Vocabulary Learning and Instruction
Volume 1, Issue 1, August 2012
http://vli-journal.org

Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?

Rie Koizumi
Tokiwa University
doi: http://dx.doi.org/10.7820/vli.v01.1.koizumi

Abstract

Lexical diversity (LD) measures have been known to be sensitive to the length of the text, and numerous revised LD measures have been proposed. This study aims to identify LD measures that are least affected by text length and can be used for the analysis of short L2 texts (50–200 tokens). This study compares the type-token ratio, Guiraud index, D, and measure of textual lexical diversity (MTLD) to assess their degree of susceptibility to text length. Spoken texts of 200 tokens from 20 L2 English learners at the lower-intermediate-level were divided into segments of 50 to 200 tokens and the text length impact was examined. It was found that MTLD was least affected by text length, and that it should be used with texts of at least 100 tokens.

Keywords: lexical diversity; text length; type-token ratio; Guiraud index; D; measure of textual lexical diversity (MTLD); speaking performance.

1 Introduction

How do L2 learners use vocabulary knowledge in their speaking and writing? Research on vocabulary use, or, as it has generally been termed, lexical richness, is a vital but under-researched topic (Schmitt, 2010; Skehan, 2009). Among several aspects of lexical richness, lexical diversity (LD, also known as lexical variation) refers to "the range and variety of vocabulary deployed in a text by either a speaker or a writer" (McCarthy & Jarvis, 2007, p. 459). Numerous measures of LD have been proposed (see Malvern, Richards, Chipere, & Durán, 2004 for a list), but perhaps the best known measure is the type-token ratio (TTR), in which the number of different words a learner writes in a text is divided by the total number of words in order to determine the degree of variation.

However, an acknowledged drawback of many such measures is that they have been known to be sensitive to the length of the text analyzed. Effects of text length on LD measures should be avoided as much as possible because they generate construct-irrelevant variances and misleading results. Numerous revised LD measures have been proposed to remedy this risk, including the Guiraud index (Guiraud, hereafter; Guiraud, 1960), D (Malvern et al., 2004), and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010), to name a few.

Previous studies (Hess, Sefton, & Landry, 1986; Malvern et al., 2004; McCarthy & Jarvis, 2007, 2010) show that like TTR, Guiraud (values) are substantially affected by text length, whereas D and MTLD (values) are affected

Please cite this article as: 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. doi: 10.7820/vli.v01.1.koizumi

only to a small degree. However, previous studies on D and MTLD examined the impact of length of texts of 100 tokens or more (McCarthy & Jarvis, 2007, 2010). Consequently, use of D and MTLD could still be problematic for analysis of texts under 100 tokens. This is unfortunate, as evaluating the quality of such short texts is sometimes necessary in studies of L2 production and vocabulary.

The present study investigates whether LD measures are usable for texts of <100 tokens, in order to determine if these measures can be used for analysis of shorter texts produced by learners. This study compares four LD measures (TTR, Guiraud, D, and MTLD) for analyzing short texts of 50 to 200 tokens and addresses the following research question: Which LD measures are least affected by text length across a range of 50 to 200 tokens?

In the next section, these four LD measures and their characteristics are summarized.

2 TTR, Guiraud, D, and MTLD

Brief explanations and examples of how to calculate the four LD measures are provided below.

Here is a spoken text from a male junior-high-school student describing a picture and comparing two pictures. His entire text consisted of 253 tokens, and 50 tokens were selected:

> *This mountain is very big and the scene is very good and there is a lake. Lake is very beautiful and beside there is a tree and over there I see the big mountain. There is a very beautiful/And this Taro's room. Before is Taro's name on textbook but*

The exemplar text has 25 types and 50 tokens. We can analyze this text by using any of the following four measures:

(a) *TTR*: the number of different words (types)/all words produced (tokens). In this example, the TTR is 0.50 (i.e. 25/50).
(b) *Guiraud*: the number of types/the square root of the number of tokens (types/$\sqrt{}$tokens). In this example, the Guiraud is 3.54 (i.e. 25/$\sqrt{}$50).
(c) *D*: Sample 35 tokens randomly from a text analyzed like this: *This*, *but*, *this*, *mountain*, *on*, *There*, *is*, *Before*, *big*, *the*, *very*, *is*, *good*, *beautiful*, *And*, *name*, *there*, *a*, *Lake*, *Taro's*, *very*, *and*, *there*, *a*, *textbook*, *and*, *there*, *see*, *big*, *mountain*, *is*, *a*, *room*, *is*, *Taro's*.

Once a word is selected, it is not selected again (i.e. without replacement). This randomly selected 35-token segment has TTR of 0.57, with 20 types and 35 tokens. Then, another segment of 35 tokens is sampled and the TTR is calculated. Sampling up to 100 times should be repeated and the mean TTR for the hundred 35-token segments (0.57) should be computed. Similarly, producing the mean TTR for hundred 36-token segments (0.56), the mean TTR for hundred 37-token segments (0.54), and so on, up to the mean TTR for hundred 50-token segments (0.50) should be continued. Figure 1 plots 16 means for each token segment and shows the empirical TTR curve. Identify the D value for which the theoretical curve best fits the empirically derived TTR curve. D for the 50-token text was 12.00.

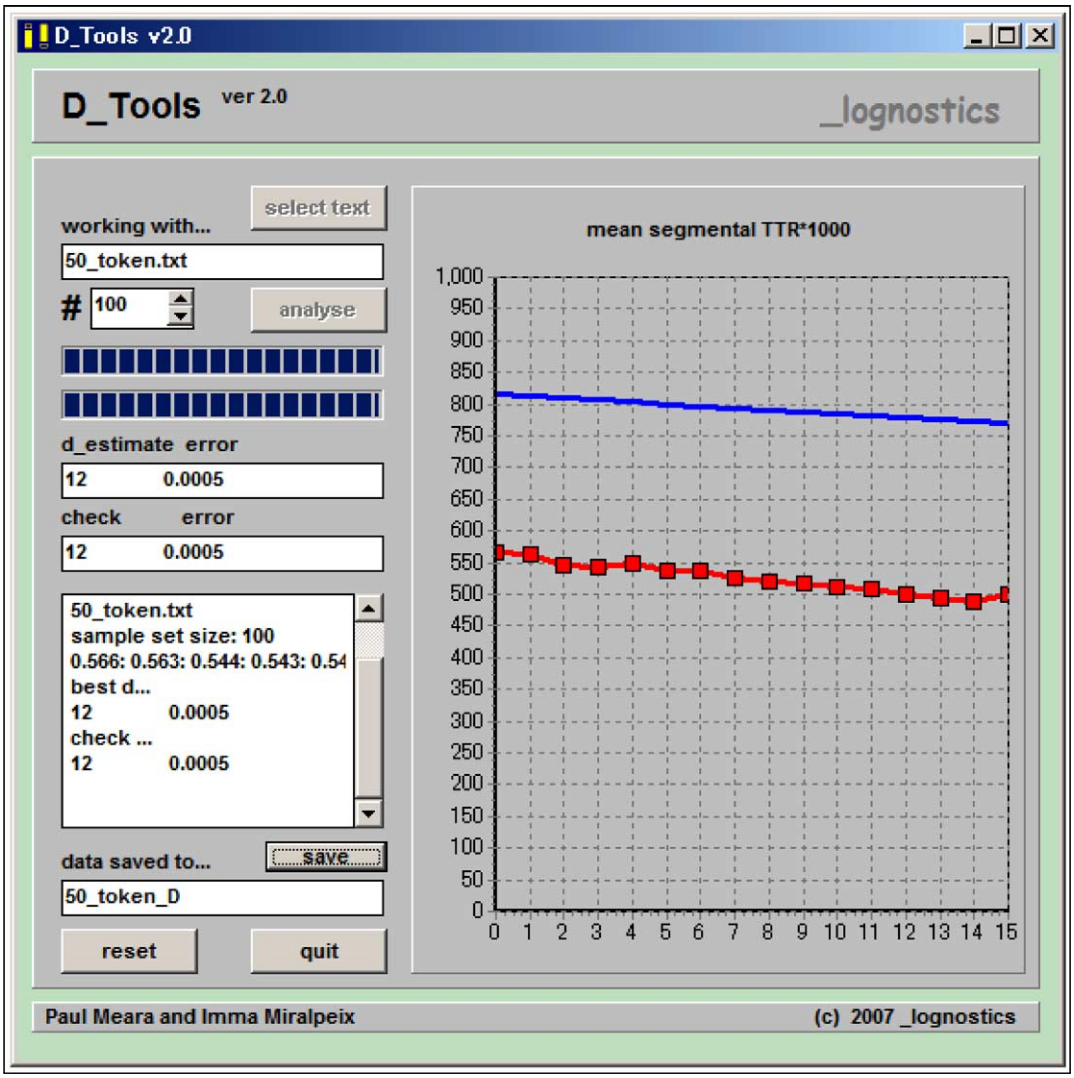

Figure 1. Example of D calculation, using D_Tool (Version 2.0; Meara & Miralpeix, 2007).

D can be computed using *vocd* (McKee, Malvern, & Richards, 2000) or D_Tool (Meara & Miralpeix, 2007).

(d) *MTLD*: "the mean length of sequential word strings in a text that maintain a given TTR value" (0.720; McCarthy & Jarvis, 2010, p. 384). Count (x) the number of times the text reaches TTR of 0.72 or below, from the beginning of the text through to the end. As seen in Table 1, TTR is calculated to obtain (x) for a segment from the beginning by increasing words successively. For example, the tenth word, *very*, is the eighth type, with TTR of 0.80. A segment up to the fourteenth word, *is*, has TTR of 0.72 or below (i.e. 0.71). This word is the last one of the first word string (factor) of the text. Then, the next word, *a*, becomes the first word of the next string. This time, the third word, *lake*, has TTR of 0.67, but this segment is not counted because it has less than 10 words. McCarthy (2005) decided not to consider "factors of less than ten words, ... as they may only represent a brief syntactical or rhetorical, textual blip" (p. 106). The second word string ends with the 37th word, *a*, with a TTR of 0.70. The third string starts with *very* but does not reach 0.72 (0.92); however, the value of 0.92 reaches 27% of the trajectory from 1.00 to 0.72

Table 1. Example of measure of textual lexical diversity Forward Calculation

	this	*mountain*	*is*	*very*	*big*	*and*	*the*	*scene*	*is*	*very*
Type	1	2	3	4	5	6	7	8	8	8
Token	1	2	3	4	5	6	7	8	9	10
TTR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.80
	good	*and*	*there*	*is*	*a*	*Lake*	*lake*	*is*	*very*	*beautiful*
Type	9	9	10	10	1	2	2	1	2	3
Token	11	12	13	14	1	2	3	1	2	3
TTR	0.82	0.75	0.77	0.71	1.00	1.00	0.67	1.00	1.00	1.00
	and	*beside*	*there*	*is*	*a*	*Tree*	*and*	*over*	*there*	*I*
Type	4	5	6	6	7	8	8	9	9	10
Token	4	5	6	7	8	9	10	11	12	13
TTR	1.00	1.00	1.00	0.86	0.88	0.89	0.80	0.82	0.75	0.77
	see	*the*	*big*	*mountain*	*there*	*is*	*a*	*very*	*beautiful*	*and*
Type	11	12	13	14	14	14	14	1	2	3
Token	14	15	16	17	18	19	20	1	2	3
TTR	0.79	0.80	0.81	0.82	0.78	0.74	0.70	1.00	1.00	1.00
	this	*Taro's*	*room*	*before*	*is*	*Taro's*	*name*	*on*	*textbook*	*but*
Type	4	5	6	7	8	8	9	10	11	12
Token	4	5	6	7	8	9	10	11	12	13
TTR	1.00	1.00	1.00	1.00	1.00	0.89	0.90	0.91	0.92	0.92

Note. TTR, type-token ratio. Underlined = TTR of 0.72 or below.

(i.e. [1.00–0.92]/[1.00–0.72] = 0.08/0.28). The remaining segment, which does not arrive at 0.72, is taken into consideration to enhance the reliability of MTLD (McCarthy, 2005). Therefore, (x), the number of times the text reaches TTR of 0.72 or below, was two plus 0.27 (i.e. 2.27). The mean number of words required is computed using the number of tokens divided by (x) as a formula; in the example, it is 21.03 (i.e. 50/2.27). Similarly, the calculation is made backward from the last word, *but*, to the first word, *this*, which produces another value (27.41). The two values derived from forward calculation and from backward calculation are averaged; the example has a value of 24.70.

MTLD can be derived easily with the use of the Gramulator (Version 5.0; McCarthy, 2011) as follows:

(a) Start the Gramulator and push the !Start! button. Figure 2 appears.
(b) Click on the folder already saved, which appears below "Browse to desired folder." Texts in the folder are shown in "Folder Contents." Click on the Click to Retain this Corpus button to add the selected folder to the corpus.
(c) Click on the folder name in "Click corpus to load." Then, the folder is selected and shown in the bottom right (e.g. "Corpus 1 is Gramulator").
(d) Go to [Modules] → [Assessment] → [The Evaluator]. Figure 3 appears.
(e) Select the folder in "Select Corpus" in the right. Select "Measures" in "Index Banks" in the left. Select "MTLD" in the middle. (You can also obtain HD-D and Maas here; see McCarthy & Jarvis, 2010).
(f) Go to [Process] → [Analyze]. Three types of MTLD values are derived on the bottom: MTLD (Raw) is a raw MTLD value described above. MTLD (Znar) and MTLD (Zsci) show z-scores; both are standardized based on norms of texts (either narrative or science) derived from the Touchstone Applied Science Associates (TASA) corpus (P.M. McCarthy, personal communication, November 9, 2011; see McCarthy, 2010, for TASA).

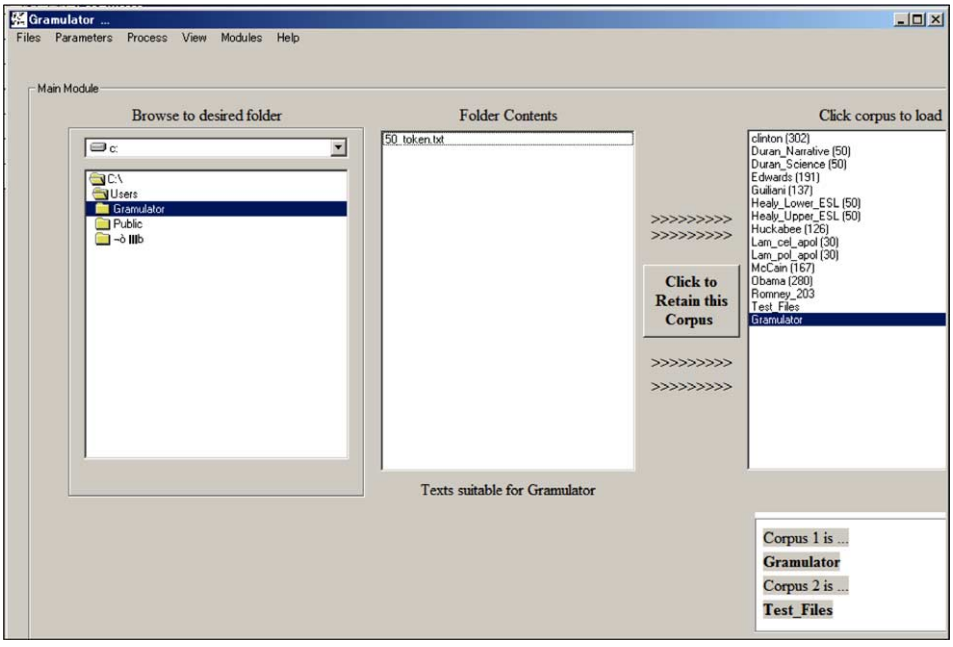

Figure 2. Example of measure of textual lexical diversity (MTLD) calculation using the Gramulator (1).

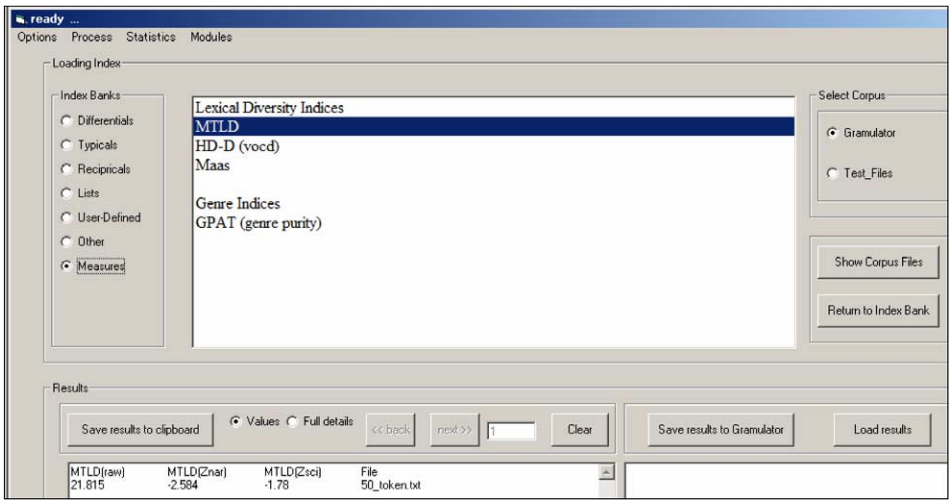

Figure 3. Example of measure of textual lexical diversity (MTLD) calculation using the Gramulator (2).

Guiraud, D, and MTLD were developed to reduce effects of text length. The three measures and TTR are all intended to assess LD, with larger values indicating more lexically diverse texts. Guiraud is a simple transformation of TTR by taking a square root of a denominator and adjusting the values in such a way that long texts are not too disadvantaged.

D and MTLD use computers but differ in principles. D is more firmly based on TTR, although random selection and curve fitting reduce the impact of text length. However, MTLD uses TTR as a cutoff point to inspect the text length for which a speaker/writer can maintain a certain level of LD. Another feature of MTLD is that MTLD "considers LD at the textual level" (McCarthy, 2005, p. 93). It takes textual patterns into account and analyzes all the words in the text from the first word to the last word and from the last to the beginning word (i.e. sequentially). McCarthy (2005) argued that "maintaining the text structure rather than sampling the text provides a more authentic measure of diversity" (p. 88).

3 Method

3.1 Language samples

Language samples from 20 Japanese learners of English at junior and senior high schools were used. A tape-mediated speaking test was employed to elicit utterances on familiar topics (e.g. self-introduction and picture description), with a total of 225 seconds of speech per learner. Of 38 students who produced 200 or more tokens, 20 representative speakers were selected ($M = 236.55$ tokens, $SD = 30.79$, minimum $= 202$, maximum $= 283$). Then, each learner's 200 tokens were separated into a text.

3.2 Analyses

Using the parallel sampling method (e.g. Hess et al., 1986), a text of 200 tokens from one participant was split into 25 segments: four segments of 50 tokens,

three segments of 60 tokens, two segments of 70, 80, 90, and 100 tokens, and one segment of 110, 120, 130, 140, 150, 160, 170, 180, 190, and 200 tokens.

Each segment was analyzed using four LD measures: TTR, Guiraud, D, and MTLD. The count of types and tokens was performed using KWIC Concordance for Windows (Version 5.0; Tsukamoto, 2011). D was computed using D_Tool (Version 2.0; Meara & Miralpeix, 2007). MTLD (raw score) was calculated using Gramulator (Version 5.0; McCarthy, 2011). Base forms and their inflected forms were considered different types in this study; for example, *table* and *tables*, and *play*, *plays*, *played*, *playing* were counted as different.

Subsequently, values from equal-sized segments were averaged. To examine the research question, five token ranges were separately examined using a one-way repeated measures ANOVA and an effect size of partial eta-squared (η_p^2): (a) 50 to 100, (b) 100 to 150, (c) 150 to 200, (d) 100 to 200, and (e) 50 to 200. Statistical assumptions were checked and handled accordingly. The alpha level was set to 0.0025 (0.05/20) because 20 ANOVAs were performed.

4 Results and Discussion

Table 2 shows that there were significant and large effects of text length on TTR, Guiraud, and D when text length changed across 50 to 100 tokens, 100 to 150 tokens, and 50 to 200 tokens (e.g. $\eta_p^2 = 0.75$, 0.49, and 0.85 for TTR; see Appendix for descriptive statistics of each measure for each segment). In the range of 150- to

Table 2. ANOVA Results for Text Length Effects ($N = 20$)

	50–100				100–150			
	F	df	*p*	η_p^2	*F*	df	*p*	η_p^2
TTR	57.54	3.02, 57.43	<0.001	0.75	17.98	2.15, 40.78	<0.001	0.49
Guiraud	81.33	2.95, 56.02	<0.001	0.81	50.68	2.28, 43.36	<0.001	0.73
D	6.49	2.66, 50.49	<0.001	0.25	10.43	2.48, 47.17	<0.001	0.35
MTLD	2.06	2.34, 44.41	0.13	0.10	1.39	3.25, 61.69	0.25	0.07
	150–200				100–200			
	F	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
TTR	16.25	1.59, 30.21	<0.001	0.46	33.92	2.06, 39.09	<0.001	0.64
Guiraud	21.22	1.57, 29.85	<0.001	0.53	57.98	2.11, 40.04	<0.001	0.75
D	1.75	1.62, 30.74	0.19	0.08	7.05	1.91, 36.23	0.003	0.27
MTLD	0.73	2.54, 48.23	0.52	0.04	2.11	2.88, 54.74	0.11	0.10
	50–200							
	F	*df*	*p*	η_p^2				
TTR	111.66	3.12, 59.21	<0.001	0.85				
Guiraud	174.72	2.89, 54.82	<0.001	0.90				
D	19.74	2.73, 51.84	<0.001	0.51				
MTLD	3.90	3.65, 69.39	0.008	0.17				

MTLD, measure of textual lexical diversity; TTR, type-token ratio.

200-token segments and 100- to 200-token segments, text length significantly affected TTR and Guiraud to a substantial degree (e.g. $\eta_p^2 = 0.46$ and 0.64 for TTR).

In contrast, MTLD was not statistically significantly affected by text length, and the degree of impact was generally small for all five ranges ($\eta_p^2 =$ 0.10, 0.07, 0.04, 0.10, and 0.17). Thus, MTLD was found to be the least affected by text length. However, it should be noted that even MTLD was sensitive to text length variations across a range of 50- to 100-token segments, 100- to 200-token segments, and 50- to 200-token segments ($\eta_p^2 =$ 0.10, 0.10, and 0.17).

These results suggest that texts of 100 tokens should be considered the minimum requirement for measuring LD using MTLD. Another finding is that the text length effect was larger when texts have more token differences. In other words, a 150-token difference ($\eta_p^2 = 0.17$) produces larger effects than a 100-token difference ($\eta_p^2 = 0.10$), which has larger effects than a 50-token difference when texts have more than 100 tokens ($\eta_p^2 =$ 0.07, and 0.04). Thus, comparisons should be made between texts of similar lengths (with an interval of 50 tokens, if possible).

The results of the present study generally concur with previous research, with the one major difference that D was affected to an excessive degree across ranges of 50 to 100 tokens, 100 to 150 tokens, 100 to 200 tokens, and 50 to 200 tokens. This shows that D is sensitive to text length when short texts up to 150 tokens are analyzed.

This study shows that text length affects MTLD to a limited degree if the texts analyzed consist of more than 100 tokens. Why does MTLD have such a characteristic in contrast to other LD measures? One reason may be that MTLD focuses on the text length required to arrive at a certain level of TTR, which the other measures do not, and such text length tends to be stable as long as an overall text has sufficient tokens (i.e. more than 100 tokens) and as long as texts are produced by the same person using similar topics. McCarthy (2005) argued that MTLD has high reliability, in the sense that shorter segments of a text generate values similar to those that a whole text produces, and that it has enough sensitivity to distinguish texts of different degrees of LD. The current study supports this argument.

5 Conclusions

This study inspected the impact of text length on four LD measures across 50- to 200-token texts and found that MTLD is least affected by text length, but that even this measure of LD should be used with texts of at least 100 tokens. This result indicates the usefulness of considering LD at the textual level.

For future research, analyses using spoken and written texts of multiple genres, with a larger number of language samples with more tokens, are needed to investigate the generalizability of this study, especially the utility of MTLD.

Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI) of the Ministry of Education, Culture, Sports, Science and Technology in Japan (No. 22720216). I am deeply indebted to Yo In'nami and reviewers for their invaluable comments.

References

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique.* Paris: Presses Universitaires de France.

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

Malvern, D.D., Richards, B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* Hampshire: Palgrave Macmillan. doi: 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)* (Unpublished PhD dissertation). University of Memphis. Retrieved from https://umdrive.memphis.edu/pmmccrth/public/Phil's%20papers.htm?uniq=-xq6brv

McCarthy, P.M. (2010). GPAT: A Genre Purity Assessment Tool. In H.W. Guesgen & C. Murray (Eds.), *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference* (pp. 241–246). Menlo Park, CA: The AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/2010/paper/view/1283

McCarthy, P.M. (2011). Gramulator (Version 5.0) [Computer software]. Retrieved from https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm

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

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–337. doi: 10.1093/llc/15.3.323

Meara, P.M., & Miralpeix, I. (2007). D_Tools (Version 2.0; _lognostics: Tools for vocabulary researchers: Free software from _lognostics) [Computer software]. University of Wales Swansea. Retrieved from http://www.lognostics.co.uk/tools/index.htm

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* New York: Palgrave MacMillan. doi: 10.1057/9780230293977

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. doi: 10.1093/applin/amp047

Tsukamoto, S. (2011). KWIC Concordance for Windows (Version 5.0) [Computer software]. Retrieved from http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html

Note: Subsequent to the writing of this article, Gramulator (Version 5; McCarthy, 2011) is no longer available, as it has been replaced with Version 6. Unfortunately the link to Version 5 is dead, so a link to Version 6 is provided above.

Appendix. Descriptive Statistics of Lexical Diversity Measures ($N = 20$)

	TTR		Guiraud		D		MTLD	
Token	M	SD	M	SD	M	SD	M	SD
50	0.62	0.03	4.33	0.19	26.21	4.38	30.44	5.35
60	0.58	0.03	4.51	0.24	27.95	5.88	29.33	5.10
70	0.57	0.04	4.80	0.34	30.42	7.90	31.38	6.78
80	0.55	0.04	4.91	0.34	29.69	6.05	29.33	4.28
90	0.53	0.03	5.07	0.28	30.84	5.73	28.28	3.67
100	0.53	0.03	5.25	0.32	31.76	6.05	28.77	4.48
110	0.52	0.05	5.44	0.52	33.56	9.41	29.23	6.13
120	0.52	0.05	5.63	0.58	35.05	10.95	28.73	4.94
130	0.50	0.05	5.72	0.53	36.03	10.49	27.97	5.08
140	0.49	0.04	5.84	0.51	36.39	10.00	27.95	5.05
150	0.49	0.04	5.96	0.49	36.31	8.65	27.72	3.90
160	0.48	0.03	6.04	0.45	36.84	7.40	27.33	3.87
170	0.47	0.03	6.10	0.41	37.00	8.03	26.69	3.19
180	0.46	0.03	6.21	0.38	37.45	6.82	26.96	2.69
190	0.46	0.03	6.32	0.38	37.87	7.08	27.25	3.66
200	0.46	0.03	6.40	0.39	38.21	6.97	26.95	3.52