Assessing Writing 47 (2021) 100505

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Investigating minimum text lengths for lexical diversity indices

Fred Zenker [a,*], Kristopher Kyle [b,c]

[a] Department of Second Language Studies, University of Hawai'i at Mānoa, United States
[b] Department of Linguistics, University of Oregon, United States
[c] Department of English Language and Literature, Yonsei University, Korea

ARTICLE INFO

Keywords:
Lexical diversity
Text length
SLA
Learner corpus research

ABSTRACT

Lexical diversity (LD) is an important feature of a second language (L2) writer's lexical knowledge, and indices of LD have been widely used in the field of writing assessment (e.g., Cumming et al., 2006; Engber, 1995). Research with longer native speaker (L1) texts has indicated, however, that many commonly used LD indices are sensitive to text length and may conflate lexical breadth and fluency. Because of the importance of measuring LD in L2 writing assessment research, it is essential to know the degree to which particular LD indices are resistant to text length effects and the minimum text lengths at which these indices produce stable values. In this study, we investigate text length effects for nine indices of LD in a corpus of 4542 L2 argumentative essays. The results indicate that MATTR (Covington & McFall, 2010) and two versions of MTLD (McCarthy, 2005; McCarthy & Jarvis, 2010) are the most stable of the indices included in the study. MATTR performs particularly well, maintaining a high degree of stability across all text lengths. Comparisons based on essay prompt and proficiency level are also discussed.

1. Introduction

Indices of lexical diversity (LD) are often used as a measure of lexical richness/complexity in second language (L2) writing assessment research (e.g., Cumming et al., 2006; Engber, 1995; Linnarud, 1986; Treffers-Daller, 2013). The simplest and most widely used LD index is the type-token ratio (TTR; Johnson, 1944), which is simply the number of unique words (types) divided by the number of running words (tokens). An important underlying assumption is that an LD score provides an indication of an L2 writer's overall lexical knowledge—specifically, that more advanced learners use a wider range of lexical items. Given the importance placed on indices of LD and their widespread use in writing assessment, it is critical for the indices to provide reliable scores.

It is well known that traditional LD indices such as simple TTR are sensitive to text length, with longer texts tending to receive lower LD scores (e.g., McCarthy & Jarvis, 2007). The fact that such indices conflate vocabulary breadth and text length is a matter of serious concern in the context of L2 writing assessment because the text lengths in any set of L2 written responses tend to vary widely, even when the writings are timed. In an attempt to address this issue, a number of revised LD indices have been proposed over the years, such as vocd-D (Malvern & Richards, 1997; Malvern, Richards, Chipere, & Durán, 2004), Maas (Maas, 1972), and the measure of textual lexical diversity (MTLD; McCarthy, 2005; McCarthy & Jarvis, 2010). While research has generally indicated that these measures are much more stable than TTR in longer L1 texts (McCarthy & Jarvis, 2007, McCarthy & Jarvis, 2010), it is less clear how stable they are in shorter texts such as those commonly produced in L2 writing assessment tasks (though see Jarvis, 2002; Koizumi & In'nami,

* Corresponding author.
E-mail addresses: fzenker@hawaii.edu (F. Zenker), kkyle2@uoregon.edu (K. Kyle).

https://doi.org/10.1016/j.asw.2020.100505
Received 2 April 2019; Received in revised form 18 September 2020; Accepted 22 September 2020
Available online 3 November 2020

2012).

The present study aims to expand this body of research by investigating the stability of LD measures with texts of different lengths with the goal of establishing guidelines for the minimum text lengths that should be used with each LD index. Nine indices were selected and used to generate LD values for texts ranging in length from 50 to 200 tokens. These indices were chosen because they included a mixture of traditional indices that are known to be susceptible to text length effects (but have been widely used) and newer ones that are thought to be more resistant to such effects.

2. Literature review

2.1. Evaluation of text length effects for traditional LD indices

LD indices have been used in a wide variety of studies related to L2 writing development and assessment (e.g., Crossley, Salsbury, McNamara, & Jarvis, 2011; Johansson, 1999; Malvern et al., 2004; Malvern & Richards, 1997; Wolfe-Quintero, Inagaki, & Kim, 1998). Researchers have long been aware that some LD indices are susceptible to text length effects. Root TTR (Guiraud, 1960) and bilogarithmic TTR (Log TTR; Chotlos, 1944; Herdan, 1960) were two early attempts to correct for TTR's sensitivity to text length using simple mathematical transformations. However, more recent developments in computer technology have allowed for the development and use of more sophisticated LD indices and facilitated more thorough research into the text length effects for different indices.

One notable early study of the effect of text length on LD values was carried out by Hess, Sefton, and Landry (1986), who analyzed 50-utterance oral language samples from 83 preschool children using five LD measures that were popular at the time: simple TTR, corrected TTR (Carroll, 1964), Root TTR, Log TTR, and Characteristic *K* (Yule, 1944). Importantly, Hess et al. observed that running simple correlations between text length and LD score using samples of different lengths that came from different children would be undesirable because it would introduce confounds into the analysis. For example, children may tend to use fewer different kinds of words in short utterances than in longer ones. To address these concerns, Hess et al. used the parallel sampling method to divide each language sample into texts of different lengths. First, each sample was clipped to the first 200 tokens. Then the resulting texts were subdivided into four texts of 50 tokens, two texts of 100 tokens, one text of 150 tokens, and one text of 200 tokens. LD scores were then calculated for each text, and values from texts of the same length were averaged. Subsequent analysis using repeated measures ANOVAs showed that all the LD measures were significantly affected by text length (e.g., $\eta^2 = .82$ for simple TTR). Hess et al. concluded that these LD measures were not suitable for making comparisons between texts of different lengths.

A follow-up study by Hess, Haug, and Landry (1989) used oral language samples from 52 elementary school children to analyze four versions of TTR: simple TTR, corrected TTR, Root TTR, and Log TTR. Each sample was divided into texts of 50–600 tokens using the parallel sampling method and then LD scores were calculated for each of the resulting texts. As was the case in the earlier study by Hess et al. (1986), results suggested that none of the TTR measures were stable across texts of different lengths.

These two studies made an important contribution to the field in that they helped to expose the susceptibility of traditional LD measures to text length effects. They were also influential in that they set a precedent for using the parallel sampling method to segment language samples into texts of different lengths for analysis.

2.2. Evaluation of text length effects for revised LD indices

Due to the sensitivity of traditional LD measures to text length effects, the creators of revised LD measures have attempted to show that their indices are less susceptible to these effects. For example, Malvern et al. (2004) evaluated the LD measure that they had developed, vocd-D, using oral language samples from 38 children aged 27–33 months in the New England Corpus (Dale, Bates, Reznick, & Morisset, 1989; Snow, 1989). Separate vocd-D values were calculated for even-numbered and odd-numbered words in each text, and then these values were compared with the vocd-D scores for the whole text using repeated measures ANOVAs. Results showed that the mean vocd-D scores for even-numbered and odd-numbered words were not significantly different from the vocd-D scores for the whole text. However, the study had several important limitations. First, the length of the texts was not reported. Second, Malvern et al. only made comparisons between vocd-D scores for half of the text and the whole text, which gives no indication of the range of text lengths that might be appropriate for use with the index. Also, the child data used in their study may not be generalizable to other populations and modes, such as written texts produced by adult L2 learners.

Observing that vocd-D was rapidly becoming the preferred LD measure in the field, McCarthy & Jarvis (2007) conducted a more thorough assessment of its sensitivity to text length using a corpus of 23 genres taken from the Lancaster-Oslo/Bergen Corpus (Johansson, Leech, & Goodluck, 1978), the Brown Corpus (Kucera & Francis, 1967), the London-Lund Corpus (Svartvik & Quirk, 1980), the Wellington Corpus of Spoken New Zealand English (Holmes, Vine, & Johnson, 1998), the Glencoe Science Corpus (Biggs et al., 2003), and the Michigan Corpus of Academic Spoken English (Swales & Malczewski, 2001). Nine representative texts were selected from each genre and divided into sections of 11 different lengths ranging from 100 to 2000 tokens using the parallel sampling method. Then vocd-D values were calculated for each of the sections and values from sections of the same length were averaged. Results of a Pearson correlation ($r = .22, p < .01$) indicated that there was a significant relationship between the vocd-D scores and text lengths, although the vocd-D scores fared better than traditional LD measures such as simple TTR ($r = -.77, p < .01$). McCarthy and Jarvis concluded that vocd-D scores were not as independent of text length as its creators had claimed.

McCarthy & Jarvis (2010) followed a similar procedure to test their own LD measure, MTLD, for text length effects. Drawing on 16 genres from the same corpus they had used in their earlier study, they selected nine texts from each genre and divided them into sections of 11 lengths ranging from 100 to 2000 tokens via parallel sampling. MTLD values were then generated for each text, and

values for texts of the same length were averaged. Analysis with a Pearson correlation did not find a significant relationship between MTLD values and text lengths ($r = -.016, p = .530$). However, when they repeated the procedure using other well-known LD measures, McCarthy and Jarvis found each of them to be subject to text length effects. For example, the values for TTR ($r = .811$, $p < .001$), vocd-D ($r = .190, p < .001$), and Maas ($r = .125, p < .001$) were all significantly correlated with text length, though the strength of the relationship varied widely from one index to another. McCarthy and Jarvis concluded that MTLD values were more independent of text length than other LD measures that had been proposed previously.

The fact that the two studies by McCarthy and Jarvis (2007, 2010) used a corpus with texts sampled from a wide range of genres marks a significant improvement from earlier research on text length effects for LD measures, which had tended to use a small number of texts representing a single genre (e.g., Hess et al., 1986, 1989). However, one limitation of these studies was that they did not attempt to investigate the minimum text lengths needed to produce stable LD values. Also, they did not analyze L2 texts or text types that are commonly produced for L2 writing assessment tasks. The fact that they did not find a significant correlation between MTLD values and text length across sections ranging from 100 to 2000 tokens suggests that MTLD provides stable values with texts at least as short as 100 tokens. However, the data give no indication of whether MTLD values would remain stable in texts of even shorter lengths.

2.3. Previous research on minimum text lengths for LD indices

One of the few studies to investigate the minimum text lengths needed to produce stable LD values was carried out by Koizumi (2012), who evaluated simple TTR, Root TTR, vocd-D, and MTLD using spoken English samples from 20 Japanese adolescents. Each text was clipped to the first 200 words and then further subdivided into 25 segments ranging in length from 50 to 200 tokens by parallel sampling. LD values were calculated for each section and then averaged across sections of the same length. Results from repeated measures ANOVAs performed on the LD scores for five different text length ranges (50–100, 100–150, 150–200, 100–200, and 50–200) indicated that MTLD values stabilized at roughly 100 tokens. None of the other indices produced stable values within the text length ranges included in the study. Another study by Koizumi and In'nami (2012) used similar methods and obtained the same pattern of results. Simple TTR, Root TTR, Maas, vocd-D, HD-D, and MTLD were evaluated using spoken English samples from 38 Japanese teenagers. The procedure for segmenting and analyzing the texts was the same as that used by Koizumi (2012). Once again, MTLD was the only index that produced stable values, with stabilization occurring at around 100 tokens.

These two studies by Koizumi (2012) and Koizumi and In'nami (2012) are significant in that they represent early attempts to systematically test the minimum text lengths needed to produce stable LD values. However, these studies also have several important limitations. First, the sample sizes were exceedingly small considering the potential of learner corpus research to analyze large numbers of texts quickly and efficiently. Second, the participants in each study represented a subset of a single learner population (namely, Japanese EFL learners). Third, only a single task/prompt was used in each study, which raises concerns about the generalizability of the findings to other tasks/prompts. Additionally, it is not clear whether these results, which were obtained using L2 speaking samples, can be generalized to L2 written samples.

2.4. Current study

While researchers have been interested in text length effects for LD measures for several decades now, there remains a pressing need for studies investigating the minimum text lengths needed to produce stable LD values, particularly in the case of L2 research, where it is often desirable and/or necessary to analyze short texts (Kyle, 2020). Previous L2 research has investigated corpora that are relatively small, represent only a single learner group, do not control for task/prompt, and have not investigated argumentative writing. The present study builds on previous research by using a large learner corpus sampled from a wide range of learner groups across two argumentative writing prompts, with the goal of establishing robust guidelines for the minimum text lengths to be used with different LD indices. Accordingly, this study is guided by the following research questions:

1 How long does an L2 argumentative essay need to be to provide stable LD values for each index?
2 To what degree do different essay prompts influence LD values for each index?
3 To what degree does L2 proficiency affect LD values for each index?

3. Method

3.1. Learner corpus

The corpus used in this study consists of 4542 written essays from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013), a freely-available database of L2 English samples from university students in 10 countries/regions in Asia (China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Taiwan, and Thailand).

Each of the essays was written in response to one of two statements: (a) 'It is important for college students to have a part-time job' ($n = 2214$) or (b) 'Smoking should be completely banned at all the restaurants in the country' ($n = 2328$). Henceforth, these shall be referred to as the 'Job' and 'Smoke' prompts. The number of texts for each country/region and prompt can be found in Table 1.

The L2 English learners who wrote the essays had been sorted into proficiency groups (A2, B1, B2) based on the Common European Framework of Reference for Languages (Council of Europe, 2001). Sorting was done using scores from an L2 vocabulary size test (Nation & Beglar, 2007) and a standardized English proficiency test (TOEFL or TOEIC). The number of learners from each

Table 1
Number of texts by country/region and prompt.

Country/Region	Job	Smoke	Total
China	355	365	720
Hong Kong	77	76	153
Indonesia	187	186	373
Japan	310	346	656
Korea	237	245	482
Pakistan	183	171	354
Philippines	189	188	377
Singapore	185	191	376
Taiwan	168	190	358
Thailand	323	370	693
Total	2214	2328	4542

country/region that were sorted into each proficiency level can be found in Table 2. Note that the numbers of learners at each proficiency level are rather unbalanced. In particular, there are far more learners at the B1 level than at either of the other levels. The majority of the learners (80.58%) responded to both essay prompts.

3.2. Lexical diversity indices

Nine lexical diversity indices were used in the present study, ranging from traditional indices such as simple TTR to newer ones such as MTLD. The procedure for calculating each index is described below (for more detailed descriptions, see Kyle, 2020).

3.2.1. Type-token ratio (TTR)

The simple type-token ratio (TTR; Johnson, 1944) is calculated as the number of unique words in a text (types) divided by the number of running words (tokens): $TTR = \frac{ntypes}{ntokens}$.

3.2.2. Root TTR

One early attempt to make LD indices more resistant to text length effects was Root TTR (Guiraud, 1960), also called Guiraud's index. Root TTR is calculated as the number of types divided by the square root of the number of tokens: $Root\ TTR = \frac{ntypes}{\sqrt{ntokens}}$.

3.2.3. Log TTR

Another simple transformation of TTR is bilogarithmic TTR (Log TTR; Chotlos, 1944; Herdan, 1960), also known as Herdan's C. Log TTR is calculated by dividing the logarithm of the number of word types by the logarithm of the number of word tokens: $Log\ TTR = \frac{Log(ntypes)}{Log(ntokens)}$.

3.2.4. Maas' index

Maas' index (Maas, 1972), often referred to simply as 'Maas,' is a more complex transformation of TTR that attempts to fit the value to a logarithmic curve. Maas is calculated as follows: $Maas = \frac{Log(ntokens) - Log(ntypes)}{Log(ntokens)^2}$. Unlike most other LD indices, lower Maas values are associated with higher diversity.

3.2.5. MATTR

Moving-average TTR (MATTR; Covington & McFall, 2010) takes another innovative approach to minimizing text length effects, in this case by calculating the moving average for all segments of a given length. For a segment length of 50 tokens, TTR is calculated on tokens 1–50, 2–51, 3–52, etc., and the resulting TTR measurements are averaged to produce the final MATTR value. A segment length

Table 2
Number of learners by country/region and proficiency level.

Country/Region	A2	B1	B2	Total
China	49	330	13	392
Hong Kong	1	74	15	90
Indonesia	32	163	3	198
Japan	145	215	17	377
Korea	69	138	72	279
Pakistan	17	175	3	195
Philippines	2	186	11	199
Singapore	0	132	66	198
Thailand	28	144	23	195
Taiwan	115	273	2	390
Total	458	1830	225	2513

of 50 tokens was used for all the moving-average calculations in the current study.

3.2.6. HD-D

The hypergeometric distribution diversity index (HDD; McCarthy & Jarvis, 2007) was developed as another (and arguably improved) method of calculating Malvern and Richards's (1997) index, vocd-D. For each word type in a text, HD-D uses the hypergeometric distribution to calculate the probability of encountering one of its tokens in a random sample of 42 tokens. These probabilities are then added together to produce the final HD-D value for the text. For ease of interpretation, we convert this to the same scale as TTR.

3.2.7. MTLD

The measure of textual lexical diversity (MTLD; McCarthy, 2005; McCarthy & Jarvis, 2010) is an estimation of the point at which TTR stabilizes in longer L1 texts. As such, MTLD is based on the average number of tokens it takes to reach a given TTR value (e.g., .720). First, MTLD moves through a text and calculates TTR on segments of increasing length (one token, two tokens, three tokens, etc.) until reaching a segment of at least 10 tokens that meets the predetermined TTR value. Then it proceeds to the next word in the text and starts the process over again. What results is a list of text segments of various lengths, which McCarthy and Jarvis refer to as 'factors.' In the simplest version of MTLD, the remaining words at the end of the text are counted as a partial factor. The final MTLD value is the average of all the factor lengths.

3.2.8. MTLD-MA-BI

Moving-average bidirectional MTLD (MTLD-MA-BI; McCarthy & Jarvis, 2010) is a revised MTLD procedure that takes a moving-average approach to calculating the index. The number of words required to create a factor (i.e., reach a TTR of .720) is calculated for each progressive word in the text until a factor cannot be completed. This procedure is then repeated in reverse, starting with the last word in the text. The final reported MTLD value comprises the average factor length for all factors.

3.2.9. MTLD-MA-Wrap

Moving-average wrapped MTLD (MTLD-MA-Wrap; McCarthy & Jarvis, 2010) is another revised method of calculating MTLD. Like MTLD-MA-BI, it takes a moving-average approach to calculating the index. However, instead of working through the text in both directions, MTLD-MA-Wrap avoids partial factors by looping back around to the beginning of the text.

3.3. Data analysis

The raw data file downloaded from the ICNALE website consisted of 5603 essays of various lengths written by a mixture of native and non-native English speakers. The first step in preparing the data for analysis was to remove the native speaker essays. This resulted in the exclusion of 400 essays, or 7% of the overall number of texts in the corpus. Next, all essays shorter than 200 tokens were also removed from the data file. This resulted in the exclusion of a further 661 essays, or 13% of the remaining texts in the corpus. These two rounds of exclusions left us with 4542 L2 essays for subsequent analysis.

The next step in preparing the data was to split the essays into texts of different lengths. First, each essay was clipped to the first 200 tokens. Then the parallel sampling method (Hess et al., 1986) was used to subdivide the essays into texts ranging from 50 to 200 tokens in length and increasing at increments of five tokens (four texts of 50 tokens, three texts of 55 tokens, etc.). Fifty tokens was chosen as the shortest text length because our procedure for calculating moving averages required a window of 50 tokens. This also aligned with previous studies (Koizumi, 2012; Koizumi & In'nami, 2012) that considered texts ranging from 50 to 200 tokens in length. LD scores were then calculated for each of the texts using a new freely available text analysis tool (Kyle, Crossley, & Jarvis, in press), and values for texts of the same length were averaged. This resulted in 31 values for each LD index per essay.

All subsequent data analyses were performed using R software (R Core Team, 2017). For initial data inspection, line graphs were generated that showed the values plotted against text length for each index. The point of stabilization was further investigated by assessing the degree to which LD values were correlated with text length. Following Cohen's (1988) guidelines for interpreting effect sizes, we considered *r*-values below .100 to represent a 'small' effect.

To explore the effect that essay prompt and proficiency level had on the LD values, linear mixed-effects models were constructed with 'prompt' and 'level' as fixed effects and with 'participant' as a random effect. The maximal model included a random intercept for by-participant variance as well as a by-participant random slope for 'prompt.' All models were fit using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) in R. The model used in the analyses was entered as follows: value ~ prompt * level + (prompt | participant).

4. Results

4.1. Analysis with raw LD values

To visually assess how text length influences LD values, we generated line graphs in which the means of the raw values were plotted against text length for each index, as shown in Fig. 1. LD values for simple TTR and its most basic transformations (i.e., Root TTR, Log TTR, and Maas) did not fully level off anywhere in the range of text lengths included in the study. All the remaining indices became relatively stable at some point, either right from the beginning in the case of MATTR and HDD or at around 100 tokens in the case of

MTLD Original, MTLD-MA-BI, and MTLD-MA-Wrap. MATTR was stable across all text lengths. For each of the plots in Fig. 1, there is a slight perturbation in the lines at 105 tokens. This discontinuity may result from the fact that 105 tokens is the first length in the parallel sampling procedure at which the original 200-token text can no longer divided into multiple shorter texts (and thus there is no longer averaging across multiple LD measurements).

Because not all indices use the same scale of measurement, the LD values vary considerably from one index to the next. For each of the panels in Fig. 1, the *y*-axis was scaled to the maximum and minimum LD values for that index in the full dataset. In the next section, we perform a *z*-score transformation on the LD values so that comparisons can be made more easily across the indices.

4.2. Analysis with *z*-scores

To facilitate comparisons across indices, additional line graphs were generated in which mean *z*-scores of the LD values were plotted against text length (see Fig. 2). The advantage of using a *z*-score transformation is that it places all the LD values on a common scale.

The line graphs with *z*-scores display the same general pattern of results as the ones with raw values. The mean *z*-scores for MATTR, HD-D, MTLD Original, and MTLD-MA-Wrap at each text length never deviate much from the grand mean (represented by a *z*-score of

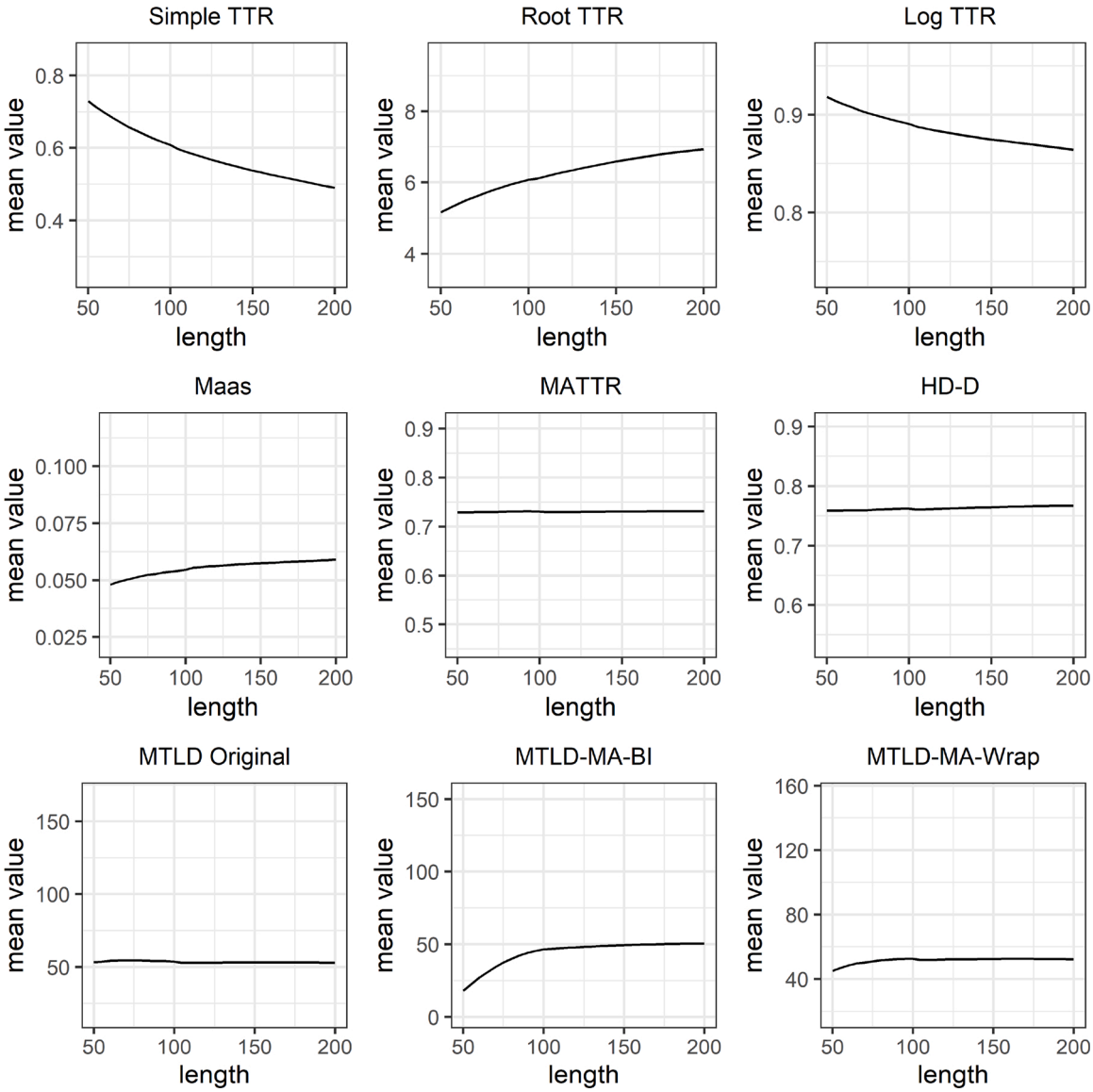

Fig. 1. Line graphs showing mean LD values plotted against text length for each index. The *y*-axis for each graph is scaled to match the full range of LD values for that index.

zero). This suggests that MATTR, HD-D, MTLD Original, and MTLD-MA-Wrap were more stable than the other indices.

The *z*-score analysis also provides an indication of the point of stabilization. Table 3 shows the mean *z*-score values organized by text length and index. The cells with asterisks indicate sequences of five or more consecutive *z*-scores that do not deviate from the final *z*-score by more than 0.10. This criterion provides a rough indication of the point at which stabilization occurs for each index. Simple TTR, Root TTR, and Log TTR do not fully stabilize at any point in the 50–200 token range. Maas, HD-D, and MTLD-MA-BI perform better, leveling off at 170 tokens, 130 tokens, and 140 tokens, respectively. MTLD Original and MTLD-MA-Wrap do better still, leveling off at 70 and 75 tokens, respectively. However, MATTR outperforms all the other indices, remaining stable for the entire 50–200 token range.

4.3. Correlational analyses

Another way to assess the stability of the indices is to examine the degree to which LD scores are correlated with text length. We performed Pearson correlations between LD scores and text lengths for each index, as shown in Table 4. If we follow Cohen's (1988) guidelines for interpreting effect sizes, in which *r*-values below .100 represent a negligible effect, then MATTR, HD-D, MTLD Original, and MTLD-MA-Wrap are the only indices that are sufficiently stable across all 31 text lengths included in the study.

Grouping the text lengths into smaller bins and repeating the correlational analysis can provide a more precise picture of the text

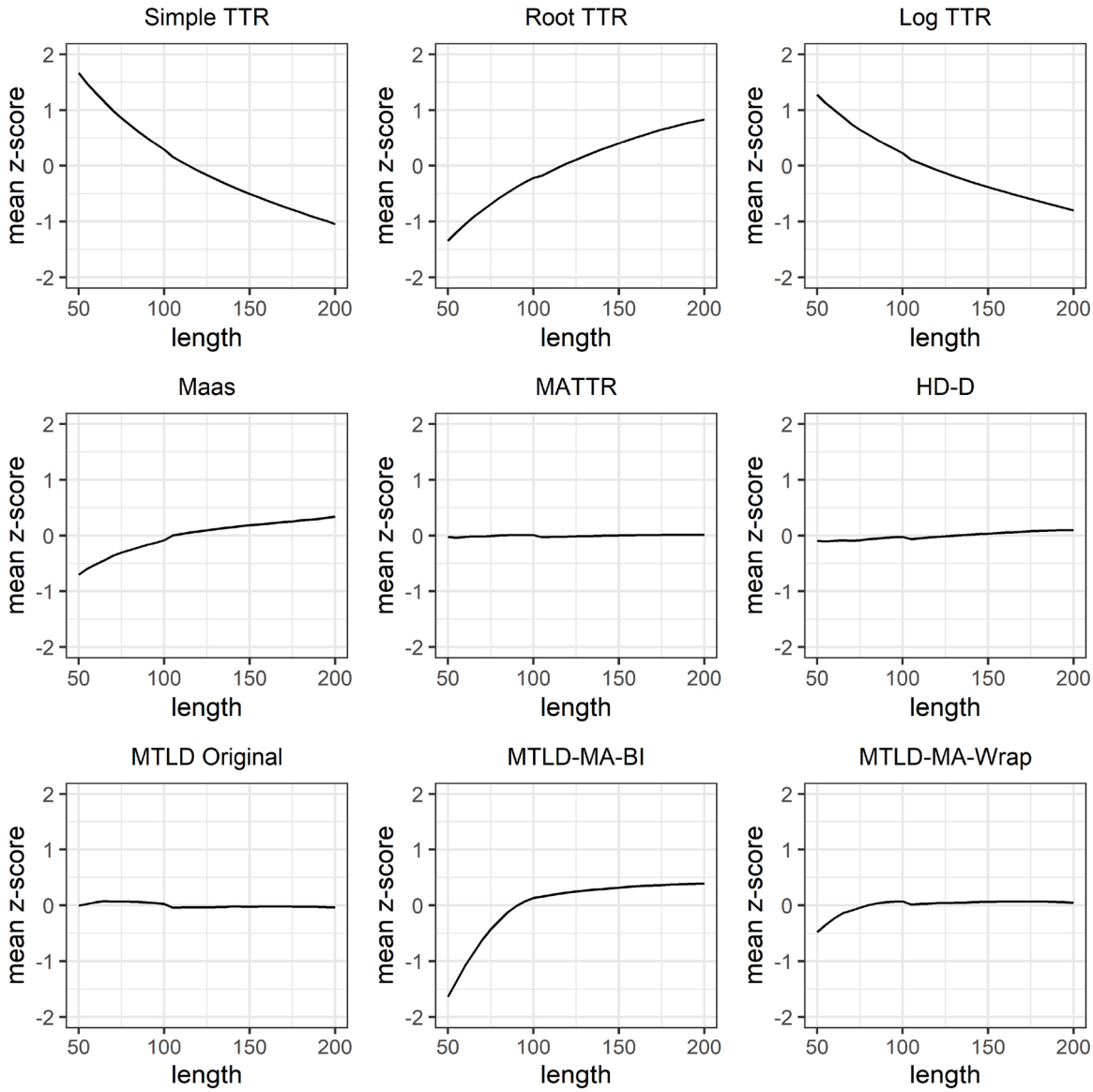

Fig. 2. Line graphs showing mean *z*-scores of the LD values plotted against text length for each index.

Table 3
Mean *z*-scores organized by length and index.

Length	Simple TTR	Root TTR	Log TTR	Maas	MATTR	HD-D	MTLD Original	MTLD-MA-BI	MTLD-MA-Wr.
50	1.67	−1.35	1.28	−0.70	−0.02*	−0.09	−0.01	−1.64	−0.48
55	1.48	−1.19	1.13	−0.59	−0.03*	−0.10	0.02	−1.36	−0.35
60	1.31	−1.05	1.00	−0.51	−0.02*	−0.09	0.06	−1.08	−0.23
65	1.15	−0.91	0.88	−0.44	−0.01*	−0.08	0.08	−0.84	−0.14
70	0.99	−0.80	0.75	−0.35	−0.01*	−0.09	0.07*	−0.62	−0.09
75	0.86	−0.69	0.65	−0.30	0.00*	−0.08	0.07*	−0.43	−0.04*
80	0.73	−0.57	0.56	−0.25	0.00*	−0.06	0.07*	−0.27	0.01*
85	0.61	−0.48	0.47	−0.21	0.01*	−0.05	0.06*	−0.12	0.04*
90	0.50	−0.38	0.39	−0.16	0.01*	−0.04	0.05*	−0.01	0.06*
95	0.40	−0.29	0.31	−0.12	0.01*	−0.03	0.04*	0.07	0.07*
100	0.29	−0.22	0.22	−0.08	0.01*	−0.02	0.02*	0.13	0.07*
105	0.16	−0.18	0.11	0.00	−0.02*	−0.06	−0.04*	0.16	0.02*
110	0.07	−0.10	0.05	0.03	−0.02*	−0.04	−0.04*	0.19	0.03*
115	−0.01	−0.03	−0.02	0.05	−0.02*	−0.03	−0.04*	0.21	0.03*
120	−0.09	0.04	−0.07	0.08	−0.01*	−0.02	−0.03*	0.23	0.04*
125	−0.17	0.11	−0.13	0.10	−0.01*	−0.01	−0.03*	0.25	0.04*
130	−0.24	0.17	−0.19	0.12	−0.01*	0.00*	−0.03*	0.27	0.05*
135	−0.31	0.23	−0.24	0.14	0.00*	0.01*	−0.03*	0.28	0.05*
140	−0.38	0.30	−0.29	0.15	0.00*	0.02*	−0.02*	0.29*	0.06*
145	−0.44	0.35	−0.34	0.17	0.00*	0.03*	−0.02*	0.31*	0.06*
150	−0.50	0.41	−0.38	0.19	0.00*	0.04*	−0.02*	0.32*	0.06*
155	−0.56	0.46	−0.43	0.20	0.01*	0.05*	−0.02*	0.33*	0.07*
160	−0.62	0.51	−0.47	0.22	0.01*	0.06*	−0.02*	0.34*	0.07*
165	−0.68	0.56	−0.51	0.23	0.01*	0.06*	−0.02*	0.35*	0.07*
170	−0.73	0.61	−0.55	0.24*	0.01*	0.07*	−0.02*	0.36*	0.07*
175	−0.78	0.65	−0.59	0.26*	0.02*	0.08*	−0.02*	0.37*	0.07*
180	−0.84	0.69	−0.64	0.27*	0.02*	0.09*	−0.02*	0.37*	0.07*
185	−0.89	0.73	−0.68	0.29*	0.02*	0.09*	−0.03*	0.38*	0.07*
190	−0.94	0.77	−0.72	0.30*	0.02*	0.10*	−0.03*	0.38*	0.06*
195	−0.99	0.80	−0.76	0.32*	0.02*	0.10*	−0.03*	0.39*	0.06*
200	−1.04	0.83	−0.80	0.34*	0.02*	0.10*	−0.03*	0.39*	0.05*

Note. Cells with asterisks indicate sequences of five or more *z*-scores that do not differ from the final *z*-score by more than 0.10.

length ranges at which indices perform well (or poorly). Following Koizumi (2012) and Koizumi and In'nami (2012), we grouped the text lengths into three bins (50–95 tokens, 100–145 tokens, and 150–195 tokens) and calculated Pearson correlations on the data in each bin (see Fig. 3). Texts of 200 tokens were excluded so that each bin would contain exactly 10 text lengths.

In the binned analysis, MATTR, HD-D, and MTLD Original all have *r*-values that remain below the .100 threshold across all bins. Maas, MTLD-MA-BI, and MTLD-MA-Wrap also have *r*-values that cross below .100 in the second and third bins. However, simple TTR, Root TTR, and Log TTR never meet this criterion, indicating that they are not stable at any of the text lengths included in the analysis. Note that in many cases the *r*-value for all three bins is lower than the *r*-value calculated on all text lengths (represented with a solid gray line in Fig. 3). This suggests that the *r*-values in the binned analysis are slightly depressed due to the reduced range of text lengths included in each bin. We do not report other analyses with larger numbers of bins here because that would have meant reducing sample sizes even further. The *r*-values for each bin and index are shown in Table 5.

The results of the correlational analyses are consistent with our initial inspection of the data. MATTR, HD-D, and MTLD Original are the most stable of the indices included in the study. Maas, MTLD-MA-BI, and MTLD-MA-Wrap produce somewhat erratic scores in the 50–95 token range but are much more stable with texts of 100 tokens or more. Conversely, simple TTR, Log TTR, and Root TTR do not meet the .100 criterion for any of the text length ranges included in our analysis.

4.4. Comparisons by essay prompt and proficiency level

To visualize how essay prompt influences LD values, we generated line graphs showing mean *z*-scores plotted against text length for each prompt, as shown in Fig. 4. We also explored the effects of proficiency level on LD values by generating line graphs in which the mean *z*-scores were plotted against text length for each proficiency level, as shown in Fig. 5. These two figures confirm that the LD values follow the same trajectories even when the data are separated by prompt or level, thereby indicating that the stabilization pattern for each index is consistent across prompts and levels. Figs. 4 and 5 also clearly show that the mean LD values are not equal for

Table 4
Absolute values of *r*-values calculated across all text lengths for each index.

Token Range	Simple TTR	Root TTR	Log TTR	Maas	MATTR	HD-D	MTLD Original	MTLD-MA-BI	MTLD-MA-Wr.
50–200	.762	.615	.581	.272	.010*	.064*	.028*	.444	.086*

Note. Cells with asterisks indicate *r*-values below .100.

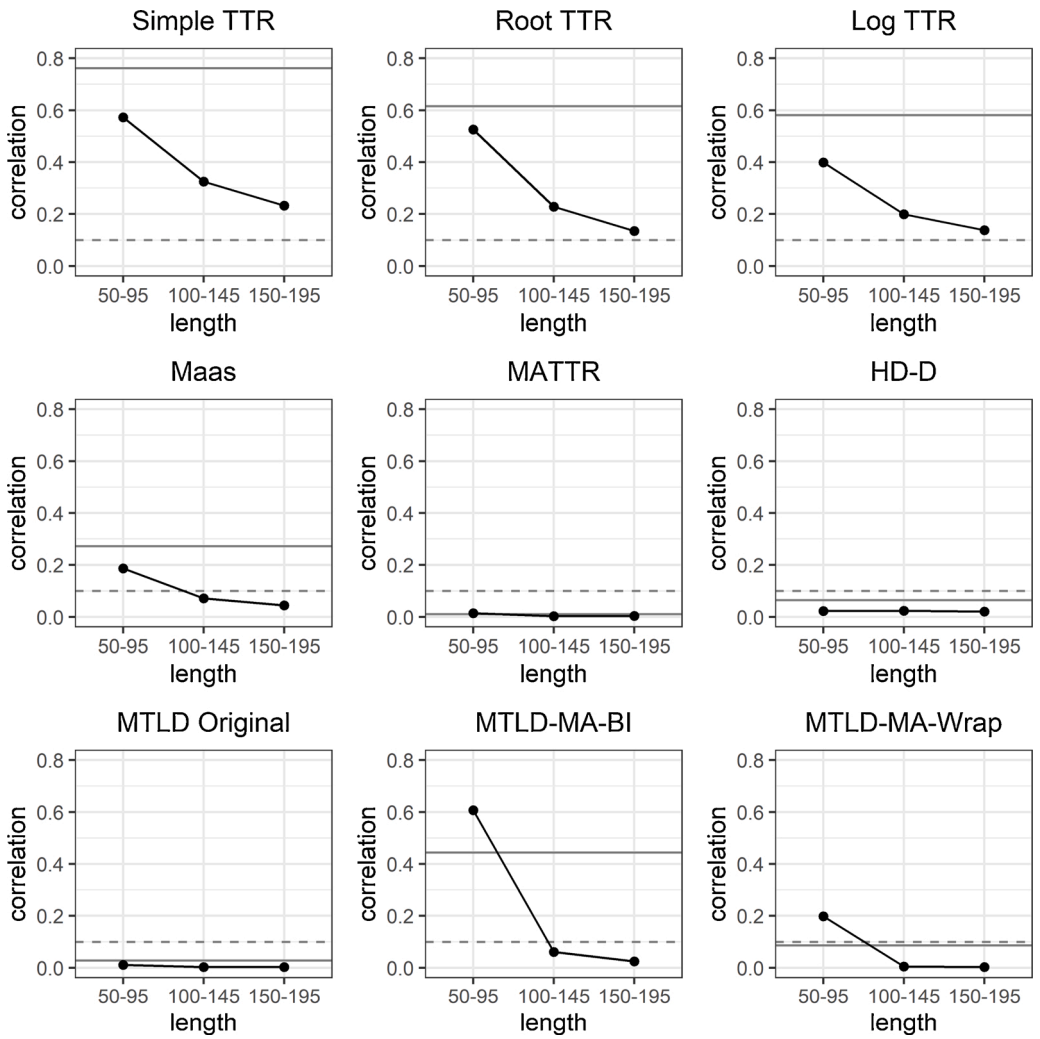

Fig. 3. Line graphs showing absolute values of *r*-values from Pearson correlations plotted against text length bins for each index. The dotted gray line indicates an *r*-value of .100 and the solid gray line represents the *r*-value from a Pearson correlation computed using all text lengths.

Table 5
Absolute values of *r*-values calculated by text length bin for each index.

Token Range	Simple TTR	Root TTR	Log TTR	Maas	MATTR	HD-D	MTLD Original	MTLD-MA-BI	MTLD-MA-Wr.
50–95	.573	.525	.399	.187	.014*	.022*	.012*	.607	.198
100–145	.325	.228	.200	.072*	.003*	.024*	.002*	.061*	.005*
150–195	.233	.135	.138	.044*	.004*	.020*	.002	.024*	.003*

Note. Asterisks indicate *r*-values below .100.

each prompt and level. In Fig. 4, the lines for the Smoke prompt are consistently higher than those for the Job prompt, except in the Maas panel where the pattern is reversed (in contrast to the other diversity indices analyzed in this study, lower Maas values indicate more diversity). Similarly, higher proficiency levels are associated with higher LD scores in Fig. 5, which is expected because writers with higher English proficiency are expected to have a wider variety of lexical items to draw on than those with lower proficiency (e.g., Engber, 1995; Treffers-Daller, Parslow, & Williams, 2016).

To investigate whether the observed differences between the essay prompts and proficiency levels were statistically significant, we constructed a linear mixed-effects model with 'prompt' and 'level' as fixed effects and with 'participant' as a random effect. The maximal model included a random intercept for by-participant variance as well as a by-participant random slope for 'prompt.' Dummy coding was used to facilitate comparisons between the different levels of the fixed effects. Since differences across the various indices were not relevant to the analysis, only a single index—MATTR—was used. The fixed effects output from the model summary is shown in Table 6.

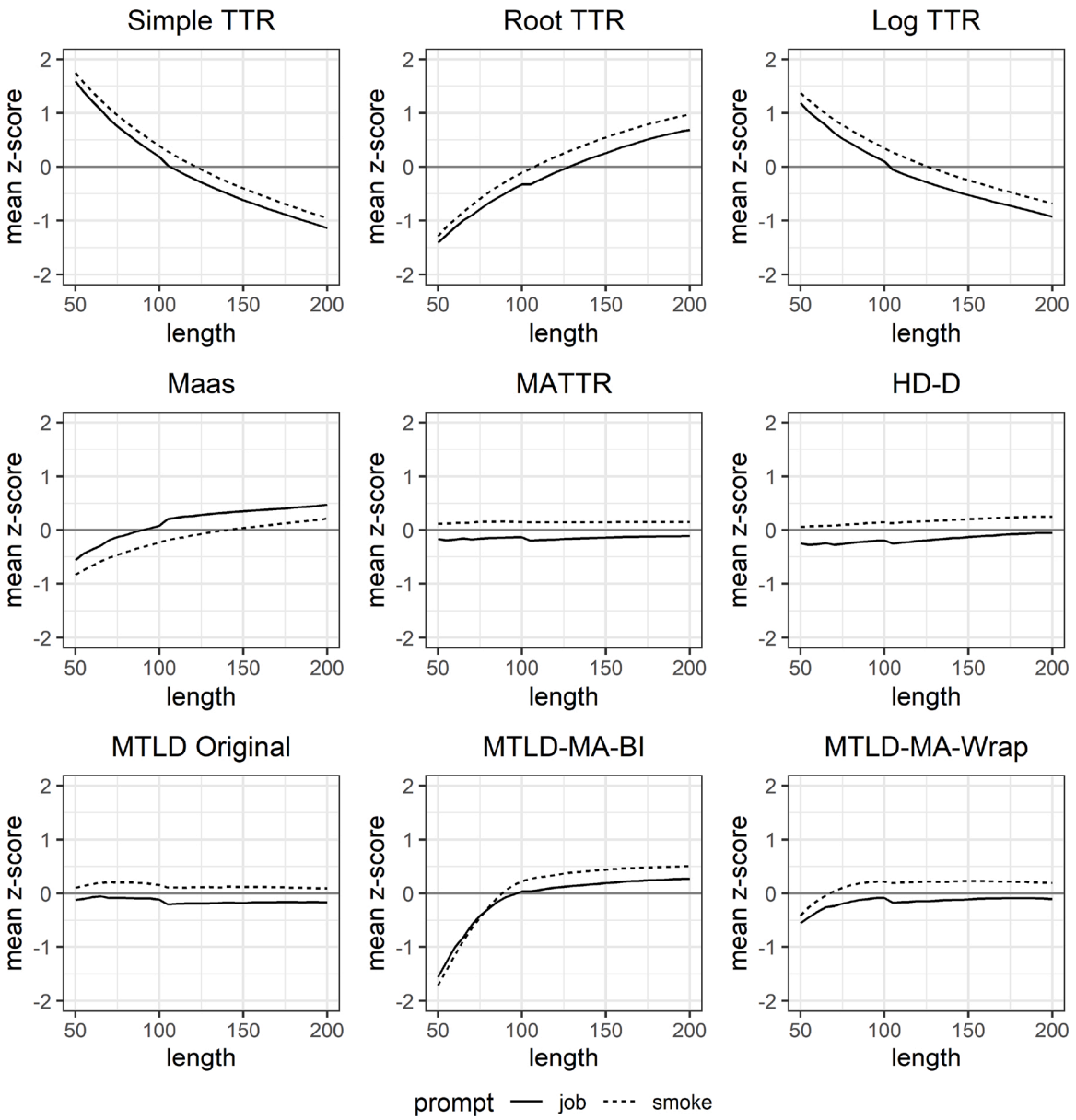

Fig. 4. Line graphs showing mean *z*-scores of the LD values plotted against text length for each essay prompt and organized by index.

The intercept estimate of 0.709 represents the mean MATTR value for the A2 level and the Job prompt. Changing the prompt from Job to Smoke was associated with an estimated increase of 0.012, which was a small but statistically significant difference ($p < .001$). The interactions between levels and prompts were not statistically significant. Going from level A2 to level B1 was associated with an estimated increase in MATTR values of 0.015, a difference which was also small but statistically significant ($p < .001$). The reference level for the analysis was then revised from A2 to B1 to determine the estimated change in LD score between B1 and B2. The results (see Table 7), indicated a significant ($p < .001$) but small increase of 0.019 in LD score between B1 and B2. Assumptions regarding normality of residuals were checked and found to be satisfied.

5. Discussion

5.1. Stabilization points for LD indices

Preliminary inspection of the mean raw values and *z*-scores suggests that values for MATTR, MTLD Original, and HD-D are highly stable across all the text lengths included in this study. MTLD-MA-Wrap, MTLD-MA-BI, and Maas are unstable at the lowest text lengths

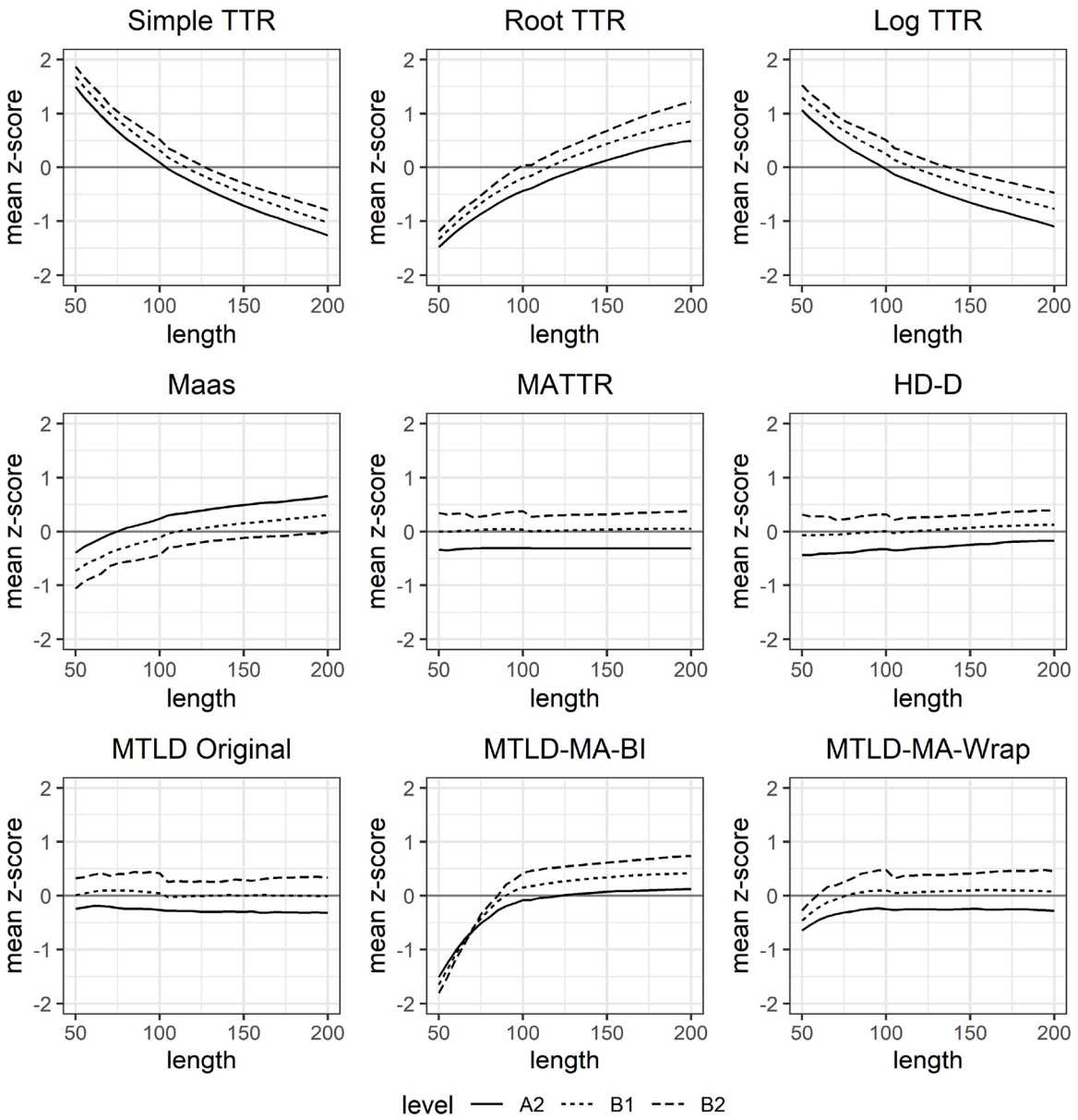

Fig. 5. Line graphs showing mean *z*-scores of the LD values plotted against text length for each proficiency level and organized by index.

Table 6
Fixed effects from linear mixed-effects model output.

	Estimate	SE	*df*	*t*-value	*p*-value
intercept	0.709	0.002	2339.2	229.16	<.001
level = B1	0.015	0.003	2329.8	5.76	<.001
level = B2	0.034	0.004	2292.9	8.57	<.001
prompt = smoke	0.012	0.003	2353.5	4.61	<.001
level = B1 × prompt = smoke	0.004	0.003	2332.5	1.49	.136
level = B2 × prompt = smoke	−0.004	0.004	2254.1	−0.88	.381

Note. The reference level is A2 for proficiency level and Job for essay prompt.

but eventually reach a point of relative stabilization. The remaining indices—simple TTR, Root TTR, and Log TTR—never level off at any point in the 50–200 token range.

The initial observations were supported by further analysis with Pearson correlations. For MATTR, MTLD Original, MTLD-MA-

Table 7
Fixed effects from linear mixed-effects model output with B1 as the reference level.

	Estimate	SE	*df*	*t*-value	*p*-value
intercept	0.724	0.001	2281.4	643.28	<.001
level = A2	−0.015	0.003	2328.4	−5.76	<.001
level = B2	0.019	0.003	2264.2	5.59	<.001
prompt = smoke	0.016	0.001	2236.4	13.13	<.001
level = A2 × prompt = smoke	0.004	0.003	2332.6	−1.49	.136
level = B2 × prompt = smoke	−0.008	0.004	2201.2	−2.19	.029

Note. The reference level is B1 for proficiency level and Job for essay prompt.

Wrap, and HD-D, the correlations between text length and LD value were negligible ($r < .100$) across all text lengths (50–200 words). MATTR, MTLD Original, and HD-D also demonstrated negligible relationships with text length across all three bins that were considered (50–95 tokens, 100–145 tokens, and 150– 195 tokens). MTLD-MA-Wrap demonstrated a small correlation with text length in the first bin (50–95 tokens), but negligible correlations in the other two bins. This suggests that MATTR, MTLD Original, and HD-D are suitable for working with short texts in the 50–200 token range and that MTLD-MA-Wrap is likely suitable for working with texts in the 100–200 token range. This finding with regard to MTLD Original contrasts with those of Koizumi (2012) and Koizumi and In'nami (2012), who found that MTLD became stable at around 100 tokens (presumably they used MTLD Original). These differences may be due to the small sample ($n = 38$) used in these two previous studies. It is also possible, however, that the disparities are related to differences in spoken vs. written production.

Significant and meaningful correlations were observed across all text lengths (50–200 tokens) and the five remaining indices (Simple TTR, Root TTR, Log TTR, Maas, MTLD-MA-BI). In the binned analysis, two of these indices (Maas and MTLD-MA-BI) demonstrated negligible relationships ($r < .100$) with text length in the two longer bins (100–145 tokens and 150–195 tokens) but demonstrated meaningful correlations in the shorter bin (50–95 tokens). This indicates that MTLD-MA-BI and Maas do not produce stable values in the 50–95 token range, but do produce stable values in texts of 100 tokens or more. For simple TTR, Root TTR, and Log TTR, the *r*-values do not dip below .100 in any of the three text length bins, suggesting that they are significantly correlated with text length across the entire 50–200 token range. This suggests that these indices are inappropriate for the analysis of short L2 texts of varying lengths, a result that was expected because McCarthy and Jarvis (2007, 2010) had found that traditional LD measures such as TTR remained sensitive to text length even at much higher token counts.

MATTR appears to be the most stable of all the indices, with an *r*-value of just .010 when all text lengths are included in the analysis. Even if a much stricter cutoff value were used (e.g., $r \leq .015$), the *r*-values for MATTR would still be sufficiently low across the three text length bins (see Table 5). The only other indices that would achieve stabilization using this stricter criterion would be MTLD Original (across all three bins) and MTLD-MA-Wrap (in the second and third bins). We therefore suggest that these three indices—MATTR, MTLD Original, and MTLD-MA-Wrap—are the most appropriate indices for working with short L2 texts.

Table 8 summarizes our recommendations regarding the use of each index for the analysis of short L2 written texts. MATTR, MTLD Original, and MTLD-MA-Wrap are recommended indices based on their low *r*-values in the binned analysis. HD-D, MTLD-MA-BI, and Maas all manage to produce *r*-values below .100 somewhere in the 50–200 token range, but they are not as resistant to text length effects as the first three indices. Log TTR, Root TTR, and simple TTR are not recommended at any of the text lengths included in the current study.

5.2. Comparisons by essay prompt and proficiency level

Visual inspection of the data using line graphs indicated small but persistent effects related to essay prompt and proficiency level on the LD values. First, the average values were higher for the Smoke prompt than for the Job prompt. The values also increased along with the writer's proficiency level. Further analysis with linear mixed-effects models on the MATTR values confirmed this pattern of observations. The increase in MATTR values associated with going from the Job prompt to the Smoke prompt was small but statistically significant. This suggests that the Smoke prompt elicited more elaborate responses than the Job prompt, thus necessitating the

Table 8
General guidelines for using LD indices with short L2 written texts.

Recommendation	Index	Minimum Text Length
Use with confidence	MATTR	50 tokens
	MTLD Original	50 tokens
	MTLD-MA-Wrap	100 tokens
Use with caution	HD-D	50 tokens
	MTLD-MA-BI	100 tokens
	Maas	100 tokens
Avoid using with short texts	Log TTR	—
	Root TTR	—
	Simple TTR	—

use of a wider range of lexical items. The increases in MATTR values associated with going from level A2 to B1 and from level B1 to B2 were also small but statistically significant. This pattern of results suggests that, as we might expect, more proficient L2 writers have a larger and more varied productive vocabulary than less proficient ones (e.g, Engber, 1995) regardless of prompt. However, as previous research has demonstrated with other linguistic indices (e.g., Kyle, Crossley, & McNamara, 2016; Reid, 1986) caution clearly needs to be taken when comparing lexical diversity scores across prompts as the analyses consistently demonstrated higher LD scores for the Smoke prompt.

5.3. Implications

This study has two main implications for writing assessment researchers and developers of automatic evaluation/scoring systems. First, the results indicate that the most stable index of lexical diversity (across texts from 50 to 200 words in length) is MATTR and that two versions of MTLD (Original and MA-Wrap) are also relatively stable. In other words, it is recommended that MATTR be used to measure LD in short texts unless a stronger validity argument can be made for MTLD. Furthermore, the results indicate that a number of well-known indices are not stable across texts from 50 to 200 words (e.g., Simple TTR, Root TTR, Log TTR, Maas, and MTLD-MA-BI) and should not be used. Second, the results indicated that lexical diversity values vary systematically across the responses to two writing prompts explored in this study. Writing researchers and developers of automatic evaluation/scoring systems should therefore use caution when comparing lexical diversity scores across samples written in response to different prompts. This issue can be mitigated by either keeping prompt constant or statistically controlling for differences across prompts.

5.4. Limitations and future directions

One limitation of this study is the fact that the data were unbalanced, especially when it came to comparisons between proficiency levels. In particular, there were many more learners in the B1 proficiency group than in either of the other groups (see Table 2). The results related to proficiency should therefore be interpreted with caution. We recommend that future researchers make exclusions where appropriate to ensure that the data are sufficiently balanced.

Another area where future research may be needed is in regard to the range of text lengths considered in the analysis. Because some indices were highly stable across all text lengths, it was not possible to precisely determine their stabilization points (i.e., their values may be stable in texts shorter than 50 words). We therefore suggest that future studies use a wider range of text lengths in their analyses. Note however that if texts shorter than 50 tokens are used, it will be necessary to adjust the widow size for indices involving moving-average calculations. For example, it would not have been possible to analyze shorter text lengths with the version of MATTR used in this study because the window size was set to 50 tokens.

A third, related area for future research pertains to lexical diversity scores across longer L2 texts, which may be common at higher proficiency levels and/or in academic and pre-academic settings. Future research should investigate the relationship between text length, proficiency, and LD scores across longer texts (e.g., 750–1500 words).

Finally, the current study only investigated the stability of indices and did not address issues related to construct validity. Jarvis (2013, 2017) has argued that most existing LD indices lack construct validity because they were not formulated with a carefully thought-out theoretical model of LD in mind. Instead, previous research validating LD indices has focused mainly on probing how resistant they are to text length effects and how strongly they correlate with other constructs. Jarvis' proposed solution involves collecting human judgments of LD and then developing a multifaceted measure of LD that comes as close as possible to replicating those judgments. However, such measures require corpus-specific calibration and thus are not appropriate for comparisons across different corpora and text lengths. Jarvis' concerns regarding construct validity are certainly important and merit further investigation. Despite such questions about what exactly LD indices should be attempting to measure, though, the fact remains that existing LD indices based on types and tokens are likely to remain widely-used tools for LD research well into the future, and thus it is useful to have information about which indices are most reliable across texts of different lengths.

6. Conclusion

LD indices are commonly used in writing assessment research. However, many LD indices are sensitive to text length effects and therefore conflate vocabulary breadth and text length. This study has explored the stability of LD indices at different text lengths and established clear guidelines for the minimum text lengths to be used with each index. Some indices, such as simple TTR, Root TTR, and Log TTR, never stabilized within the entire 50–200 token range and therefore should not be used for the analysis of short L2 written texts. Others, such as Maas, MTLD-MA-BI, and MTLD-MA-Wrap exhibited turbulence in the 50–95 token range but became more stable at around 100 tokens. The remaining three indices—HD-D, MTLD Original, and MATTR—remained relatively stable across the entire 50–200 token range. We recommend using MATTR, MTLD Original, and MTLD-MA-Wrap for the analysis of short L2 written texts because these indices displayed the smallest degree of text length effects, although the MTLD-MA-Wrap is not appropriate for use with texts shorter than 100 tokens. MATTR performed particularly well, displaying a high degree of stability across the entire 50– 200 token range. As expected from previous L2 production research (e.g., Engber, 1995; Treffers-Daller et al., 2016), a significant relationship was found between LD values and proficiency, indicating that LD is an indicator of proficiency. This is not to say that writing assessment can be reduced to automated measures of LD; instead, our data simply suggest that LD is an important component of L2 writing ability. Differences in LD scores across proficiency levels were small, likely due to the relatively small range of proficiency levels analyzed (CEFR A2–B2). Additionally, as has been found with regard to other indices related to linguistic proficiency (e.g., Kyle

et al., 2016; Reid, 1986) the results of this study indicated that there were small but systematic differences in LD scores across different prompts. This suggests that it may be important to control for prompt in studies that use LD as an index of proficiency. We hope that these findings will be of use to writing researchers and teachers who have a need to analyze short L2 English essays on a regular basis.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgements

We would like to express our gratitude to the editors, Martin East and David Slomp, and to our anonymous reviewers for their helpful comments and suggestions. We would also like to thank the students of the Spring 2018 section of SLS 750 'Learner Corpus Research' at the University of Hawai'i at Mānoa for their constructive feedback.

References

Biggs, A., Daniel, L., Feather, R. M., Ortleb, E., Rillero, P., Snyder, S. L., et al. (2003). *Glencoe science: Level green*. New York: Glencoe/McGraw-Hill.
Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.
Chotlos, J. W. (1944). Studies in language behavior IV: A statistical and comparative analysis of individual written language samples. *Psychological Monographs, 56*(2), 77–111.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203771587
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
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. https://doi.org/10.1080/09296171003643098
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561–580. https://doi.org/10.1177/0265532210378031
Cumming, A., Kantor, R., Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL (TOEFL Monograph No. MS-30)*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01990.x
Dale, P., Bates, E., Reznick, S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. *Journal of Child Language, 16*(2), 239–249. https://doi.org/10.1017/s0305000900010394
Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7
Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique [Problems and methods of linguistic statistics]*. Dordrecht: Reidel.
Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.
Hess, C. W., Haug, H., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research, 32*(3), 536–540. https://doi.org/10.1044/jshr.3203.536
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*(1), 129–134. https://doi.org/10.1044/jshr.2901.129
Holmes, J., Vine, B., & Johnson, G. (1998). *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.
Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies around the world, 1* pp. 91–118). Kobe, Japan: Kobe University.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa
Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x
Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537–553. https://doi.org/10.1177/0265532217710632
Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo.
Johansson, V. (1999). Word frequencies in speech and writing: A study of expository discourse. In R. A. Aisenman (Ed.), *Working papers in developing literacy across genres, modalities, and languages* (Vol. 1, pp. 182–198). Tel Aviv: Tel Aviv University Press.
Johnson, W. (1944). Studies in language behavior 1: A program of research. *Psychological Monographs, 56*, 1–15.
Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction, 1*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi
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. https://doi.org/10.1016/j.system.2012.10.012
Kucera, H., & Francis, W. N. (1967). Computational analysis of Swedish learners' written English. *Lund Studies in English, 74*. Malmö: Liber Förlag.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13
Kyle, K. (2020). Measuring lexical richness. In S.Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454–476). New York: Routledge. https://doi.org/10.4324/9780429291586-29.
Kyle, K., Crossley, S. A., & Jarvis, S. (in press) Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*. doi:10.1080/15434303.2020.1844205.
Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319–340. https://doi.org/10.1177/0265532215587391
Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmö: CWK Gleerup.
Maas, H. D. (1972). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes [On the relationship between vocabulary and the length of a text]. *Zeitschrift für Literaturwissenschaft und Linguistik, 2*(8), 73.
Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Hampshire, UK: Palgrave Macmillan. https://doi.org/10.1057/9780230511804

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual lexical diversity (MTLD)* (Doctoral dissertation). Memphis, TN: University of Memphis https://www.aaai.org/ocs/index.php/FLAIRS/2010/paper/view/1283.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488. https://doi.org/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. https://doi.org/10.3758/brm.42.2.381

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://doi.org/10.26686/wgtn.12552197

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167–188). Alexandria, VA: TESOL.

Snow, C. E. (1989). Imitativeness: A trait or a skill? In G. E. Speidel, & K. E. Nelson (Eds.), *The many faces of imitation in language learning* (pp. 75–88). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-1011-5_4.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation*. Lund: CWK Gleerup.

Swales, J. M., & Malczewski, B. (2001). Discourse management and new-episode flags in MICASE. In R. C. Simpson, & J. M. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 145–164). Ann Arbor, MI: University of Michigan Press.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD, and HD-D as measures of language ability. In S. Jarvis, & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). Amsterdam: Benjamins. https://doi.org/10.1075/sibil.47.05ch3.

Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics, 39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Quantifying lexical diversity in the study of language development*. Honolulu, HI: Second Language Teaching and Curriculum Development Center, University of Hawaii.

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

Fred Zenker is a PhD student in the Department of Second Language Studies at the University of Hawai'i at Mānoa. His research interests include experimental syntax, applied psycholinguistics, and learner corpus research. He is particularly interested in investigating pronominal resumption in L1/L2 English.

Kristopher Kyle is an Assistant Professor in the Department of Linguistics at the University of Oregon and holds a joint Assistant Professorship in the English Department at Yonsei University. His research interests include second language writing and speaking, assessment, and second language acquisition. He is especially interested in applying natural language processing (NLP) and corpora to the exploration of these areas.