Language Assessment Quarterly

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

An Empirical Evaluation of Lexical Diversity Indices in L2 Korean Writing Assessment

Hakyung Sung, Sooyeon Cho & Kristopher Kyle

To cite this article: Hakyung Sung, Sooyeon Cho & Kristopher Kyle (09 Feb 2024): An Empirical Evaluation of Lexical Diversity Indices in L2 Korean Writing Assessment, Language Assessment Quarterly, DOI: 10.1080/15434303.2024.2311728

To link to this article: https://doi.org/10.1080/15434303.2024.2311728

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2024.2311728

Routledge
Taylor & Francis Group

ARTICLE

An Empirical Evaluation of Lexical Diversity Indices in L2 Korean Writing Assessment

Hakyung Sung[a], Sooyeon Cho[b], and Kristopher Kyle[a]

[a]University of Oregon, Eugene, OR, USA; [b]Zurich University of Applied Sciences, Zurich, Switzerland

ABSTRACT

Lexical diversity (LD) is an important indicator of second language lexical development. Much research has investigated LD indices, with a focus on learners of English. However, further research is needed in languages that are typologically distinct from English, such as Korean. In this study, we evaluated the reliability and validity of LD indices applied to argumentative writing produced by Korean learners. The results indicated that HD-D, MATTR, and MTLD were reliable across different text lengths and were correlated with holistic proficiency scores. However, the meaningful differences were found across Korean-specific tokenization types related to the way morphemes are processed.

INTRODUCTION

A major task of second language (L2) lexical research is to discover what L2 learners know about the lexicon of their L2. Accordingly, a focus of research in this domain has been developing and evaluating methods of measuring productive lexical knowledge. One construct of productive lexical knowledge is breadth (i.e., how many lexical items an L2 users know; e.g., Nation, 1990; Schmitt et al., 2011), which is often measured using indices of lexical diversity (LD) (e.g., Malvern et al., 2004; Thomson & Thompson, 1915). Research on LD indices has primarily focused on aspects of reliability (e.g., text-length independence; Hess et al., 1986; Koizumi, 2012) and validity (relationship with human judgements of proficiency and/or LD; e.g., Kyle et al., 2021; Treffers-Daller et al., 2018). Studies have indicated that a number of proposed LD indices such as Root TTR (Guiraud, 1960) and D (MacWhinney, 2000) intrinsically vary depending on text length (e.g., Malvern et al., 2004), while others such as MATTR (Covington & McFall, 2010) and MTLD (McCarthy & Jarvis, 2010) are relatively stable across text lengths (e.g., Kyle et al., 2023; Zenker & Kyle, 2021). Studies have also indicated that text-length stable indices such as MATTR and MTLD are correlated reasonably well with human judgements of proficiency (e.g., Zenker & Kyle, 2021) and LD specifically (e.g., Kyle et al., 2021). One limitation to the generalizability of extant research on LD is that it has focused on L2 learners of English (e.g., Crossley et al., 2011; Kyle et al., 2021; Zenker & Kyle, 2021). Although some research investigated L2s other than English (e.g., French: Treffers-Daller, 2013; Spanish: Castañeda-Jiménez & Jarvis, 2014), less is known about the reliability and validity of LD indices for L2s that are typologically different from English, such as Korean, which is agglutinative.

CONTACT Hakyung Sung hsung@uoregon.edu Linguistics, University of Oregon, 1600 Millrace Dr., Suite 201, Eugene, OR 97403, USA

This article has been republished with minor changes. These changes does not impact on the academic content of the article.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15434303.2024.2311728

The current study aims to expand LD research to L2 Korean writing assessment by exploring a sizable Korean learner corpus and comparing the performance of different types of Korean-specific tokenizers which parse texts based on *eojeols* (i.e., a sequence of Korean characters separated by white space, consisting of both lexical and functional morphemes) or morphemes (i.e., the smallest meaningful unit). The study will evaluate six LD indices, including Type-token ratio (TTR), log TTR, root TTR, HD-D (the Hypergeometric Distribution Diversity index), MATTR (the Moving-Average TTR), and MTLD (the Measure of Textual Lexical Diversity). To be specific, we focus on evaluating aspects of the reliability and validity of LD indices when applied to L2 Korean argumentative writing. We aim to examine the degree to which indices are stable across text lengths and predictive of Korean holistic proficiency levels.

LITERATURE REVIEW

Evaluating reliability and validity

Reliability

An important issue in the measurement of LD is the relationship between index scores and text length. It can be expected that learners of different proficiency levels will produce texts of different lengths. While text length and LD may be positively correlated because they are both hypothesized to increase with proficiency, LD indices should not have an intrinsic relationship with text length. In other words, differences in LD scores across texts of different lengths should represent differences in the diversity of lexical items used in a text and should not simply be an artifact of text length. Accordingly, much of the English LD research has focused on creating more indices that are reliable (i.e., consistent) across texts of different lengths (e.g., Carroll, 1964; Guiraud, 1960; Hess et al., 1986; Johnson, 1944; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021).

Although other methods have been used (see Fergadiotis et al., 2015; Tweedie & Baayen, 1998), most studies that have evaluated the stability of LD indices across text lengths have used the parallel sampling method (Hess et al., 1986), which involves splitting a text into a series of different text lengths, averaging the LD scores for a particular sample length, and then analyzing the relationship between sample length and LD score. For example, a 200-word text might be divided into four 50-word samples, three 55-word samples, three 60-word samples, and so on. Researchers have applied this method to both first language (L1) (e.g., McCarthy & Jarvis, 2007, 2010) and L2 English texts (e.g., Koizumi & In'nami, 2012; Vidal & Jarvis, 2020; Zenker & Kyle, 2021). For example, Zenker and Kyle (2021) used the method to evaluate LD indices within L2 English written texts with samples of 50–200 words in five-word increments. Results indicated that traditional LD indices such as TTR, log TTR, and root TTR were strongly related to text length. On the other hand, recently revised indices such as MATTR (the Moving-Average TTR; Covington & McFall, 2010), and MTLD (the Measure of Textual Lexical Diversity; McCarthy & Jarvis, 2010) were stable across texts ranging from 50–200 words. It is unclear, however, whether these results are generalizable to agglutinative languages such as Korean.

Validity

One piece of support for an LD index validity argument is an alignment between index scores and human judgements of LD and/or productive proficiency (e.g., writing quality; Chapelle, 1998; Chapelle & Douglas, 2006; Chapelle et al., 2011). Such an alignment provides support for an index's construct validity. If index scores are correlated with human judgements of LD, an argument for construct validity can be made, and without such evidence it is difficult to make such an argument. Accordingly, a body of research has investigated the relationship between the LD indices and human judgments (e.g., Jarvis, 2013, 2017; Kyle et al., 2021). For example, Kyle et al. (2021) examined the correlations between human judgments of LD in L2 argumentative essays and several LD indices (MATTR, MTLD, HD-D), which were previously identified as being resistant to text length effects (Zenker & Kyle, 2021). Notably, the correlations with human judgments of LD were moderate to large: MATTR (r = .492), MTLD (r = .505), and HD-D (r = .602).

When direct judgements of LD are not available, holistic writing scores (e.g., Engber, 1995; Jarvis, 2002; Vidal & Jarvis, 2020) or holistic L2 proficiency level (e.g., Treffers-Daller et al., 2018; Zenker & Kyle, 2021) can be used as an alternative way to assess the construct validity. This assumes that more proficient learners would have access to and produce more sizable and diverse lexicon (Meara, 2005). For instance, Treffers-Daller et al. (2018) used L2 English learners' essays to examine the correlations between the LD indices and (1) the L2 English proficiency level (i.e., Common European Framework of Reference); (2) a vocabulary score; and (3) a writing score by using L2 English learners' essays. The results showed that the indices showed no significant correlation to the L2 proficiency level, but showed small to moderate correlations to the vocabulary/writing scores. Recently, Zenker and Kyle (2021) investigated the validity by comparing the LD indices with the English holistic proficiency level using L2 English argumentative essays and found a statistically significant relationship between LD values and proficiency levels.

To summarize, while the reliability and validity of LD indices are widely recognized in the context of English language assessment, there has been limited exploration of their applicability and validity in other languages. This sets the stage for the next section, which delves into LD indices in the context of Korean language assessment.

Previous Korean LD studies

Some previous studies have used LD indices to analyze written data produced by L2 Korean language learners (e.g., Bai, 2012; Hur & Lee, 2019; Kang, 2018; Lee, 2017). For example, Bai (2012) analyzed six writing samples of L2 Korean learners by calculating TTR and comparing it based on the L1 backgrounds of the learners. The researcher pointed out that while TTR was significantly different across L1 backgrounds, this measurement was flawed due to its negative correlation with text length. Meanwhile, Kang (2018) examined the development of vocabulary production by using 60 text samples produced by Korean language learners whose holistic proficiency ranged from level 1 to 6. The result suggested that the number of tokens and types differed across proficiency levels, with TTR showing a meaningful difference only between the most proficient (level 6) and less proficient learners (levels 1 to 5). Despite these initial efforts, the studies have only used TTR, token counts, or type counts as measures of LD, and most importantly, there is a lack of research examining the reliability and validity of the LD index in the assessment of L2 Korean writing.

OPERATIONAL DEFINITION OF A LEXICAL ITEM

Agglutinative nature of Korean

An important challenge in computing lexical production indices like LD is defining what constitutes a lexical item and coming up with a valid operational definition of it (e.g., Kyle, 2019). Researchers' assumptions about a learner's knowledge can be influenced by the chosen operational definition of a lexical item. For example, in past English LD studies, lexical items were typically defined as words, typically delimited by white space and word boundary punctuation marks such as periods and commas. However, the definition of a tokenized lexical item can vary, ranging from orthographic word forms to lemmas, flemmas, or word families (Jarvis & Hashimoto, 2021; Kyle, 2019). For instance, if word family (i.e., a base word form and all its derived and inflected forms; Bauer & Nation, 1993, p. 253) has been chosen as an operational definition of a lexical item, the researcher assumes that learners can use all derived and inflected forms of a base word. Accordingly, when we calculate an LD index, we consider the derived, inflected, and base forms to be equal, counting them as one type. On the other hand, if lemma (i.e., a set of word forms comprising all inflected variants of a word; Leech et al., 2001, pp. 4–5) has been chosen, then the researcher assumes that learners have knowledge of word inflection systems. Consequently, the inflected and the base forms of a word are treated equally and count as a single type.

In examining Korean, it is important to consider the unique structure and characteristics of lexical items, especially with respect to morpheme parsing at the eojeol (a token bounded by white space) level. While English is an isolating language with minimal inflection (Haspelmath & Sims, 2013), Korean takes on the properties of an agglutinative language. This agglutination introduce ambiguity when determining the optimal word-token boundary (Chung & Gildea, 2009; Haspelmath & Michaelis, 2017).

Delving deeper into the structure of the Korean eojeol, it often consists of multiple morphemes. These not only include a lexical morpheme but also various functional morphemes. Examples of these functional morphemes are particles, which include case markers and postpositions, as well as sentence pre-final and final endings. Particles play an instrumental role of showing grammatical relationships between nouns and other elements in a sentence (e.g., subject, object, or oblique roles). Furthermore, the structure of Korean predicates offers insight into the language's complexity. These predicates typically consist of a root, which serves as the central semantic unit for both adjective and verbs and presents in a base form or as a stem. After the stem, sentence pre-final endings are introduced to indicate tense or aspect. These are then followed by sentence final endings, which provide cues for whether the sentence is declarative, imperative, or interrogative. Connective markers sometimes occupy the positions for sentence final endings, reflecting their role in connecting clauses.

Taken together, two main operational definitions can be considered when choosing a lexical item for L2 Korean production research. The first involves choosing a lexical item based on the eojeol, which tokenizes a text in the same way as the previous English LD studies (i.e., separation by a white space unit). This assumes that correctly combining lexical and functional morphemes within an eojeol indicates learners' lexical knowledge, given that the adept combination of morphemes (and using various inflected and derived forms of the stem) is important for their lexical knowledge. The second definition involves

choosing a lexical item based on the morpheme unit, which is to tokenize each eojeol based on morpheme boundaries. This approach treats each morpheme (whether it is lexical or functional) as a unique token, rather than considering the combination of morphemes within an eojeol. This option can be further refined into several fine-grained options, depending on which type of functional morphemes (such as case markers, tense markers, or sentence final endings) are considered individual lexical items. For example, researchers might decide to operationalize functional morphemes related to case markers as separate lexical items, while considering tense markers and sentence final endings as a single lexical item when combined correctly (i.e., consider every inflected form of a predicate to be a unique token type). From a methodological standpoint, this requires determining the most suitable Korean tokenizer for text processing before conducting LD calculations.

CURRENT STUDY

The current study evaluates aspects of the reliability and validity of six LD indices (TTR, log TTR, root TTR, HD-D, MATTR, and MTLD) in L2 Korean written text samples. In evaluating these indices, we will compare four different types of Korean tokenizers based on the assumption that the different methods of operationalizing a lexical item would influence the calculation of LD, as well as reliability and validity evaluations. Our research questions (RQs) are as follows:

RQ 1. What is the relationship between LD indices and text length?

RQ 2. To what degree are text-length stable LD indices predictive of Korean proficiency levels?

METHODOLOGY

Learner corpus

In this study, data samples were sourced from the L2 Korean learner corpus collected and maintained by the National Institute of Korean Language (NIKL).[1] The corpus includes written and spoken data from Korean learners between 2015 and 2020, with a total of 27,299 written and 2,541 spoken texts. We focus on 4,208 argumentative essays drawn from the written section of the corpus, which range in length from 54 to 295 eojeols. Each essay addresses a prompt that asks test takers to express their opinion on social issues such as school violence or environmental protection. The most common L1 backgrounds among the participants were Chinese and Japanese, followed by Vietnamese, Cantonese, and English. Participants' Korean proficiency was assessed based on their performance across three distinct sections: listening, reading, and writing, as outlined in the Test of Proficiency in Korean (TOPIK, n.d..) for levels 3 to 6 (see Y. Won, 2016 for a detailed overview of TOPIK's structure and evaluation). Table 1 shows the distribution of learners across

[1]The National Institute of Korean Language (NIKL) is a government institution established with the aim of developing the Korean language and enhancing its usage in daily life. Its primary objectives include conducting research on language policies, managing linguistic data, and publishing Korean dictionaries.

Table 1. The number of learners by proficiency level and mean (sd) of eojeol count per text.

Level	The number of learners	Mean (sd) of eojeol per text
3	296	139.22 (30.09)
4	1,167	135.81 (31.71)
5	1,645	146.79 (25.22)
6	1,110	151.95 (36.14)
Total	4,208	

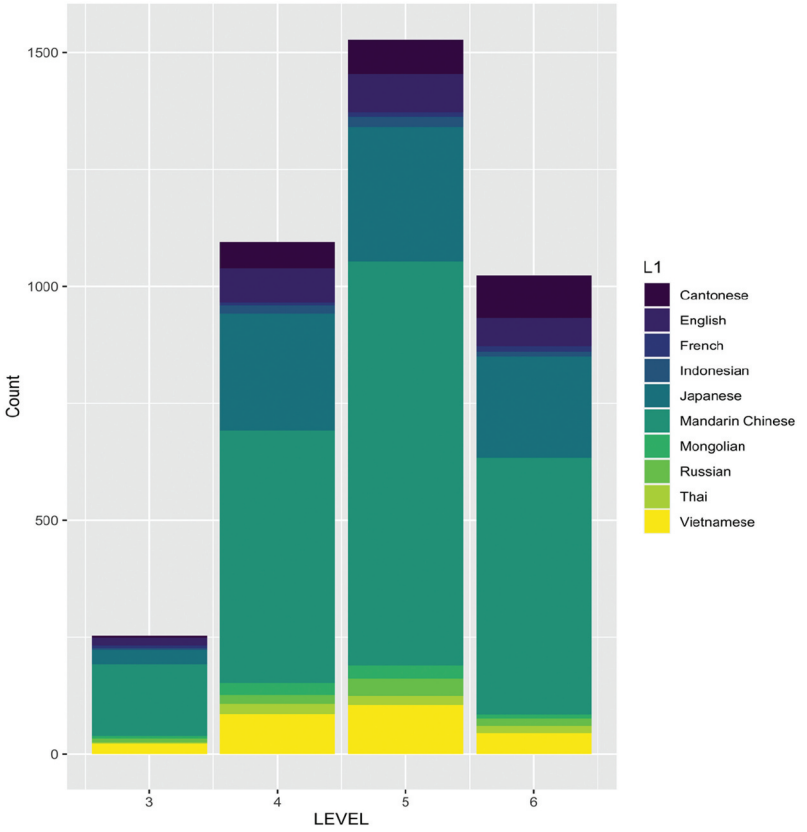

Figure 1. An overview of the top ten L1s of the learners included in the study.

proficiency levels 3 to 6, which were the focus of this study. For each proficiency level, the table details the total number of learners and provides the mean (along with the standard deviation) of eojeols per text produced by individual learners. Figure 1 shows the distribution of the ten most frequently spoken L1 backgrounds of the Korean learners in this study (and the full information on the participants can be found in the Appendix).

The corpus metadata does not specify whether the essays were collected from in-class tasks or were completed as homework with potential outside support (Shin & Jung, 2021). To address this, two researchers with a Korean L1 background manually reviewed the essays. Texts were excluded if both agreed that the content appeared to reference external resources, especially if they cited detailed statistics or other numerical details.

Data cleaning and tokenization

A Python script automatically removed misspelled words from the corpus using *win32com.client* package (pywin32; Mhammond, n.d..). On average, 5.19 eojeols were removed from each text, with a standard deviation of 3.67. Proportionally, 3.66% eojeols were removed from the original text, with a standard deviation of 2.57%. Next, tokenization was performed using tokenizers from *Stanza* (Qi et al., 2020) and *KoNLPy* (Park, 2018). Stanza was used to tokenize the texts into an eojeol unit, which separates tokens based on white space. This approach mirrors preprocessing methods used in English LD studies and serves as our study's baseline. KoNLPy is an open-source Python package for Korean natural language processing. It contains five different morpheme-based tokenizers (i.e., morpheme analyzers): Hannanum, Komoran, Kkma, Mecab, and Okt (in alphabetical order). The package suggests choosing a tokenizer based on factors such as execution and loading time, the algorithm for token segmentation, the relationship between a word and surrounding words, and the L1 Korean dictionaries that each tokenizer has trained on (Park, 2018). The most important difference between the tokenizers is the specificity of morpheme tokenization and the resulting part-of-speech tags assigned (H. Won et al., 2020). Table 2 illustrates the differences in tokenization with an example sentence '나는 *na-nun* (1SINGULAR-TOPIC) 읽었던 *ilk-ess-ten* (read-PAST-RELATIVE) 책을 *chayk-ul* (book-ACCUSATIVE) 또 *tto* (again) 읽었다 *ilk-ess-ta* (read-PAST-DECLARATIVE)', which translates to "I read [the book I had read] again" in English. Okt tokenizes functional morphemes related to particles, but not tense markers or sentence final suffixes. Mecab sub-tokenizes stems, but not functional morphemes in the predicates such as sentence pre-final/final endings. Kkma has the most fine-grained tokenization scheme as it splits the predicate pre-final and final endings. The example output from Kkma demonstrates that some of the morphemes (especially functional morphemes) can be split into even smaller unit than a syllable, even down to a consonant level. Note that Hannanum and Komoran were excluded from subsequent analyses. While they use the same tokenization rule as Kkma, Kkma outperformed Hannanum and Komoran in a small-scale accuracy evaluation, specifically in terms of segmenting eojeols into morphemes.

As demonstrated by the example above, the count of tokens and types can vary depending on the operationalization of functional morphemes (e.g., case markers, predicate endings) and the choice of tokenizer. For instance, if Okt is selected as the tokenizer, it considers each inflected form of a verb as a separate word. This distinction is evident in the example where '읽었던' and '읽었다' share the same stem '읽-', "read", but Okt treats them as separate lexical items due to the different functional morphemes: '었-던' indicates past tense and a sense of relative clause (to modify "the book"); '었-다' indicates past tense

Table 2. Comparison of different types of Korean tokenizers.

Case marker	Predicate stem	Predicate (pre-) final endings	Example	*n*tokens	*n*types	Tokenizer
X	X	X	나는, 읽었던, 책을, 또, 읽었다	5	5	Stanza
O	X	X	나, 는, 읽었던, 책, 을, 또, 읽었다	7	7	Okt
O	O	X	나, 는, 읽, 었던, 책, 을, 또, 읽, 었다	9	8	Mecab
O	O	O	나, 는, 읽, 었, 더, ㄴ, 책, 을, 또, 읽, 었, 다	12	10	Kkma

and a declarative statement. On the other hand, if we choose Kkma, it counts every inflectional morpheme as a separate lexical item, but at the same time, does not count the same stem repeatedly even though it is inflected in different ways. To illustrate this difference in tokenization, consider the calculation of TTR for each tokenizer option: TTR = 7/7 = 1.00 for Okt, but TTR = 8/9 = 0.89 for Mecab and TTR = 10/12 = 0.83 for Kkma (rounded to two decimal places).

LD indices

After tokenizing the text samples, LD indices were calculated using TAALED (version 0.32) for Python. The package offers various LD indices that have been tested and validated for English LD analyses. In this study, we used six LD indices based on the previous studies: simple TTR (TTR), Log TTR (LTTR), Root TTR (RTTR), HD-D, MATTR, and MTLD (For a more detailed descriptions of each index, see Kyle et al., 2021).

1. *TTR*:

$$nTypes/nTokens$$

Type-token ratio (TTR) is calculated by dividing the number of unique tokens (i.e., types) in a text by the total number of tokens (Johnson, 1944).

2. *Log* TTR:

$$\log(nTypes)/\log(nTokens)$$

Log TTR (LTTR) is calculated by the logarithm of the number of types divided by the logarithm of the number of tokens (Herdan, 1960).

3. *Root TTR*:

$$nTypes/\sqrt{nTokens}$$

Root TTR (RTTR), also known as *Guiraud's index*, is calculated by dividing the number of types by the square root of the total tokens (Guiraud, 1960).

4. *HD-D*:

The Hypergeometric Distribution Diversity (HD-D) index uses the hypergeometric distribution to calculate the probability of running across one of its tokens in a randomly selected sample of 42 tokens. These probabilities are then added to calculate the overall value for the entire text (McCarthy & Jarvis, 2007).

5. *MATTR*:

The Moving-Average Type-Token Ratio (MATTR) calculates the moving-average TTR after segmenting the text into specified lengths. Following previous studies, we segmented the text into lengths of 50 tokens, and then computed the TTR for segments 1–50, 2–51, 3–52, and so forth. Calculated TTR values are averaged to produce the final value (Covington & McFall, 2010).

6. *MTLD*:

The Measure of Textual Lexical Diversity (MTLD) measures the average number of tokens required to achieve a pre-designated TTR value. We followed the previous

studies and set the TTR value as .72. The number of tokens that maintain the TTR values is called a factor. Factors are computed in one direction from the beginning to the end of the given text, and the tokens are not overlapped while segmenting the text. Tokens not included in a factor are termed a "partial factor", and they adjust the final value (McCarthy & Jarvis, 2010).

PRELIMINARY ANALYSIS: EVALUATING THE PERFORMANCE OF KOREAN TOKENIZERS

To analyze LD in Korean, it is crucial to assess the performance of Korean tokenizers in accurately parsing and tokenizing morphemes. Recently, Sung and Shin (2023) evaluated four Korean morphological analyzers on a sizable L2 Korean corpus; however, no studies have directly compared the morpheme parsing and tagging accuracy of the Korean tokenizers used in this study. To address this research gap, we calculated precision, recall, and F1 scores for each tokenizer using a sample of 100 randomly selected sentences from the NIKL corpus.[2] Two researchers, both native Korean speakers, hand-annotated the sentences based on the tokenization rules of each tokenizer. Each sentence was independently annotated, and the results were cross validated. Any disagreements were adjudicated as we reviewed each case. Calculating the reliability between two annotators before adjudication was skipped due to the small sample size. Tokenization accuracy was calculated by comparing the tokens hand-annotated by the researchers with those automatically parsed by the tokenizers.

The results of this analysis indicated that all four tokenizers achieved relatively high tokenization accuracy with F1 scores ranging from 0.905 to 0.967 (as shown in Table 3). The results showed that Stanza was the most accurate, followed by Mecab, Okt, and Kkma. As all tokenizers showed F1 scores above 0.9, we concluded that they are suitable for use in our LD analysis.

ANALYSES

To address RQ 1—determining the LD indices that are reasonably independent of the text length effects, we carried out a parallel sampling approach. Drawing on previous studies that employed a parallel sampling approach (Hess et al., 1986; Koizumi, 2012; Zenker & Kyle, 2021), we set a cut-off point at 200 tokens. According to these studies, 200 tokens provide an adequate sample size for investigating the relationship between text length and LD measures. Subsequently, the texts were divided into segments with lengths ranging from

Table 3. Comparative evaluations of Korean tokenizers.

	Predictive results		
Tokenizer	Precision	Recall	F1 score
Stanza	0.978	0.959	0.967
Okt	0.915	0.900	0.907
Mecab	0.944	0.929	0.936
Kkma	0.918	0.893	0.905

[2]Precision measures how many of the parsed tokens were correctly identified. Recall indicates how many of the actual tokens were correctly identified and parsed by the tokenizer. To encapsulate these two metrics into a single performance measure, we calculated the F1 score using the formula: 2 * (precision * recall)/(precision + recall).

50 to 200 tokens, incrementing by 5 tokens. Essays with fewer than 200 tokens were excluded from analysis. For each essay, index values were calculated for 31 different text segment lengths (50-55-60-...-200). This type of analysis has been used in previous studies to evaluate indices in English L2 written and spoken corpora (e.g., Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007; Treffers-Daller, 2013; Vidal & Jarvis, 2020). We conducted correlation analyses between the segmented text length and the value of each index, and visualizations were also produced. The stabilization point was determined by evaluating the correlation values, following the guidelines by Cohen (2013), who suggested that *r*-values below .100 represent a small effect (Zenker & Kyle, 2021).

To answer RQ 2—validating the text-length stable indices in terms of predicting holistic Korean proficiency, we started from descriptive statistics using violin plots. Then, we performed ordinal regression analysis between the LD values and the proficiency levels. The statistical analysis was carried out using R software (R Core Team, 2021), and the ordinal logistic regressions were computed using the MASS package (Ripley et al., 2013). Odds ratios and effect sizes were calculated using the RMS package (Harrell et al., 2017). The full data and code for the analysis can be found in the online supplementary material (https://osf.io/mz39f/?view_only=abc0cfc0254f432b842ea3b63fc2c7cb).

RESULTS

RQ 1

To examine the correlation between LD values and text length and assess the reliability of the indices, we conducted parallel sampling and correlation analyses. Descriptive statistics for the number of analyzed essays, the mean, and standard deviation of each index from the parallel analysis are presented in Table 4. It is important to note that the different tokenizers adopt different operational definitions of a lexical item and essays shorter than 200 tokens were excluded during the parallel analysis, thus resulting in variations in the number of essays included in each tokenizer option. With Stanza, we adjusted the minimum tokens to be 125, encompassing segments ranging from 50 to 125 tokens. This decision was made because Stanza's eojeol-based tokenization approach led to a significant exclusion of texts when the minimum token threshold was set at 200.

We employed mean z-scores for the LD values obtained by various measures, Pearson correlation values, and bin analysis to analyze the results. To facilitate comparison of the tokenizers and indices, z-scores were used as a common scale, following a previous English LD study on text length (Zenker & Kyle, 2021). Figure 2 depicts a solid blue line indicating the best-fit correlation for the entire sample, with closer proximity to zero and horizontality

Table 4. Descriptive statistics for the parallel analysis.

	*n*essays	*n*types	TTR	LTTR	RTTR	HD-D	MATTR	MTLD
Stanza	2662	71.78	0.83	0.96	7.64	0.90	0.88	164.77
		(17.62)	(0.07)	(0.02)	(0.98)	(0.04)	(0.05)	(85.86)
Okt	2124	78.72	0.65	0.91	7.04	0.82	0.79	77.04
		(22.94)	(0.08)	(0.02)	(0.93)	(0.04)	(0.05)	(24.05)
Mecab	3394	70.64	0.59	0.89	6.34	0.78	0.75	59.32
		(19.36)	(0.09)	(0.03)	(0.76)	(0.04)	(0.05)	(14.75)
Kkma	3809	68.45	0.57	0.88	6.15	0.76	0.73	54.89
		(18.37)	(0.09)	(0.03)	(0.70)	(0.04)	(0.05)	(12.63)

The numbers in the parenthesis are standard deviation.

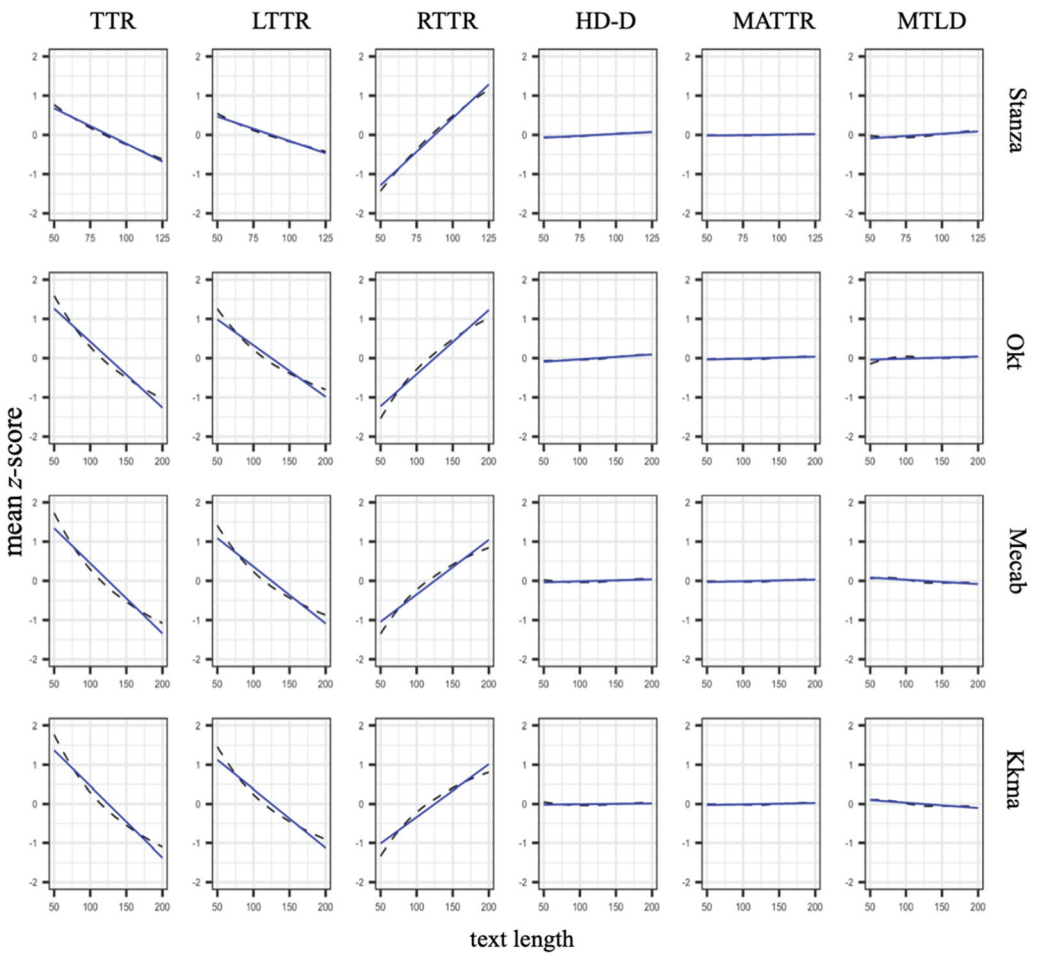

Figure 2. Relationship between LD values and text length.

Table 5. Correlation *r*-values across all text lengths.

	TTR	LTTR	RTTR	HD-D	MATTR	MTLD
Stanza	−.417	−.287	.790	**.045**	**.012**	**.053**
Okt	−.752	−.587	.732	**.055**	**.022**	**.024**
Mecab	−.780	−.648	.622	**.023**	**.019**	**−.050**
Kkma	−.818	−.673	.605	**.009**	**.016**	**−.061**

Bold numbers indicate absolute *r*-values below .100.

indicating greater stability in the LD index. A dashed black line refers to the Loess line (i.e., a smooth curve that fitted to the data points). Table 5 displays the correlation *r*-values between all text lengths and the indices.

In the following step, we adopted a binning approach for a more detailed examination of the correlations between the LD values and text length, which specifically aims to investigate the LD indices that showed minimal correlations with text length in the previous step (i.e., HD-D, MATTR, MTLD). Following previous studies (Koizumi, 2012; Koizumi & In'nami, 2012; Zenker & Kyle, 2021), data samples were grouped into three bins: Okt, Mecab, & Kkma – bin

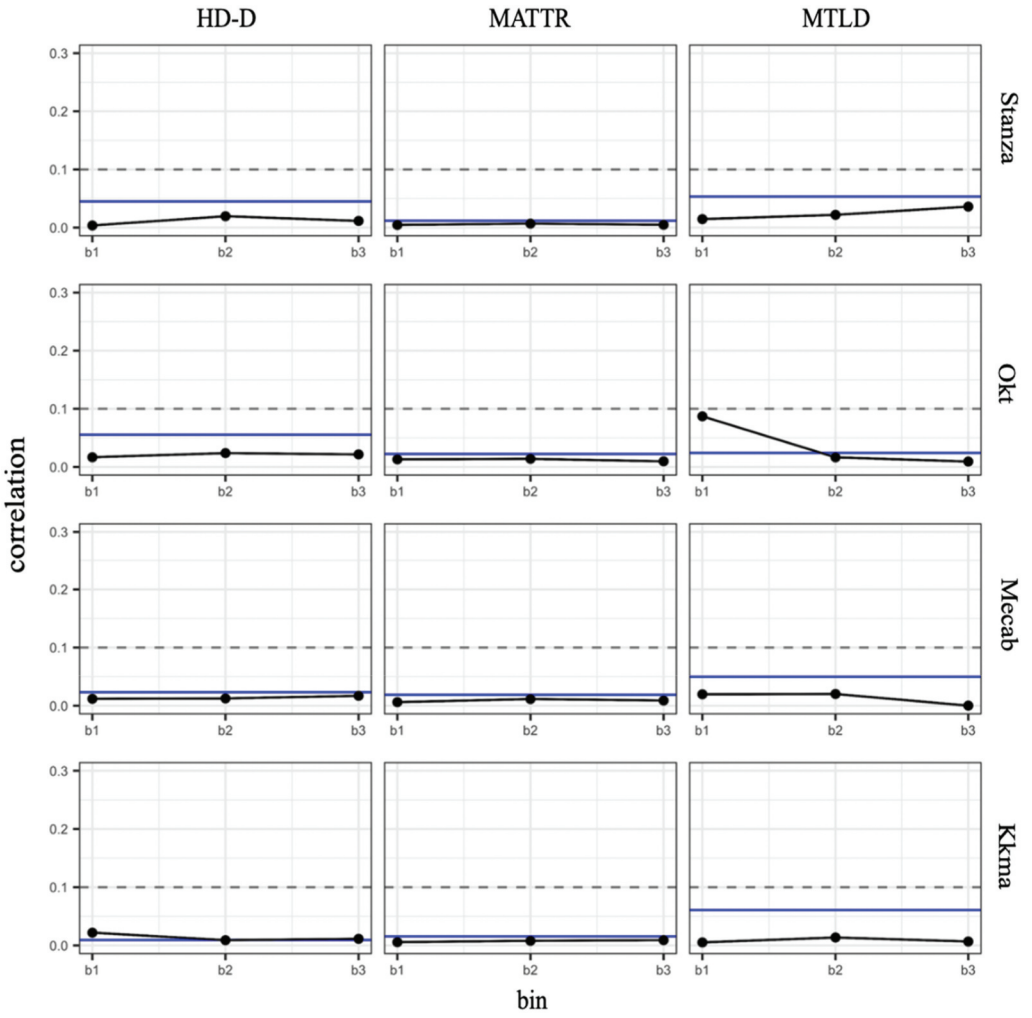

Figure 3. Relationship between LD values and text length bins.

1: 50–95 tokens; bin 2: 100–145 tokens; bin 3: 150–195 tokens; Stanza – bin 1: 50–70 tokens; bin 2: 75–95 tokens; bin 3: 100–120 tokens. This process examines the relationship between the LD values and text lengths within smaller sections. Pearson correlations in each bin were calculated. Figure 3 displays the results, with Table 6 providing the correlations between LD indices and text length across bins and tokenizers. The solid blue line in the graph indicates the overall correlation between the index and text length for texts ranging from 50–200 tokens (Stanza: 50–125 tokens), while the dashed grey line shows the threshold for a small correlation ($r = .100$). The black dots indicate the correlation between LD scores and text length within each text bin.

RQ 1 interim discussion

The results indicate that HD-D, MATTR, and MTLD are stable across the text lengths. Conversely, TTR, LTTR, RTTR showed strong correlations with text length, primarily with large Pearson correlation values (i.e., absolute r-values > .600; Cohen, 2013). The result for

Table 6. Correlations between LD indices and text length within bins.

	Bin	HD-D	MATTR	MTLD
Stanza	1	.003	.005	.015
	2	.020	.007	.022
	3	.011	.005	.036
Okt	1	.017	.013	.087
	2	.024	.013	.016
	3	.022	.010	.010
Mecab	1	.012	.006	.020
	2	.013	.011	.020
	3	.017	.009	.000
Kkma	1	.022	.006	.005
	2	.009	.008	.014
	3	.011	.009	.007

the following bin analysis shows that HD-D, MATTR, and MTLD remain below a correlation of $r = .100$ across all three bins, confirming their reliability in the presence of text length effects. No noticeable differences were observed in the reliability of LD values among the different types of tokenizers, even though each tokenizer produced varying LD values for different text lengths and bins. This suggests that the operational definition of a lexical item does not significantly impact the reliability of the LD indices.

RQ 2

To answer the question on the validity of LD indices, we conducted ordinal logistic regression analyses between text-length stable LD (HD-D, MATTR, and MTLD) indices and holistic language proficiency levels. Table 7 provides descriptive statistics, including the number of tokens and types, for the text-length reliable indices. Before conducting the regression analysis, we performed a visual inspection of the data using violin plots to examine the distribution of each LD value at every level as in Figure 4. The result shows that there was no difference in LD values across proficiency levels when tokenized by Stanza. Hence, the regression analysis was carried out using the remaining three tokenizers – Okt, Mecab, and Kkma, as shown in Table 8.

Table 7. Descriptive statistics.

		*n*tokens	*n*types	HD-D	MATTR	MTLD
Stanza	mean	138.57	105.58	0.90	0.88	171.79
	(sd)	(30.97)	(23.24)	(0.04)	(0.05)	(70.68)
	min	54	43	0.67	0.64	28.99
	max	295	230	0.98	1.00	700.00
Okt	mean	205.81	113.50	0.82	0.79	76.70
	(sd)	(47.51)	(22.98)	(0.04)	(0.04)	(22.74)
	min	69	48	0.67	0.61	28.80
	max	441	230	0.91	0.91	196.16
Mecab	mean	250.88	112.55	0.78	0.75	58.01
	(sd)	(56.79)	(21.94)	(0.03)	(0.04)	(12.98)
	min	91	52	0.64	0.58	25.59
	max	551	220	0.87	0.86	136.55
Kkma	mean	270.58	113.28	0.76	0.74	53.14
	(sd)	(60.67)	(21.45)	(0.03)	(0.04)	(10.57)
	min	94	51	0.63	0.58	25.50
	max	585	216	0.85	0.84	107.53

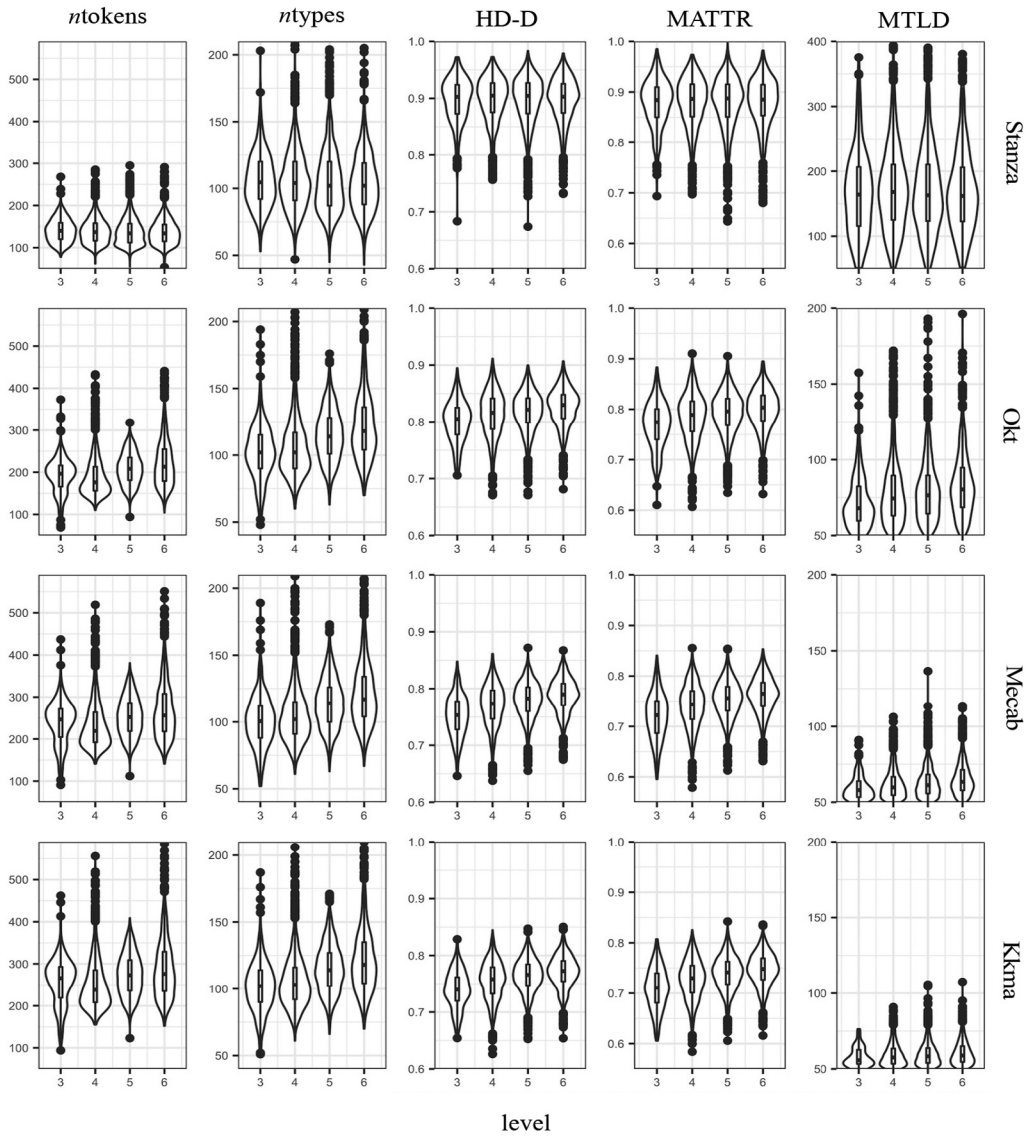

Figure 4. Violin plots comparing correlations between holistic proficiency level and LD values.

Considering that differences across various LD indices were not the focal point of this analysis, we carried out separate ordinal logistic regression analyses for each LD index (HD-D, MATTR, MTLD) against the proficiency levels. In other words, each LD index was examined in relation to the proficiency levels in isolation. Because differences across the different indices were not relevant to this analysis, a single LD index was compared to the levels one by one. Finally, based on the ordinal regression analysis, we compared the effect sizes and odds ratios of each index and tokenizer pair, as demonstrated in Table 9. Note that all the LD measures were computed over full texts.

Table 8. Ordinal logistic regression.

			Estimate	SE	*t*-value	*p*-value
Okt	HD-D	coefficients	8.966	0.788	11.381	<0.001*
		level = 3|4	4.690	0.641	7.317	<0.001*
		level = 4|5	6.690	0.643	10.400	<0.001*
		level = 5|6	8.393	0.648	12.952	<0.001*
	MATTR	coefficients	8.088	0.668	12.113	<0.001*
		level = 3|4	3.749	0.525	7.140	<0.001*
		level = 4|5	5.756	0.527	10.913	<0.001*
		level = 5|6	7.465	0.533	14.017	<0.001*
	MTLD	coefficients	0.015	0.001	11.184	<0.001*
		level = 3|4	−1.527	0.112	−13.666	<0.001*
		level = 4|5	0.469	0.103	4.569	<0.001*
		level = 5|6	2.172	0.108	20.126	<0.001*
		level = 5|6	4.263	0.170	25.059	<0.001*
Mecab	HD-D	coefficients	14.932	0.866	17.241	<0.001*
		level = 3|4	8.908	0.666	13.367	<0.001*
		level = 4|5	10.963	0.672	16.308	<0.001*
		level = 5|6	12.714	0.679	18.717	<0.001*
	MATTR	coefficients	12.880	0.734	17.543	<0.001*
		level = 3|4	6.949	0.544	12.779	<0.001*
		level = 4|5	9.010	0.550	16.397	<0.001*
		level = 5|6	10.766	0.557	19.335	<0.001*
	MTLD	coefficients	0.041	0.002	17.356	<0.001*
		level = 3|4	−0.348	0.140	−2.476	0.013
		level = 4|5	1.698	0.137	12.414	<0.001*
		level = 5|6	3.458	0.145	23.885	<0.001*
Kkma	HD-D	coefficients	14.812	0.925	16.005	<0.001*
		level = 3|4	8.595	0.699	12.302	<0.001*
		level = 4|5	10.636	0.704	15.114	<0.001*
		level = 5|6	12.376	0.710	17.423	<0.001*
	MATTR	coefficients	13.285	0.789	16.844	<0.001*
		level = 3|4	7.071	0.574	12.326	<0.001*
		level = 4|5	9.125	0.579	15.758	<0.001*
		level = 5|6	10.874	0.586	18.552	<0.001*
	MTLD	coefficients	0.047	0.003	16.691	<0.001*
		level = 3|4	−0.183	0.154	−1.184	0.237
		level = 4|5	1.859	0.151	12.272	<0.001*
		level = 5|6	3.611	0.159	22.694	<0.001*

Table 9. Effect size and odds ratio.

		Effect size	S.E.	Lower .95	Upper 0.95	Odds ratio
Okt	HD-D	0.420	0.037	0.348	0.492	1.522
	MATTR	0.451	0.037	0.378	0.524	1.570
	MTLD	0.410	0.037	0.338	0.482	1.507
Mecab	HD-D	0.669	0.039	0.593	0.745	1.952
	MATTR	0.677	0.039	0.601	0.753	1.968
	MTLD	0.668	0.038	0.592	0.743	1.950
Kkma	HD-D	0.608	0.038	0.534	0.683	1.837
	MATTR	0.650	0.039	0.575	0.726	1.916
	MTLD	0.657	0.039	0.580	0.733	1.928

RQ 2 interim discussion

In contrast to the findings from RQ1, the results first show that the validity of LD indices is influenced by way of tokenizing texts. When the text samples were tokenized using the eojeol-unit (Stanza), the text-length stable LD values showed no correlation with the

proficiency levels. Considering the agglutinative nature of the Korean language, our findings suggest that a more valid approach to derive LD values may be to count lexical items based on a morpheme unit.

We hypothesized a positive correlation between the number of lexical items an L2 user knows and their holistic proficiency levels, based on previous L2 studies (e.g., Bulté & Roothooft, 2020; Vidal & Jarvis, 2020). These studies suggest that as language learners become more proficient, their productive vocabulary expands, leading to a greater variety of lexical items used. To explore this hypothesis further, we conducted an ordinal logistic regression analysis. The results indicated that, overall, text-length stable indices were significant predictors of holistic proficiency levels. However, exceptions were found between levels 3 and 4 when the texts were tokenized using Mecab and Kkma, and when LD was calculated using MTLD. This implies that LD values measured by MTLD may be less valid among the lower-level groups of L2 Korean learners. Because the ordinal logistic regression model assumes that the relationships between the predictors (i.e., LD values) and each level of holistic proficiency are proportional, it can be interpreted that a one-level proficiency increase in HD-D is expected to result in a 14.932 increase on the log odds scale.[3] When the models were converted into odds ratios, the results confirmed that the text-length stable indices had stronger predictions when tokenized by Mecab or Kkma than Okt.

DISCUSSION

In this study, we aimed to examine the reliability and validity of the LD indices in a large corpus of L2 Korean written texts. Six different LD indices were included in the analysis: TTR, LTTR, RTTR, HD-D, MATTR, and MTLD. To account for the agglutinative nature of Korean, we used four different types of tokenizers to define a lexical item. Specifically, these tokenizers were based on four methods: (1) recognizing an eojeol (separated by white space) as in Stanza, (2) parsing only particles as in Okt, (3) parsing a predicate stem as in Mecab, and (4) parsing every functional morpheme in a predicate as in Kkma. As a preliminary step, we evaluated the accuracy of each tokenizer by examining the F1 scores of 100 hand-annotated sample sentences compared to tokens automatically tagged by each tokenizer. The results showed that all F1 scores were above 0.9, and thus, the four tokenizers were included for the subsequent analyses.

To determine the independence of LD indices from text length, we used a parallel sampling approach. The results demonstrated that TTR, RTTR, and LTTR were highly related to text length, while HD-D, MATTR, and MTLD were relatively stable. Even though the stable LD indices have been examined in several empirical studies in relation to L1 and L2 English texts (Koizumi, 2012; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007, 2010; Zenker & Kyle, 2021), the findings of this study suggest that HD-D, MATTR, MTLD could also be employed in L2 Korean writing assessment as reliable LD measures. In addition, the reliabilities of the indices were not affected by the way we operationalized the lexical items (i.e., type of tokenizer).

[3]An odds ratio shows the change in the odds of an outcome with one-unit (i.e., proficiency level) in the predictor, assuming all other factors remain unchanged.

Next, we explored the degree to which the text-length stable LD indices predicted L2 holistic proficiency. The results indicated that the choice of tokenization significantly impacted the effect sizes. Both Mecab and Kkma exhibited large effect sizes, with Kkma being slightly smaller. In contrast, when the texts were tokenized using Stanza, there was little to no relationship between proficiency levels and LD values. This suggests that when a lexical item is calculated based on an eojeol unit (the combined form of lexical and functional morphemes), the resulting value may overestimate the diversity of the lexical items used by Korean learners. This may especially be true for low-level learners who tend to use the same lexical morpheme repeatedly but combine it with different functional morphemes (e.g., text sample_510 (level 3) – the word "happiness" is used in various forms: 행복 *hayng-pok* "happiness", 행복할 *hayng-pok-hal* "will be happy", 행복하다 *hayng-pok-ha-ta* "be happy", '행복하려면' *hayng-pok-ha-lye-myen* "in order to be happy"). These findings are consistent with previous research suggesting that the definition of a lexical item could be language-specific and that written units separated by white space may not reflect important lexical units in agglutinative languages like Korean (Chung & Gildea, 2009; Haspelmath & Michaelis, 2017).

LIMITATIONS AND FUTURE DIRECTIONS

Although we identified potential applications of LD indices in L2 Korean writing assessment, the processing of tokenizers and analysis of Korean learner corpora warrant further investigation. First, despite our efforts to remove misspelled words by using an automated Python package, we need to be more confident in the accuracy of the automatic spelling checker, particularly with word-spacing errors. Such errors can create ambiguities in the lexical interpretation (Kwon et al., 2004), potentially impacting the tokenizer's performance and, consequently, LD calculations. Future studies should employ more sophisticated Korean spelling checkers to enhance reliability in the text preprocessing stage. Second, this study focused solely on one type of writing task (argumentative writing). Because previous research suggests that task type may affect LD scores (e.g., Alexopoulou et al., 2017), future research should investigate LD scores across different writing tasks (e.g., expository writing, narrative writing, descriptive writing, etc.). Third, our calculation of LD indices incorporated both lexical and functional morphemes. Future research might benefit from analyzing LD indices by considering these two types of morphemes separately. This separation could provide a more nuanced understanding of the usage patterns at various proficiency levels and offer insights into whether proficiency differences stem primarily from the usage of lexical and/or functional morphemes. Fourth, while we investigated the validity of the indices by using holistic Korean proficiency measures, this method is still indirect. It would be worthwhile for future research to directly validate these indices by comparing them to native Korean speaker judgements of LD. Lastly, considering the limited scope of this study, a key future direction is to broaden the investigation into how well the LD indices developed for English (and possibly other Indo-European languages) function with typologically different languages, moving beyond Korean. A valuable extension of our findings could examine potential variations based on the L1's typological background. For instance, an intriguing comparison might involve contrasting L2 Korean

speakers from an English L1 background (which is typologically different from Korean) with L2 Korean speakers from a Turkish L1 background (more similar typologically to Korean). Both approaches could enrich our comprehension of lexical diversity in diverse L2 contexts.

Disclosure statement

We used the NIKL dataset with permission required. We strictly followed all stipulated guidelines to respect the interests of the data providers. No potential conflict of interest was reported by the author(s).

ORCID

Hakyung Sung http://orcid.org/0000-0002-5860-7353
Kristopher Kyle http://orcid.org/0000-0001-5415-9672

REFERENCES

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Bai, D.-Y. (2012). *Hankwuke haksupcauy ssukiey nathanan ehwi tayangto mich ehwi milto yenkwu* [A study on the lexical variation and lexical density shown in writing of KFL learners]. *Journal of Language Sciences*, *19*(1), 99–117.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Bulté, B., & Roothooft, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, *91*, 102246. https://doi.org/10.1016/j.system.2020.102246

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

Castañeda-Jiménez, G., & Jarvis, S. (2014). Exploring lexical diversity in second language Spanish. In K. L. Geeslin (Ed.), *Handbook of Spanish second language acquisition* (pp. 498–513). Wiley. https://doi.org/10.1002/9781118584347.ch28

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a foreign language*. Routledge.

Chung, T., & Gildea, D. (2009). Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 718–726). https://doi.org/10.3115/1699571.1699606

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

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

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

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

Guiraud, P. (1960). Problèmes et méthodes de la statistique linguistique *[Problems and methods of linguistic statistics]*. Reidel.

Harrell, F. E., Jr., Harrell, M. F. E., Jr., & Hmisc, D. (2017). *Rms* R Package. https://CRAN.R-project.org/package=rms

Haspelmath, M., & Michaelis, S. M. (2017). Analytic and synthetic: Typological change in varieties of European languages in language variation. In I. Buchstaller & B. Siebenhaar (Eds.), *European perspectives VI: Selected papers from the 8th international conference on language variation in Europe (ICLaVE 8), Leipzig 2015* (pp. 3–22). Benjamins. https://doi.org/10.1075/silv.19.01has

Haspelmath, M., & Sims, A. (2013). *Understanding morphology*. Routledge.

Herdan, G. (1960). *Type-token mathematics: A textbook for mathematical linguistics*. Mouton.

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, Language, & Hearing Research*, *29*(1), 129–134. https://doi.org/10.1044/jshr.2901.129

Hur, W.-J., & Lee, M. (2019). *hankwuke haksupcauy swuktaltopyel ehwi sayong yangsang* [Study on Korean language learner's vocabulary usage]. *Bilingual Research, 77,* 215–239.

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). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–43). Benjamins.

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

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

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

Kang, J.-H. (2018). *Hankwuke haksupcauy ehwilyek paltalkwa ehwi sayong yangsang yenkwu -ssuki theyksuthuey nathanan ehwi chukcengul cwungsimulo-* [A study on the vocabulary development and the lexicon use aspect of Korean learners – focusing on the lexicon development appearing in the writing test]. *Bilingual Research*, *71*, 31–64.

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

Kwon, H.-C., Kang, M.-Y., & Choi, S.-J. (2004). Stochastic Korean word-spacing with smoothing using Korean spelling checker. *International Journal of Computer Processing of Languages*, *17*(4), 239–252. https://doi.org/10.1142/S0219427904001103

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–476). Routledge.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2023). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, 1–22. Advance online publication. https://doi.org/10.1017/S0272263123000402

Lee, S.-M. (2017). *hankwuke haksupcauy malhakiwa ssukiey nathanan ehwi sayonguy congtancek yenkwu* [A longitudinal study of vocabulary usage presented in speaking and writing of Korean learners]. *The Korean Language and Literature*, *74*, 183–214.

Leech, G. N., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Longman.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing Talk, vol. 2, the database*. Lawrence Erlbaum.

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

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

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, S. Christopher, M. Á. G. González, & S. M. Doval-Suárez (Eds.), *The dynamics of language use* (pp. 271–285). John Benjamins.

Mhammond. (n.d.). *Pywin32 Python Package*. https://github.com/mhammond/pywin32

Nation, P. (1990). *Teaching and learning vocabulary*. Newbury House.

Park, L. (2018). *KoNLPy Python Package* Documentation. https://buildmedia.readthedocs.org/media/pdf/konlpy/v0.3.3/konlpy.pdf

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv Preprint arXiv: 200307082*. https://doi.org/10.18653/v1/2020.acl-demos.14

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). *Mass R Package*. https://cran.r-project.org/web/packages/MASS/index.html

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shin, G-H, & Jung, B. K. (2021). Automatic analysis of passive constructions in Korean: Written production by Mandarin-speaking learners of Korean. *International Journal of Learner Corpus Research*, *7*(1), 53–82. https://doi.org/10.1075/ijlcr.20002.shi

Sung, H., & Shin, G-H (2023). Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023) (pp. 72–82). https://doi.org/10.18653/v1/2023.bea-1.6

Thomson, G. H., & Thompson, J. R. (1915). Outlines of a method of the quantitative analysis of writing vocabularies. *British Journal of Psychology*, *8*(1), 52–69. https://doi.org/10.1111/j.2044-8295.1915.tb00128.x

TOPIK. (n.d.). Retrieved July 31, 2022, from https://www.topik.go.kr/HMENU0/HMENU00018.do

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. In S. Jarvis & M. Daller (Eds.), Vocabulary knowledge: Human ratings and automated measures (pp. 79–104). John Benjamins. https://doi.org/10.1075/sibil.47.05ch3

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. https://doi.org/10.1093/applin/amw009

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. https://doi.org/10.1023/A:1001749303137

Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568–587. https://doi.org/10.1177/1362168818817945

Won, Y. (2016). Common European framework of reference for language (CEFR) and Test of Proficiency in Korean (TOPIK). *International Journal of Area Studies*, *11*(1), 39–58. https://doi.org/10.1515/ijas-2016-0003

Won, H., Lee, H., & Kang, S. (2020). Multi-prototype morpheme embedding for text classification. In *The 9th International Conference on Smart Media and Applications* (pp. 295–300). https://doi.org/10.1145/3426020.3426095

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

Appendix. The number and type of L1s of the Korean learners included in the study

The table shows the number and type of L1s of Korean learners included in this study. It is organized into three columns showing the rank, the L1, and the number (N) of learners.

Rank	L1	N	Rank	L1	N	Rank	L1	N
1	Mandarin Chinese	2105	18	Arabic	14	35	Slovak	3
2	Japanese	784	19	Burmese	14	36	Madagascar	2
3	Vietnamese	258	20	Italian	14	37	Nepali	2
4	English	232	21	Sinhala	14	38	Tajik	2
5	Cantonese	225	22	Turkish	14	39	Amharic	1
6	Russian	82	23	Dutch	9	40	Armenian	1
7	Mongolian	67	24	Portuguese	9	41	Estonian	1
8	Thai	59	25	Azerbaijani	6	42	Finnish	1
9	Indonesian	53	26	Persian	6	43	Javanese	1
10	French	35	27	Hindi	5	44	Kannada	1
11	Spanish	31	28	Norwegian	5	45	Kurdish	1
12	Kazakh	29	29	Polish	5	46	Luxembourgish	1
13	German	22	30	Ukrainian	5	47	Rumanian	1
14	Kyrgyz	22	31	Khmer	4	48	Swahili	1
15	Malay	18	32	Tagalog	4	49	Tamil	1
16	Swedish	15	33	Hungarian	3	50	Tetun-Dili	1
17	Uzbek	15	34	Lao	3	51	Tibetan	1