$, $p = .113$). The between-level differences in the mean values of LogTTR ($F(3,404) = 3.527$, $p = .015$) were also nonsignificant after the Bonferroni correction. The D measure revealed a stronger effect than the original TTR, as well, with significant correlations in seven groups, and a weak, significant combined correlation ($\rho = .352$, $p < .001$). The mean values of D decreased linearly from Level A to Level D, and the between-level differences were significant ($F(3,404) = 16.205$, $p < .001$).

Table 13 shows the correlations among these seven measures. The correlation coefficients ranged from moderate to perfect ($r = .525$–1.000). Among these, CTTR and RTTR were perfectly correlated. These two forms of transformation were essentially the same. The original TTR most strongly correlated with LogTTR ($r = .953$) and Uber ($r = .882$); the latter two also strongly correlated with each other ($r = .949$). Compared with the other five measures, MSTTR and D had weaker correlations with the other measures. However, both were highly correlated with CTTR, RTTR, and Uber, as well as with each other, and less strongly correlated with TTR and LogTTR.

The nine measures based on TTR applied to one or more word classes revealed mixed results. Among the seven nontransformed ratios, only AdvV ($\rho = .203$, $p < .001$) and ModV ($\rho = .173$, $p = .001$) had significant, albeit very low, combined correlations with the test takers' rankings, whereas the lexical word, verb, noun, and adjective measures had nonsignificant correlations with them. Similarly, the results for the one-way ANOVAs indicated that between-level differences in the mean values of AdvV ($F(3,404) = 5.807$, $p = .001$) and ModV ($F(3,404) = 2.970$, $p = 0.032$) were significant, with the means decreasing linearly from Level A to Level D, but between-level differences in the mean values of the other five measures were not significant. The two transformed verb measures, SVV1 ($\rho = .291$, $p < .001$) and CVV1 ($\rho = .290$, $p < .001$), had similar weak, significant

TABLE 11
Comparison of Level Means of Lexical Variation Measures

	Level Mean				ANOVA*	
Measure	A ($N = 96$)	B ($n = 102$)	C ($n = 103$)	D ($n = 107$)	*F*	Sig.
NDW	136.630	121.520	116.910	105.960	41.675	.000
NDW–50	35.780	35.310	34.100	34.410	4.757	.003
NDW–ER50	37.697	37.045	36.477	36.182	11.719	.000
NDW–ES50	35.459	34.659	33.591	33.324	13.122	.000
TTR	.412	.419	.400	.423	3.110	.026
MSTTR–50	.710	.696	.670	.672	14.899	.000
CTTR	5.281	5.017	4.810	4.695	24.108	.000
RTTR	7.468	7.095	6.802	6.640	24.097	.000
LogTTR	.846	.845	.837	.843	3.527	.015
Uber	27.456	26.825	24.963	25.485	6.499	.000
D	61.883	57.917	51.808	51.256	16.205	.000
LV	.578	.580	.560	.575	1.369	.252
VV1	.586	.585	.559	.583	1.267	.285
SVV1	15.183	13.758	12.751	12.141	10.119	.000
CVV1	2.724	2.592	2.497	2.427	10.122	.000
VV2	.187	.195	.190	.200	2.382	.069
NV	.586	.597	.592	.586	.285	.836
AdjV	.110	.110	.108	.103	.697	.554
AdvV	.047	.044	.040	.038	5.807	.001
ModV	.158	.154	.148	.142	2.970	.032

*The alpha value for the analysis of variance (ANOVA) was adjusted to 0.05/28, or 0.0018, by the Bonferroni correction for multiple tests, as 28 tests were done (including those reported in Table 8).

TABLE 12
Correlations Among Lexical Variation Measures Based on Number of Different Words (NDW)

	NDW	NDW–50	NDW–ER50	NDW–ES50	Words
NDW	1.000	.274**	.579**	.478**	.808**
NDW–50	.274**	1.000	.448**	.448**	.014
NDW–ER50	.579**	.448**	1.000	.731**	.160**
NDW–ES50	.478**	.448**	.731**	1.000**	.064
Words	.808**	.014	.160**	.064	1.000

**Correlation is significant at the .01 level (2-tailed).

TABLE 13
Correlations Among Lexical Variation Measures Based on Type–Token Ratio (TTR) of Entire Vocabulary

	TTR	MSTTR–50	CTTR	RTTR	LogTTR	Uber	D
TTR	1.000	.592**	.525**	.526**	.953**	.882**	.528**
MSTTR–50	.592**	1.000	.737**	.737**	.697**	.746**	.741**
CTTR	.525**	.737**	1.000	1.000	.719**	.845**	.805**
RTTR	.526**	.737**	1.000	1.000	.719**	.845**	.806**
LogTTR	.953**	.697**	.719**	.719	1.000	.949**	.656**
Uber	.882**	.746**	.845**	.845**	.949**	1.000	.762**
D	.528**	.741**	.805**	.806**	.626**	.762**	1.000

**Correlation is significant at the .01 level (2-tailed).

combined correlations. Between-level differences in the mean values for both SVV1 ($F(3{,}404) = 10.119$, $p < .001$) and CVV1 ($F(3{,}404) = 10.122$, $p < .001$) were significant, with the means decreasing linearly from Level A to Level D. The better performance of the transformed verb measures again suggests that these types of standardization procedures may improve the traditional TTR measure.

Table 14 presents the correlations among the nine measures in this group. Among these, LV correlated strongly with VV1 ($r = .786$) and NV ($r = .714$). The two transformations of VV1, SVV1, and CVV1 were strongly correlated ($r = .994$). AdjV and ModV also correlated strongly ($r = .886$). VV2 and AdvV did not show strong correlations with other measures.

Research Question 4

As discussed above, lexical density failed to show a significant relationship to test takers' rankings, two of the five measures of lexical sophistication, both of which measure verb sophistication, revealed significant but low combined correlations ($r = .133$–$.166$), and 13 of the 20 measures of lexical variation showed significant combined correlations that ranged from very low ($r = .173$) to moderate ($r = .526$). It appears, then, that at least on the basis of existing measures proposed in the language acquisition literature, lexical variation may have the strongest effect for the quality of ESL learners' oral narratives among the three dimensions of lexical richness.

To understand the relationships among the three dimensions of lexical richness, we examined two groups of measures separately. The first group consisted of eight measures that apply to the total vocabulary or all lexical words, including LD, LS1, LS2, NDW, NDW–ES50, CTTR, D, and LV. Correlations among these measures appear in Table 15. Lexical density had trivial to weak correlations with lexical sophistication and variation, with correlation coefficients ranging from .042 (with D) to .279 (with NDW–ES50). Similarly, the lexical sophistication measures revealed trivial to very low correlations with the lexical variation measures. The correlation coefficients ranged from .079 (between LS1 and LV) to .237 (between LS2 and CTTR). The second group consisted of the seven verb sophistication and variation measures. Correlation coefficients for these measures are in Table 16. Although these measures were more strongly correlated with each other than the measures in the first group, the correlations among them were at best weak. VV2 showed the lowest correlations with the verb sophistication measures, with correlation coefficients ranging from .167 (between CVS1 and VV2) to .188 (between VS2 and VV2). Correlations among the other verb sophistication and variation measures were all weak, with correlation coefficients ranging from .282 (between CVS1 and VV1) to .433 (between CVS1 and SVV1). Results from both groups indicated that lexical density, lexical sophistication, and lexical variation are indeed three different constructs.

CONCLUSION

This study investigated the relationship of 26 measures of lexical richness to the quality of ESL learners' oral narratives, as well as to each other, by computationally analyzing the lexical richness of 12 groups of TEM–4 spoken test data from the Spoken English Corpus of Chinese Learners. By examining the effect of lexical richness on the quality of oral narratives in 12 different groups and applying a meta-analysis of the results from these groups, we obtained more reliable and objective results than would have been possible through smaller scale, single-group studies. In addition, by investigating a large set of measures with a large dataset, we have avoided the inconsistency and variability found among previous studies in terms of choice and definition of measures, language task used, sample size, corpus length, and so on. This makes our findings less problematic than those obtained through research syntheses.

Implications for L2 Teaching, Learning, and Research

This large-scale investigation yielded a number of substantive findings with important implications for L2 teaching, learning, and research. First, the three dimensions of lexical richness posited in this and previous research did not correlate strongly with each other, which suggests that they are indeed different constructs. Of these dimensions, lexical variation correlated most strongly with the raters' judgments of the quality of ESL learners' oral narratives. No effect for lexical density emerged, and a very small effect was found for lexical sophistication. This constitutes evidence for the relationship between lexical variation and the quality of test taker's oral productions claimed in the rating scale of the TEM–4 spoken test. For vocabulary instruction and learning, this implies that teachers and learners should

TABLE 14
Correlations Among Lexical Variation Measures Based on Type–Token Ratio (TTR) of Word Classes

	LV	VV1	SVV1	CVV1	VV2	NV	AjdV	AdvV	ModV
LV	1.000	.786**	.452**	.457**	.524**	.714**	.511**	.208**	.558**
VV1	.786**	1.000	.639**	.649**	.443**	.283**	.443**	.153**	.475**
SVV1	.452**	.639**	1.000	.994**	.530**	.120*	.103*	.212**	.192**
CVV1	.457**	.649**	.994**	1.000	.535**	.119*	.105*	.215**	.196**
VV2	.524**	.443**	.530**	.535**	1.000	.324**	−.089	.173**	−.002
NV	.714**	.283**	.120*	.119*	.324**	1.000	.321**	.137**	.354**
AdjV	.511**	.443**	.103*	.105*	−.089	.321**	1.000	−.038	.886**
AdvV	.208**	.153**	.212**	.215**	.173**	.137**	−.038	1.000	.414**
ModV	.558**	.475**	.192**	.196**	−.002	.354**	.886**	.414**	1.000

*Correlation is significant at the .05 level (2-tailed).
** Correlation is significant at the .01 level (2-tailed).

TABLE 15
Correlations Among General Lexical Density, Sophistication, and Variation Measures

	LD	LS1	LS2	NDW	NDW–ES50	CTTR	D	LV
LD	1.000	.173**	.127**	.057	.281**	.236**	.165**	.042
LS1	.173**	1.000	.637**	.085	.103	.169**	.079	.178**
LS2	.127**	.637**	1.000	.162**	.166**	.237**	.141**	.196**
NDW	.057	.085	.162**	1.000	.455**	.811**	.078	.582**
NDW–ES50	.281**	.103	.166**	.455**	1.000	.720**	.492**	.731**
CTTR	.236**	.169**	.237**	.811**	.720**	1.000	.566**	.805**
D	.165**	.079	.141**	.078	.492**	.566**	1.000	.484**
LV	.042	.178**	.196**	.582**	.731**	.805**	.484**	1.000

**Correlation is significant at the .01 level (2-tailed).

TABLE 16
Correlations Among Verb Sophistication and Variation Measures

	VS1	CVS1	VS2	VV1	SVV1	CVV1	VV2
VS1	1.000	.966**	.909**	.403**	.371**	.366**	.174**
CVS1	.966**	1.000	.935**	.282**	.433**	.427**	.188**
VS2	.909**	.935**	1.000	.295**	.403**	.389**	.167**
VV1	.403**	.282**	.295**	1.000	.639**	.649**	.443**
SVV1	.371**	.433**	.403**	.639**	1.000	.994**	.530**
CVV1	.366**	.427**	.389**	.649**	.994**	1.000	.535**
VV2	.174**	.188**	.167**	.443**	.530**	.535**	1.000

**Correlation is significant at the .01 level (2-tailed).

focus more on the range and less on the degree of sophistication of vocabulary.

Second, among the 20 measures of lexical variation, the following nine performed the best and are, therefore, recommended for teachers and researchers to use for assessing lexical variation in learners' oral narratives: NDW, NDW–ER50, NDW–ES50, CTTR, RTTR, MSTTR–50, D, SVV1, and CVV1. These measures were all significantly correlated with test takers' rankings, with correlations ranging from .282 to .526. Although only low to moderate by our definition, these significant correlations were nevertheless substantial, given the numerous factors that raters took into account in judging the quality of an oral narrative, such as accuracy, fluency, grammar, intonation, pronunciation, sociolinguistics, and so on (see, e.g., Yu, 2010). In addition, all nine measures decreased linearly across the four performance levels. NDW showed the strongest effect among all of the measures examined. As previously discussed, NDW and sample length were strongly correlated, and the dependence of NDW on sample length presents a problem for comparing samples of different lengths. However, the results suggest that in the case of timed oral narratives, NDW is potentially a very useful measure to consider. The two standardized versions, NDW–ER50 and NDW–ES50, are good alternatives when the sample size effect needs to be carefully controlled.

Third, the patterns of correlations among different measures identified in this study can be useful in guiding the measure-selection process in cases in which teachers and researchers wish to consider more than one measure. In general, it is plausible to choose a small set of measures that approach lexical richness in different ways, that have significant effects for quality, and that are not strongly correlated with each other. For example, in assessing lexical variation, it may be useful to consider NDW, along with the D measure. These measures gauge lexical variation from different perspectives: One looks at the number of different words while the other is a transformation of the type–token ratio. Both are among the measures with a stronger effect for quality, and they have a relatively lower correlation with each other than with other measures of lexical variation that show a significant effect for quality.

Fourth, another useful observation that can be made from our results is that the transformed TTR measures proved superior to their untransformed counterparts. This is evident in the list of the measures of lexical variation showing the strongest effect for quality. Among the seven TTR measures applied to the total vocabulary, the three transformations of the original TTR, namely, CTTR, RTTR, and the D measure, performed the best. Also, among the nine TTR measures applied to one or more word classes, the two transformed verb measures, SVV1 and CVV1, exhibited the strongest effect. This observation suggests that, in cases in which some form of the TTR measure is needed, it is generally recommendable for researchers to consider a transformed version of it.

Finally, the computational system (Lexical Complexity Analyzer) and methodology presented in this research can be used to validate similar claims in other language tests, or more genrally, to investigate the relationship of the linguistic code, in particular the lexicon, to the quality of a spoken or written product. The computational tool may also prove useful to language teachers and researchers for assessing L2 lexical proficiency or tracking L2 lexical development.

Limitations and Future Research

Given the scope of this research and the nature of the dataset, a number of important issues were not addressed in this study. These issues will be taken up in future research. First, lexical richness was conceptualized as a multidimensional feature following Read (2000), and each of the 26 metrics was considered a measure of a particular dimension following the studies that proposed or utilized the metrics. The fact that measures generally tended to correlate much more strongly with measures within the same dimension than with those of a different dimension offers support for this conceptualization. However, it will be very useful to investigate whether the dimensions posited are the same as the ones that would emerge from a factor analysis.

Second, the corpus used in this study contained the students' rankings within their groups, but not their actual scores or final ratings. As such, we were unable to run regression models to eliminate measures contributing no real independent variance to student performance. In addition, the division of the students into four levels based on their rankings instead of actual scores was not without problems. For example, the difference in performance between Levels A and B could be much larger (or smaller) than that between Levels B and C. The research results should be interpreted with this limitation in mind.

Third, it would be interesting to use the computational tool and methodology to investigate the relationship between lexical richness and both speaking and writing proficiency using a range of different conceptualizations of quality or proficiency. Previous studies suggested that lexical richness differs for spoken and written texts (see, e.g., Halliday, 1985; Yu, 2010), and a replication of this study with large-scale written data hopefully would provide useful insight into the extent to which this is the case. In addition, given the information that is available in our dataset, the quality of ESL learners' oral narratives was operationalized as rankings within the random subgroups. It would be very useful to replicate this study with large-scale data that allow for other operationalizations of quality (e.g., actual performance ratings) or language proficiency in general (e.g., program-level or school-level). This latter case may enable us to examine empirically the extent to which the findings obtained for the particular group of learners included in our dataset can be extended to learners at other levels of proficiency.

Fourth, a systematic examination of the effects that task-related variables may have on lexical richness will be useful. For example, the topics that participants talked about varied from year to year, and a one-way ANOVA of the four different years indicated that topic had a significant effect ($p < .05$) for the mean values observed for all but one (i.e., NV) of the 26 lexical richness measures. A systematic evaluation of such effects is important for our understanding of the relationship between lexical richness and language proficiency. It is important to note, however, that these potential effects had minimal influence on the reliability of the results obtained in this study. This is the case because in the meta-analysis approach adopted, the 12 groups were examined individually and within each individual group and all participants performed on the same task, that is, they received the same instructions and were given the same amount of time to talk on the same topic.

Fifth, methodologically, the current computational system can be enhanced with flexible options for different definitions of various measures to meet the needs of different teachers and researchers. As an example, the BNC word list was used for defining sophisticated words because it was derived from a 100 million–word collection of written and spoken samples from a wide range of sources. However, in cases where small slices of the BNC word list do not represent the language that a particular group of learners are exposed to, teachers and researchers may wish to provide a better list tailored to their own students or participants.

Finally, and as discussed earlier, the quality of a speaking or writing product is clearly determined by a multiplicity of factors, including accuracy, fluency, grammatical complexity, lexical richness, sociolinguistics, and so on. It would be useful to examine the relationship of lexical richness to the many other factors.

NOTES

[1] According to the national syllabus for English majors (Steering Committee for Foreign Language Major Education in Institutions of Higher Education, 2000), by the time they sit for TEM–4, second-year English majors in 4-year colleges in China are expected to be able to recognize 5,500–6,000 words and use 3,000–4,000 of those words appropriately. They are also expected to be able to communicate effectively with native speakers on social occasions with correct pronunciation, natural intonation, and language that is pragmatically appropriate and free from major grammatical errors. The national syllabus also details expectations in other areas of English language proficiency.

[2] Group (Column 1) is an arbitrary identifier assigned to the different groups of data in our dataset, whereas Group ID (Column 3) is the actual identifier provided in the corpus for a group of data within a particular year. For example, Group 7 in our dataset represents Group 01 in the year 2001 in the corpus, and Group 10 in our dataset represents Group 01 in the year 2002 in the corpus.

[3] CHAT is the standard transcription system for the CHILDES database (MacWhinney, 2000). The system offers options for both basic discourse transcription and detailed phonological and morphological analysis. Files in the CHAT format can be analyzed by the CLAN programs.

[4] This system is implemented in python and runs on UNIX-like systems (UNIX, LINUX, and Mac OS). The Web-based version of the system can be accessed at http://www.personal.psu.edu/xxl13/download.html.

[5] Each of the 10 subsamples was drawn independently from the sample. In other words, after a subsample was drawn and the number of different words in it counted, all of the words in it were returned to the sample before the next subsample was drawn. Averaging the number of different words in 10 such subsamples allowed us to approximate the expected value of NDW and was therefore more reliable than counting the number of different words in one subsample only. The procedure for calculating NDW–ES50 was similar in that all words in a subsample were returned to the sample before the next subsample was drawn.

[6] There are two statistical models for combining results in a meta-analysis. The fixed-effect (as opposed to random-effects) model was appropriate here because the effect being examined across the 12 groups was the same, that is, Spearman's rank order correlations between the measures and test takers' rankings; in addition, homogeneity tests showed no significant statistical heterogeneity among the correlations found across the groups.

[7] More exactly, the average correlation for a particular measure across the 12 groups was calculated as follows. We first weighted the correlation for each group by the inverse of its variance, that is, $1/\sigma^2$. For example, if for the ith group, the correlation $\rho_\mathrm{i} = .472$ and its variance $\sigma_\mathrm{i}{}^2 = .021$, then the weight w_i of the correlation ρ_i was $1/\sigma_\mathrm{i}{}^2 = 1/.021 = 47.619$, and the weighted correlation $\rho_\mathrm{i}\text{'} = \rho_\mathrm{i} * w_\mathrm{i} = .472 * 47.619 = 22.476$. The average correlation across the 12 groups was calculated as the sum of the 12 weighted correlations divided by the sum of the 12 weights, that is, $\sum\rho_\mathrm{i}\text{'} \ / \sum w_\mathrm{i}$.

REFERENCES

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In

P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 133–145). London: MacMillan.

Borenstein, M., Hedges, L. V., Higging, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing*, *3*, 179–201.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition*, *12*, 43–64.

Crystal, D. (1982). *Profiling linguistic disability*. London: Edward Arnold.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*, 197–222.

De Jong, J. H. A. L., & van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In J. H. A. L. De Jong (Ed.), *The construct of language proficiency* (pp. 112–140). Philadelphia: Benjamins.

Dugast, D. (1979). *Vocabulaire et stylistique. I Théâtre et dialogue* [*Vocabulary and style. Vol. 1 Theatre and dialogue*]. Geneva, Switzerland: Slatkine-Champion.

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

English Language Institute. (2001). Michigan English Language Assessment Battery Speaking Test rating scale reference sheet. Ann Arbor: University of Michigan. Retrieved September 5, 2011, from http://www.lsa.umich.edu/eli/testing/melab/testtakers

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique* [*Problems and methods of statistical linguistics*]. Dordrecht, The Netherlands: D. Reidel.

Halliday, M. A. K. (1985). *Spoken and written language*. Melbourne, Australia: Deakin University Press.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*, 415–440.

Herdan, G. (1964). *Quantitative linguistics*. London: Butterworths.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type–token ratios for oral language of preschool children. *Journal of Speech and Hearing Research*, *29*, 129–134.

Higgs, T., & Clifford, R. (1982). The push towards communication. In T. V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57–79). Lincolnwood, IL: National Textbook Company.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development*, *9*, 67–84.

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

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*, 57–84.

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*, 1–15.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*, *12*, 28–41.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439–448.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*, 21–33.

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara. *Applied Linguistics*, *26*, 581–587.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund, Sweden: CWK Gleerup.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

MacWhinney, B. (2000). *The CHILDES project*: *Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave MacMillan.

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*, *2*, 141–154.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*, 323–337.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, *26*, 32–47.

Miller, J. F. (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research in child language disorders*: *A decade of progress* (pp. 211–220). Austin, TX: Pro-Ed.

Minnen, G., Carroll, J., & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*, *7*, 207–223.

O'Loughlin, K. (1995). Lexical density in candidate output of direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*, 217–237.

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

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language*, *14*, 201–209.

Steering Committee for Foreign Language Major Education in Institutions of Higher Education. (2000). *National syllabus for English majors*. Beijing: Foreign Language Teaching and Research Press.

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

Thordardottir, E., & Ellis Weismer, S. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language and Communication Disorders*, *36*, 221–244.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 252–259). Edmonton, Alberta, Canada: The Association for Computational Linguistics.

Ure, J. (1971). Lexical density: A computational technique and some findings. In M. Coultard (Ed.), *Talking about text* (pp. 27–48). Birmingham, England: English Language Research, University of Birmingham.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65–83.

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

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing*: *Measures of fluency, accuracy, and complexity* (Report No. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*,215–229.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*, 236–259.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency—variable sensitivity. *Studies in Second Language Acquisition*, *27*, 567–595.