Language Assessment Quarterly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Assessing the Validity of Lexical Diversity Indices Using Direct Judgements

Kristopher Kyle, Scott A. Crossley & Scott Jarvis

LANGUAGE ASSESSMENT QUARTERLY
2021, VOL. 18, NO. 2, 154–170
https://doi.org/10.1080/15434303.2020.1844205

Routledge
Taylor & Francis Group

Assessing the Validity of Lexical Diversity Indices Using Direct Judgements

Kristopher Kyle[a,b], Scott A. Crossley[c], and Scott Jarvis[d]

[a]Department of Linguistics, University of Oregon; [b]Department of English, Yonsei University; [c]Department of Applied Linguistics and ESL, Georgia State University; [d]Department of Linguistics, University of Utah

ABSTRACT

Indices of lexical diversity have been used to estimate the size of a writer's vocabulary and/or a writer's lexical proficiency for some time. One issue with many commonly used indices of lexical diversity (e.g., TTR and index) is that they vary as a function of text length. Accordingly, much research has been devoted to the development of indices that are text length independent. However, very little research has investigated the degree to which indices of lexical diversity are reflective of human ratings of diversity itself. In this study, the relationship between indices related to three dimensions of lexical diversity (abundance, variety, and volume) and human ratings of lexical diversity are explored in a corpus of L1 and L2 argumentative essays. The results indicated that abundance was the strongest predictor of lexical diversity ratings, followed by volume and variety. Multivariate models indicated that 74% of the variance in lexical diversity ratings could be explained by abundance and variety, and that speaker status had a negligible effect. Implications for research, writing instruction, and writers are discussed.

Introduction

Lexical measures have been used for over a hundred years as a way to estimate the size of a language user's vocabulary (Jarvis, 2002, 2013b; Thomson & Thompson, 1915; Yule, 1944). Much of this research has used indices that measure the diversity of lexical items produced in a particular text. One of the simplest measures of lexical diversity is the type token ratio (TTR), which is calculated as the number of unique words (types) in a text divided by the total number of words (tokens) in the text. It is widely known, however, that TTR values are intrinsically skewed by the length of a text, wherein longer texts tend to have lower TTR scores because the proportion of repeated words increases as the text grows longer (e.g., Koizumi & In'nami, 2012; Tweedie & Baayen, 1998). Accordingly, much research has been devoted to the development of indices of lexical diversity that do not vary as a function of text length (e.g., Covington & McFall, 2010; Malvern & Richards, 1997; McCarthy & Jarvis, 2007, 2010). In addition to testing the independence of these measures from text length, studies often use criteria related to proficiency (e.g., holistic writing or speaking quality scores, program level, etc.) to validate candidate indices (e.g., Engber, 1995; Jarvis, 2002; Treffers-Daller, Parslow, & Williams, 2018). What is rare, however, is for

CONTACT Kristopher Kyle kkyle2@uoregon.edu Kristopher Kyle, University of Oregon, Straub (John) Memorial Hall, 161, Eugene , OR 97405-1290, United States of America

measurements of diversity to be validated using human judgements of diversity itself (c.f., Jarvis, 2013b, 2017)

While much is known about text-length effects, much less is known about which indices (or set of indices) best measure the construct of lexical diversity itself. Knowing that lexical diversity is an important aspect of language assessment, it becomes particularly important to better understand which index (or indices) best measure lexical diversity and in what tasks. While a few studies have investigated the validity of text-length independent lexical diversity indices (e.g., MATTR, HD-D, MTLD) using judgments of lexical diversity in written tasks, none (to our knowledge) examine lexical diversity in argumentative writing tasks. Additionally, many lexical diversity studies focus on language samples produced by first language (L1) or second language (L2) speakers, but none focus on both groups. Importantly, calculating indices of lexical diversity has traditionally required a background in computer science (to develop automated measures of lexical diversity), limiting access to the construct and perhaps the number of lexical diversity studies. In this study we address these gaps by investigating the relationship between six indices of lexical diversity and human judgements of lexical diversity in argumentative essays written by both first language (L1) and second language (L2) users. We also introduce a freely available tool (the Tool for the Automatic Analysis of Lexical Diversity [TAALED]) to measure a variety of lexical diversity indices.

Indices of lexical diversity and text length

Indices of lexical diversity have been used to measure the variety of words used in a text, which is argued to be reflective of the size of a writer's or speaker's vocabulary. Given a particular language production task (e.g., an essay on a particular topic), more proficient language users, who presumably have a larger vocabulary, will use a wider variety of lexical items than less proficient users. A wide range of indices of lexical diversity have been proposed, but one of the simplest (and perhaps the most well-known) is the type-token ratio (TTR), which is calculated as the number of types (i.e., different words) in a text divided by the number of tokens (i.e., total number of words) in a text (Johnson, 1944). A well-known issue with indices such as TTR is their relationship with text length. In the case of TTR, longer texts tend to earn lower diversity scores than shorter texts (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007). This is problematic for at least two related reasons. First, it is presumed that in a particular timed task more proficient language users will be more fluent (resulting in longer texts) in addition to having a larger vocabulary (resulting in the use of more diverse vocabulary items). When using indices such as TTR to measure lexical diversity, more proficient language users may be penalized because they will write longer texts that will earn lower diversity scores.[1] Alternatively, when using indices with positive relationships with text length (e.g., Guiraud's index, see below) more proficient language users may earn inflated lexical diversity scores (e.g., Koizumi & In'nami, 2012). This latter case may be less problematic – there is some evidence that human perceptions of lexical diversity are positively correlated with text length (e.g., Jarvis, 2013b).

[1] Some transformations of TTR, such as Guiraud's (1960) index, over-correct for TTR's relationship with text length and therefore are likely to inflate diversity scores for more proficient language learners.

The second potential issue concerns measurement precision – ideally indices of lexical diversity should reflect diversity itself irrespective of other textual (or language user) characteristics. This is particularly important for diagnostic purposes wherein language learners and/or their teachers may want to determine the strengths and/or weakness of their language production skills on a particular task. One advantage of text length is that it tends to be strongly correlated with production quality scores (see, e.g., Chodorow & Burstein, 2004). However, text length is grossly imprecise with regard to diagnostic information – a wide range of (both positive and negative) textual features can potentially affect the length of a production. When indices of lexical diversity vary intrinsically due to text length, it cannot clearly be determined whether lexical diversity scores are a reflection of 'true' diversity or some other feature (or collection of features).

Due to these issues, a number of indices of lexical diversity have been proposed that attempt to disambiguate text length and lexical diversity using statistical transformations (Carroll, 1964; e.g., Guiraud, 1960) and more complex methods (e.g., voc-D [Malvern & Richards, 1997]; MTLD [McCarthy & Jarvis, 2010]). Unfortunately, many of these indices (including commonly used ones such as Guiraud's [1960] index) have been repeatedly shown to vary widely based on the length of a text (e.g., Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007). For instance, using a parallel sampling technique, Koizumi and In'nami found a large effect of text length for a number of common indices of lexical diversity in a small (n = 38) sample of short spoken L2 texts, with the strongest effects observed for TTR and Guiraud's (1960) index. McCarthy and Jarvis (2007) also used a parallel sampling technique to investigate the relationship between text length and lexical diversity indices in an L1 corpus of speaking and writing. They found that most indices (including TTR and Root TTR) were strongly correlated with the length of text segments investigated. Research has, however, provided some evidence that a small set of lexical diversity indices may be relatively independent of text length. These include Moving Average TTR (MATTR; Covington & McFall, 2010), Measure of Lexical Textual Diversity (MTLD; McCarthy & Jarvis, 2010), and D (Malvern & Richards, 1997; McCarthy & Jarvis, 2007).

Multiple dimensions of lexical diversity

While most studies have conceptualized lexical diversity as a unidimensional construct (using indices such as TTR, MATTR, etc.), Jarvis (2013a, 2017) argued that lexical diversity is a multi-dimensional phenomenon. Influenced by the multiple ways in which ecologists measure biodiversity, Jarvis (2013a, 2017) introduced seven properties of lexical diversity, including *volume* (number of tokens), *abundance* (number of types), *variety* (the relative proportion of unique words), *evenness* (the degree to which types have an equal number of repetitions), *disparity* (the semantic relatedness of words), *specialness* (the presence of specific words perceived as enhancing diversity), and *dispersion* (the size of intervals between repetitions of the same word). In the present study, we focus on the effects of the first three of these dimensions: volume, abundance, and variety. The first two – volume (tokens) and abundance (types) – have the most straightforward measures, and the third (variety) is the dimension whose measures have been subjected to the greatest degree of validation testing using measures of lexical diversity. The other four dimensions are likely also to contribute to the overall measurement of lexical diversity, but we have not included them in the empirical portion of this study for reasons of space and because appropriate measures for these dimensions are still under development (Jarvis, 2017).

Validating indices of lexical diversity

Although a great deal of lexical diversity research has investigated relationships between lexical diversity measures (primarily with indices of lexical variety) and text-length, relatively few studies have investigated the degree to which various indices align with judgements of proficiency and/or lexical diversity itself. Of these studies, most have used judgements of writing and/or speaking proficiency as the dependent variable (Engber, 1995; Jarvis, 2002; Treffers-Daller et al., 2018). Engber (1995), for example, examined the relationship between a number of lexical indices (including lexical variety) and holistic writing quality scores. Engber's study indicated a moderate positive relationship between lexical variety (both with and without lexical errors) and writing quality scores. In a more recent study, Treffers-Daller et al. (2018) investigated the relationship between three simple and three more complex indices of lexical diversity and proficiency scores based on the Common European Framework of Reference for Languages (CEFR) using essays written by L2 English users of varying proficiency levels. They found that the simple indices (e.g., number of types per essay) were better indicators of proficiency than the more complex ones (e.g., MTLD). The results most likely reflect the relationship between the simple measures and text length, which is a strong predictor of essay quality. However, the authors did find that the number of types per essay and the number of words per essay uniquely contributed to the variance in proficiency scores, indicating that these two dimensions of lexical diversity are not tautological.

Jarvis (2013b, 2017) directly investigated the relationship between indices of lexical diversity and human holistic judgements of lexical diversity using narrative retellings of a silent film (n = 50). Jarvis (2013b) found meaningful correlations ($|r| \geq .100$) between indices that represented six of the diversity features (disparity, dispersion, evenness, specialness, variety, and volume) identified in Jarvis (2013a) and holistic lexical diversity scores. A multiple regression analysis indicated that a large proportion of the variance in lexical diversity judgements (49%) could be explained by volume (number of words) and specialness (referred to as "rarity" and measured as the mean frequency rank for all words based on the British National Corpus). However, most of the variance explained by this model could be attributed to volume. When volume was excluded from the model, however, a similar amount of variance in lexical diversity judgments was still explained (47%) using the five remaining indices (disparity, dispersion, evenness, specialness, and variety). One of the limitations of this study is that each retelling was rated by only two or three raters, and no rater rated more than 20 essays. Interrater reliability between pairs of raters ranged from $r = -.04$ to $r = .65$, with a mean $r = .31$. Low levels of interrater reliability may have had deleterious effects on the validity of the regression model.

To address this problem, Jarvis (2017) focused on determining whether human judgements of lexical diversity found in learners' and native speakers' written narrative retellings of a silent film were reliable and replicable. He also explored the relationship between these judgments and indices of lexical variety. In Jarvis' study, three groups of raters – each having 20 or 21 raters – rated a corpus of 50 (one group) or 60 (two groups) retellings.

Levels of interrater reliability among the raters was found to be exceedingly high, with Cronbach's alphas of .90 (first group), .96 (second group), and .95 (third group). The study also found that the mean lexical diversity ratings of each group of raters were nearly identical. Jarvis found moderate to strong correlations between the human judges' mean ratings of lexical diversity and indices of variety that have been found in previous studies to be relatively independent of text length, including HD-D (r = .669), MTLD (r = .483), and MATTR (r = .577). The results show that under the right conditions, human judgments of lexical diversity can indeed be both highly reliable and replicable and can therefore serve as a gold standard against which other measures of lexical diversity can be assessed and validated. Jarvis pointed out that although these correlation values were relatively high, much higher correlations would be expected if lexical variety (as measured by these indices) were fully representative of the construct of lexical diversity indicating that a multidimensional model of lexical diversity may be needed (Jarvis (2013a, 2013b). To highlight this point, research has shown that a combination of at least five of the seven dimensions of lexical diversity Jarvis proposed accounts for as much as 89% of the variance in human judgements of lexical diversity (Jarvis, 2017).

Current study

The current study investigates the relationship between three dimensions of lexical diversity (volume, abundance, and variety) and human judgments of lexical diversity in argumentative essays written by both first language (L1) and second language (L2) users of English. With regard to variety, four indices that have been shown to be relatively independent of text length are used. Relationships between the six indices and human judgements of lexical diversity are first examined. Linear models are then used to determine which indices of volume, abundance, and variety are the most predictive of human scores (while controlling for L1 status). Accordingly, this study is guided by the following research questions:

1. What is the relationship between human judgements of lexical diversity and objective measures of lexical volume, abundance, and variety?
2. To what degree are the relationships between human judgements of lexical diversity and objective measures of lexical volume, abundance, and variety complementary?

Method

Corpora

In this study, two corpora that shared a register (argumentative writing) were used. The first comprises argumentative essays written by individuals for whom English was their first language (L1). The second comprises argumentative essays written by individuals for whom English was an additional or second language (L2). All essays were given holistic lexical diversity scores by trained raters using the same scale. See below for more information regarding the corpora and the ratings. Descriptive statistics can be found in Table 1

L1 corpus

The L1 corpus comprised 315 argumentative essays taken from Crossley and McNamara (2011). The essays were written by undergraduate freshmen composition students at

Table 1. Descriptive statistics for holistic lexical diversity scores.

Corpus	*n*	Min	Max	Mean	Standard Deviation	Median
L1	315	4.50	10.00	6.49	1.04	6.50
L2	300	3.00	10.00	6.11	1.01	6.50
Combined	615	3.00	10.00	6.31	1.04	6.50

Mississippi State University (MSU) on two SAT prompts. Students were randomly assigned to either prompt. One prompt was about originality and uniqueness, while the other was about admiring heroes versus celebrities. The students were given 25 minutes to write an essay during which no outside referencing was allowed. All essays were written on computers.

L2 corpus

The L2 corpus comprised 300 argumentative independent essays taken from the TOEFL public use data set. Participants wrote one independent argumentative essay on one of two prompts that did not overlap with the prompts in the L1 corpus. In each independent essay prompt, the test-takers were given two opposing viewpoints, and they had to support their positions with arguments in a timed 30-minute writing task. The two prompts dealt with selecting majors and cooperation. The selected essays were stratified by writing scores so that a range of scores were represented, and equal numbers of essays were selected from the two prompts (i.e., 150 essays from each prompt).

Lexical diversity ratings

All essays were scored for lexical diversity by two trained raters. The raters were instructed to rate each sample on a scale of 1–10 with 1 being a text with low lexical diversity and 10 being a text with high lexical diversity. Lexical diversity was defined as “the variety of word use that can be found in a person’s speech or writing.” Prior to scoring the essays, the raters were trained in the following manner: First, they were given three sample texts that had already been scored for lexical diversity, and they were asked to read these texts and attempt to calibrate themselves with the scores already assigned to them. The three sample texts were selected from the Chaplin corpus used in Jarvis (2017), which had been rated for lexical diversity by several dozen human raters and had shown exceptionally high levels of inter-rater reliability (Cronbach’s alpha = 0.99). The assessed levels of lexical diversity in these texts therefore appear to be trustworthy as well as rather precise. The first sample text chosen for the present study is the one with the lowest lexical diversity rating in the Chaplin corpus (2.89 on a scale of 1–10). The second sample text has a lexical diversity rating of 5.00 on the same scale, and the third text has a lexical diversity rating of 8.04, which is the highest lexical diversity score in the Chaplin corpus.

After being shown the sample texts and their associated lexical diversity scores, the raters were trained on 100 unrelated text samples and once acceptable inter-rater reliability (Kappa > .70) was reached, the raters scored the entire set of written essay samples independently. The initial Kappa value for the dataset was .667. After scoring, raters were given the opportunity to adjudicate any ratings that varied by more than 1 point (on a 10-point scale) between the two raters. At this point, the raters could revise their scores after talking through their initial ratings. The average scores across these adjudicated ratings were used as the final ratings of lexical diversity for the dataset. Kappa for the adjudicated ratings was .748.

Indices of lexical diversity

All indices of lexical diversity were calculated using the Tool for the Automatic Analysis of Lexical Diversity (TAALED),[2] which calculates a range of classic (e.g., TTR and Guiraud's index) and more robust indices of lexical variety (e.g., MATTR and MTLD), in addition to other diversity indices such as volume and abundance. TAALED is freely available and is open-source and is available as an easy to use graphical user interface (GUI) that works on both Mac and Windows operating systems. The underlying analysis code is also available as a Python package. While the GUI version of TAALED was designed for English texts, the Python package can be used to calculate lexical diversity indices for a wide variety of languages. For this study, we selected three dimensions of lexical diversity based on Jarvis (2013a, 2017) framework: volume, abundance, and variety. One index was selected to measure volume and one was selected to measure abundance. Four indices of variety that have been shown to be relatively independent of text length including MATTR (Covington & McFall, 2010); an instantiation of D (Malvern & Richards, 1997; McCarthy & Jarvis, 2007), and two versions of MTLD (McCarthy & Jarvis, 2010) were considered. Further details regarding the operationalization of each is included below.

Volume

Volume refers simply to the number of words in a text.

Abundance

Abundance refers to the total number of different types in a text. Note that types refer to lemmas. For instance, if the verbs *run* and *ran* occur in a text, they would be counted as instances of a single type *run*.

Variety

Four indices of lexical variety were selected based on their relative independence of text length.

HD-D

HD-D (McCarthy & Jarvis, 2007), which is a more reliable calculation of voc-D (Malvern & Richards, 1997) relies on the probability that a word in a text would be included in a random sample from that text. Probabilities are calculated using the hypergeometric distribution, and the HD-D scores used in this study comprise the combined probabilities for all words in a text (McCarthy & Jarvis, 2007 refer to this variant of HD-D as ATTR). Using this operationalization, higher HD-D scores are associated with greater lexical variety. McCarthy and Jarvis (2007) found a small relationship (r = .282) between text length and HD-D for longer texts (up to 2000 words). Koizumi and In'nami (2012), using a relatively small sample (n = 38) of spoken L2 texts, found that HD-D was less affected by text length

[2] The compiled tool is available at www.linguisticanalysistools.org. The open source Python code is available at https://github.com/kristopherkyle/TAALED.

than other commonly used indices, while still reporting meaningful and significant effects. Zenker and Kyle (2021), using a large corpus (n = 4,542) of written L2 argumentative texts (similar to those used in the current study) found a negligible relationship (r = .064) between HD-D and text length. Because HD-D has been demonstrated to have a smaller relationship with text length than many other commonly used indices and due to the negligible relationship with text length in Zenker and Kyle (2021), it was included in the current study.

MATTR

Moving average type-token ratio (MATTR; Covington & McFall, 2010) minimizes the relationship between type-token ratios and text length by averaging type-token ratios across multiple, overlapping, equally sized windows (sections) in the text. In this study, we use a 50-word moving window, which means that TTR values are calculated for words 1–50, 2–51, 3–52, etc., until all words in the text have been included in at least one window. The final MATTR score comprises the average of all TTR values. Higher MATTR scores are associated with greater lexical variety. Fergadiotis, Wright, and Green (2015) found that MATTR scores were independent of length in a variety of spoken L1 registers, and Zenker and Kyle (2021) found similar results for L2 argumentative essays.

MTLD (original)

The measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010) represents the average number of words it takes to reach a point of TTR stabilization (TTR = .720) while iterating through a text (which are referred to as factors). One issue with operationalizing MTLD is that there are usually words at the end of the text that have not yet reached a TTR of .720. McCarthy and Jarvis (2010) dealt with this issue by estimating the final factor length based on the parameters of the remaining words. To help achieve index stability, MTLD is run both forwards and backwards, and the final reported value represents the average of these two. Research has demonstrated that MTLD is independent of text length in a variety of L1 genres (Fergadiotis et al., 2015; McCarthy & Jarvis, 2010) and Zenker and Kyle (2021) in L2 argumentative essays.

MTLD-W

MLTD moving average is a variant of MTLD (see above; McCarthy & Jarvis, 2010) that uses a moving average (or moving window) approach. Factor lengths are calculated starting with the first word, then the second, then the third, etc., until the last word in the text has been reached. Instead of calculating the final factor based on the parameters of a partial factor (as in the original version), the final factor is calculated by wrapping around to the beginning of the text. For example, in a 200-word text, if the final factor does not reach a TTR of .720 by word 200, then the tool wraps around to the beginning of the text and continues adding words until a full factor has been reached. Research has indicated that MTLD-W is also independent from text length in L2 argumentative essays Zenker and Kyle (2021).

Statistical analysis

All statistical analyses were conducted in R 3.6.1 (R Core Team, 2016). To address RQ1, bivariate Pearson correlations were calculated using the cor.test() function (R Core Team,

2016) between each index of lexical diversity and the human lexical diversity ratings, both for the corpus as a whole and for each subcorpus. To address RQ2, linear models were used to predict human lexical diversity scores from lexical diversity indices using the lm() function (R Core Team, 2016). The relative importance of the indices in each model was calculated using the calc.relimp() function in the relaimpo package (Grömping, 2006). Specifically, the metric lmg (Lindeman, Merenda, & Gold, 1980), which takes into account both the direct relationship between the independent and dependent variable (i.e., the bivariate correlation) and the indirect relationship between the independent and dependent variable (i.e., the amount of variance explained when included in a multivariate model) was used.

Results

A preliminary visual analysis of distribution plots and scatter plots indicated that all indices of volume and variety were roughly normally distributed and demonstrated a reasonably linear relationship with the human ratings of lexical diversity.

RQ1

The results of a correlation analysis indicated that there were medium to large correlations between the indices of volume, abundance, variety and the human ratings of lexical diversity. Table 2 includes descriptive statistics, and Table 3 includes the results of the correlation analysis between diversity indices and the holistic LD scores.

The results indicate that all indices of lexical diversity considered demonstrate moderate to large correlations with holistic human judgements of lexical diversity. These relationships are roughly consistent across the L1 and L2 corpora, though stronger relationships are consistently found between LD indices and LD scores in the L2 corpus. The strongest relationship was found between Abundance and LD scores (r = .847 in the combined

Table 2. Descriptive statistics for all variables investigated.

	Combined		L1		L2	
Index	mean	SD	mean	SD	mean	SD
Holistic LD scores	6.30	1.04	6.49	1.04	6.11	1.01
Volume	330.75	100.48	348.28	115.97	312.35	77.12
Abundance	140.87	36.64	146.84	39.88	134.61	31.78
MATTR	0.75	0.04	0.75	0.04	0.74	0.04
HD-D	0.78	0.03	0.79	0.03	0.78	0.03
MTLD	58.14	13.53	59.51	13.90	56.70	13.00
MTLD-W	57.44	13.50	58.88	13.92	55.94	12.90

Table 3. Correlations between lexical diversity indices and human judgements of lexical diversity.

	Combined		L1		L2	
Index	*r*	*p*	*r*	*p*	*r*	*p*
Volume	0.687	<.001	0.683	<.001	0.695	<.001
Abundance	0.847	<.001	0.815	<.001	0.890	<.001
MATTR	0.492	<.001	0.402	<.001	0.566	<.001
HD-D	0.602	<.001	0.522	<.001	0.666	<.001
MTLD	0.505	<.001	0.438	<.001	0.566	<.001
MTLD-W	0.524	<.001	0.433	<.001	0.612	<.001

Table 4. Correlation matrix for all variables considered in the combined corpus.

	Holistic LD Score	Volume	Abundance	MATTR	HD-D	MTLD
Volume	0.687					
Abundance	0.847	0.884				
MATTR	0.492	0.163	0.493			
HD-D	0.602	0.295	0.619	0.879		
MTLD	0.505	0.189	0.525	0.901	0.837	
MTLD-W	0.524	0.202	0.543	0.918	0.863	0.954

corpus), followed by Volume (*r* = .687 in the combined corpus), HD-D (*r* = .602 in the combined corpus), and the other three indices of variety.

Given the conceptual and operational similarities between the indices investigated (and in particular, the indices of variety), a correlation matrix was constructed to determine the degree to which indices of lexical diversity were collinear. The correlation matrix, which is reported in Table 4, indicates that abundance is strongly correlated with volume (*r* = .884). As expected, all indices of variety are also strongly correlated (ranging from *r* = .837 to *r* = .954).

RQ2

A number of linear models were constructed to investigate the relationship between holistic lexical diversity scores and abundance, speaker status, and indices of lexical variety. We selected abundance and not volume because, while both were strongly multicollinear, abundance yielded a stronger correlation with human ratings of lexical diversity. We developed separate models for each index of lexical variety because of the strong multicollinearity found among the lexical variety indices. In addition, we opted to include a separate model for each index of variety to demonstrate changes in lexical diversity predictor models when different measures of lexical variety were included.

Abundance and MATTR

The two indices abundance and MATTR were used in a linear model to predict holistic LD scores. The model indicated a significant relationship with a large effect ($p < .001$, $R^2_{adjusted} = .736$), and explained approximately 74% of the variance in LD scores. The model parameters are summarized in Table 5. The relative importance metrics indicate that approximately 59% of the explained variance can be attributed to abundance, while approximately 12% of the explained variance can be attributed to MATTR. There was no

Table 5. Abundance + MATTR.

	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		1.788	0.646	2.766	0.006
Abundance	0.588	0.020	0.001	24.470	< 0.001
MATTR	0.121	2.287	0.912	2.508	0.012
SpStatusL2	0.017	−0.661	0.920	−0.718	0.473
Abundance:SpStatusL2	0.011	0.007	0.001	4.578	< 0.001
MATTR:SpStatusL2	0.001	−0.477	1.345	−0.355	0.723

main effect for speaker status (SpStatus), though there was a significant (but negligible) interaction between abundance and speaker status.

Abundance and HD-D

The two indices abundance and HD-D were used in a linear model to predict holistic LD scores. The model indicated a significant relationship with a large effect (p < .001, $R^2_{adjusted}$ = .737), and explained approximately 74% of the variance in LD scores. The model parameters are summarized in Table 6. The relative importance metrics indicate that approximately 53% of the explained variance can be attributed to abundance, while approximately 18% of the explained variance can be attributed to HD-D. There was no main effect for speaker status (SpStatus), though there was a significant (but negligible) interaction between abundance and speaker status.

Abundance and MTLD (original)

The two indices abundance and MTLD (original) were used in a linear model to predict holistic LD scores. The model indicated a significant relationship with a large effect (p < .001, $R^2_{adjusted}$ = .735), and explained approximately 74% of the variance in LD scores. The model parameters are summarized in Table 7. The relative importance metrics indicate that approximately 58% of the explained variance can be attributed to abundance, while approximately 13% of the explained variance can be attributed to MTLD (original). There was a significant (but negligible) main effect for speaker status (SpStatus), and a significant (but negligible) interaction between abundance and speaker status.

Abundance + MTLD-W

The two indices abundance and MTLD-W were used in a linear model to predict holistic LD scores. The model indicated a significant relationship with a large effect (p < .001, $R^2_{adjusted}$ = .735), and explained approximately 74% of the variance in LD scores. The model parameters are summarized in Table 8. The relative importance metrics indicate that

Table 6. Abundance + HDD.

	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		1.004	0.867	1.157	0.248
Abundance	0.532	0.020	0.001	21.719	< 0.001
HDD	0.181	3.286	1.187	2.768	0.006
SpStatusL2	0.016	−0.566	1.215	−0.466	0.641
Abundance:SpStatusL2	0.009	0.006	0.002	3.958	< 0.001
HDD:SpStatusL2	0.001	−0.535	1.716	−0.312	0.755

Table 7. Abundance + MTLD.

	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		3.172	0.145	21.864	< 0.001
Abundance	0.580	0.020	0.001	23.668	< 0.001
MTLD	0.127	0.006	0.002	2.402	0.017
SpStatusL2	0.017	−0.964	0.210	−4.584	< 0.001
Abundance:SpStatusL2	0.011	0.007	0.001	4.769	< 0.001
MTLD:SpStatusL2	0.002	−0.002	0.004	−0.509	0.611

Table 8. *Abundance* + MTLD-W.

	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		3.208	0.144	22.315	< 0.001
Abundance	0.571	0.020	0.001	23.751	< 0.001
MTLD (wrap)	0.136	0.005	0.002	2.040	0.042
SpStatusL2	0.017	−1.034	0.208	−4.976	< 0.001
Abundance:SpStatusL2	0.010	0.006	0.002	4.178	< 0.001
MTLD (wrap):SpStatusL2	0.003	0.001	0.004	0.250	0.803

approximately 57% of the explained variance can be attributed to abundance, while approximately 14% of the explained variance can be attributed to MTLD-W. There was a significant (but negligible) main effect for speaker status (SpStatus), and a significant (but negligible) interaction between abundance and speaker status.

Discussion and conclusion

This study investigated the relationship between human judgements of lexical diversity and indices related to three dimensions of lexical diversity in argumentative essays written by both L1 and L2 speakers using TAALED. The results of the study, which are discussed below, have important implications for understanding and measuring the construct of lexical diversity, and also provide directions for future research.

RQ1

Research question one investigated the individual relationships between human judgments of lexical diversity and various indices of lexical diversity. The results indicated that Abundance (number of types) exhibited the strongest relationship with human judgements of diversity in both the L1 and the L2 corpora, and when the datasets were combined ($r > .800$). Volume (number of tokens) was the next strongest index in all three datasets ($r > .683$), but was strongly collinear with abundance ($r = .884$). The four indices of variety exhibited medium to strong correlations ($r = .402$ to $.666$) with holistic lexical diversity scores across the L1, L2 and combined corpora. The variety indices demonstrated consistently stronger effects in the L2 corpus than in the L1 corpus (on average, r values were approximately .125 higher in the L2 corpus). This is likely due to the fact that there was a wider range of lexical diversity scores in the L2 corpus than the L1 corpus. The indices of variety were highly collinear ($r = .837$ to $.954$), which suggests that they all measure a very similar aspect of lexical diversity. These indices exhibited medium to large correlations ($r = .493$ to $.619$) with abundance but were weakly correlated with volume ($r = .163$ to $.295$). These results generally align with previous related studies that have investigated the relationship between objective measures of lexical diversity and holistic judgements of lexical diversity in L2 narrative retelling tasks (Jarvis, 2013b, 2017). This suggests that similar lexical features affect judgements of lexical diversity across argumentative essay and narrative retelling tasks.

Overall, the results indicate that, of the indices investigated, abundance (the number of types in an essay) demonstrates the strongest relationship with human judgments of lexical diversity. This suggests that human judgments of diversity are affected by the total number of types in addition to the proportion of types to tokens. The introduction of new types occurs via at least two possible ways: a) through the introduction of new ideas to the text and b) through paraphrasing or summarizing ideas that have already been introduced. Indices of abundance and variety are sensitive to both methods of type introduction. However, while indices of variety should not be affected by the total number of ideas in a text, abundance will be. In other words, abundance appears to highlight a rater's dependence on the total number of ideas in a text when assigning a score. It would be easy to attribute this relationship to the length of an essay (i.e., longer essays tend to have more arguments, more support for arguments, etc.). However, the stronger correlations between judgements of lexical diversity and abundance as compared to volume suggests that raters are attuned to more than just length – they appear to be affected by differences in both idea generation and paraphrasing/summarizing.

RQ2

The second research question was concerned with predicting human judgements of lexical diversity using objective indices related to abundance, volume, and variety. Because abundance and volume were strongly collinear (r = .847) and because abundance was more strongly correlated with human judgements, volume was excluded from the predictor models. Four predictor models were constructed using abundance and each of the indices of variety respectively. The results indicated that all four models explained approximately 74% of the variance in human judgements of lexical diversity (see Table 9 for a summary of these results).

In each model, speaker status had only a negligible effect on the prediction, suggesting that the trends are stable across first and second language writers. The relative importance metrics indicate that abundance contributed the most explanatory power to each model, but that the indices of variety also contributed explanatory power. Again, these results generally align with previous studies involving narrative retelling tasks (Jarvis, 2013b, 2017). Jarvis' (2013b) model, for example, explained 49% of the variance in human judgements of lexical diversity, with most of the variance being explained by volume. While abundance was not used in Jarvis (2013b), our results indicate that volume and abundance are strongly correlated. However, a second multivariate model that excluded volume also explained a similar amount of the variance. Relatedly, Jarvis (2017) reported that when a large number (15+) raters are used, up to 89% of the variance in human judgements of lexical diversity can be explained using similar diversity indices (abundance, variety, evenness, dispersion, & specialness). Taken together, it is clear that powerful explanatory models of human

Table 9. Summary of regression models.

Model	$R^2_{adjusted}$	Relative importance of abundance	Relative importance of variety index	Relative importance of other features
Abundance + MATTR	0.736	0.588	0.121	0.029
Abundance + HD-D	0.737	0.532	0.181	0.026
Abundance + MLTD (original)	0.735	0.580	0.127	0.030
Abundance + MTLD-W	0.735	0.571	0.136	0.030

judgements of lexical diversity can be constructed using fairly straightforward objective measures.

Implications for research

Overall, the results indicated that the best objective indicator of human judgments of lexical diversity is abundance (the total number of types in an essay). This finding is important for understanding the construct of lexical diversity – namely that human perceptions of lexical diversity may be affected by the total number of different ideas in a text. This suggests that the most effective method of measuring lexical diversity (in isolation) would be to use the total number of types in a language sample, and possibly dispense with the more complex indices of diversity that are independent of text length. However, researchers most often use an index of lexical diversity as one of many variables (along with indices of lexical sophistication, syntactic complexity, etc.) to measure larger, more complex constructs such as writing quality. In this case, the correlation between abundance and volume (i.e., total number of words) becomes potentially problematic – the number of words in an essay is known to indiscriminately account for a wide range of potential features (including writer fluency, degree to which ideas are developed and/or arguments supported, etc.). For this reason, many automatic scoring models have attempted to disambiguate the various correlates of text length (e.g., Chodorow & Burstein, 2004). From this perspective, measures of variety that are more length independent (such as HD-D, MATTR, and MTLD) still hold an important place in the measurement of lexical diversity.

Furthermore, it may be the case that abundance, much like volume, is capturing a wide range of features that contribute to perceptions of lexical diversity but does so rather indiscriminately. If this is the case, then it may be important to determine which components affect abundance (and human judgments of lexical diversity themselves) and how they should be measured. As Jarvis (2013a, 2013b, 2017) argues, the construct of lexical diversity consists of a number of components. Included in this paper are only three of these components (volume, abundance, and variety), while at least four other components (disparity, dispersion, evenness, and specialness) are not included, primarily because they are still under development and are in need of validation. In order for lexical diversity to be clearly (and more fully) indexed when measuring more complex constructs (such as lexical diversity), it is important to find appropriate measures of the other constructs (and in particular, disparity).

Implications for writers, test takers, and assessment design

While L2 writers are often encouraged to use a diverse array of lexical items in writing classrooms, it should be noted that paraphrasing ideas that have already been addressed is only one way to increase perceptions of lexical diversity (e.g., Crossley, Muldner, & McNamara, 2016). Given the relationship between holistic lexical diversity scores and both abundance and volume, it would appear that judgements of lexical diversity are also affected by the total number of different ideas in a text. Adequately displaying lexical diversity therefore likely involves both demonstrating one's ability to use a variety of words to discuss a particular issue/argument and one's ability to discuss a variety of issues and arguments. For assessment design, task type may affect a test-taker's diversity of lexical items. If tasks are intended to measure lexical diversity, it may be important for designers to ensure that the tasks lend themselves to elaboration and/or idea generation.

Limitations and future directions

This study suggests that abundance is the best predictor of holistic judgements of lexical diversity in argumentative essays regardless of L1/L2 status. This study also suggests that text length independent indices of variety are also reasonable predictors of holistic judgements of lexical diversity, and may be preferred when used with other indices (e.g., of lexical and phrasal sophistication) to measure more complex constructs (such as holistic writing quality scores). Nonetheless, these findings should be tempered through a number of limitations that should be addressed in future studies. First, although these results align with similar findings with narrative retelling writing tasks (Jarvis, 2013b, 2017), they cannot necessarily be extended to other writing tasks (e.g., descriptive and letter writing tasks). Future research should continue to expand this research with other writing tasks and in a variety of oral tasks. Second, in this study we have made a tentative connection between lexical diversity scores and idea generation based on at least one previous study (Crossley et al., 2016). While Crossley et al. demonstrated this connection in L1 argumentative writing samples, future qualitative and/or quantitative research (e.g., using latent semantic analysis, latent Dirichlet allocation, or related methods) should more directly investigate this presumed relationship. Other potentially informative approaches would be to conduct follow-up interviews with raters and/or record and analyze the rater adjudication sessions to identify textual features that were affecting raters' judgements. Third, although this study followed a common practice of employing two to three raters when assigning holistic scores to essays (e.g., Enright & Quinlan, 2010), and although high agreement was achieved after rater adjudication, previous research by Jarvis (2013b, 2017) has used a large number of raters (> 15) to obtain highly reliable ratings of lexical diversity. It is possible, therefore, that the ratings of lexical diversity in this study reflect the biases of the raters themselves and do not reflect the "true" diversity scores. Future research should compare the effect of training + adjudication (as was done in this study) and minimal training + larger rater samples (as was done by Jarvis, 2013b, 2017). It may also be reasonable to employ a post-rating session to elicit information from the raters about the salient text features that informed their ratings. Alternatively (or in addition), it may be helpful to record the adjudication sessions to further capture insight into how text features influence ratings of lexical diversity.

Finally, this study explored the relationship between holistic lexical diversity scores and three components of lexical diversity that have been clearly established in the field (Abundance, Variety, and Volume). However, Jarvis (2013a, 2013b, 2017) has outlined at least four other components of lexical diversity that need to be more clearly operationalized and tested. This may be particularly important when using lexical diversity as a predictor of writing proficiency and/or production quality in concert with other indices and for automated and semi-automated feedback systems where Abundance may be indicative of various constructs beyond lexical diversity (e.g., fluency).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Kristopher Kyle http://orcid.org/0000-0001-5415-9672
Scott A. Crossley http://orcid.org/0000-0002-5148-0273
Scott Jarvis http://orcid.org/0000-0001-5191-5487

References

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, *2*(1), 80.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, *2004*(1), i–38. doi:10.1002/j.2333-8504.2004.tb01931.x

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. doi:10.1080/09296171003643098

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2–3), 170–191. doi:10.1504/IJCEELL.2011.040197

Crossley, S. A., Muldner, K., & McNamara, D. S. (2016). Idea generation in student writing: Computational assessments and links to successful writing. *Written Communication*, *33*(3), 328–354. doi:10.1177/0741088316650178

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. doi:10.1016/1060-3743(95)90004-7

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334. doi:10.1177/0265532210363144

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, *58*(3), 840–852. doi:10.1044/2015_JSLHR-L-14-0280

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27. doi:10.18637/jss.v017.i01

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: Reidel.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. doi:10.1191/0265532202lt220oa

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106. doi:10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–34). John Benjamins. https://doi.org/10.1075/sibil.47.03ch1

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. doi:10.1177/0265532217710632

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*(2), 1–15. doi:10.1037/h0093508

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. doi:10.1016/j.system.2012.10.012

Lindeman, R. H., Merenda, P., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis* (pp. 119). Glenview, IL. Scott: Foresman and Company.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58–71.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. doi:10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. doi:10.3758/BRM.42.2.381

R Core Team. (2016). *A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Thomson, G. H., & Thompson, J. R. (1915). Outlines of a method for the quantitative analysis of writing vocabularies. *British Journal of Psychology*, *8*(1), 52.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352. doi:10.1023/A:1001749303137

Yule, C. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. Assessing Writing, 47, 100505. 2021.100505 47 doi:10.1016/j.asw.2020.100505