

# Evaluating Jarvis and Hashimoto’s Operationalizations of Word Types and Their Influence on Lexical Diversity Measures

Yajie LI

Graduate School of Integrated Arts and Sciences

Hiroshima University

Simon FRASER

Institute for Foreign Language Research and Education

Hiroshima University

Jon CLENTON

Graduate School of Integrated Arts and Sciences

Hiroshima University

A flurry of recent papers (Brown et al., 2020; Kremmel & Schmitt, 2016; McLean, 2018; Stoeckel et al., 2018; Treffers-Daller et al., 2018) highlight how different operationalizations of what constitutes a ‘word’ influence learner vocabulary knowledge measures. In the field of second language writing in particular, the vocabulary used in learner texts is an important means of evaluating language ability. However, before making any quantitative analyses of this vocabulary, it is first necessary to ensure that appropriate word counting units are used. One means of examining word counting is through the lens of lexical diversity measures. Lexical diversity (LD) is a measure of the variety of word knowledge exhibited in speaking or writing, and is used in a number of assessment tools to predict learner proficiency levels. In this review article, we examine a recent study (Jarvis & Hashimoto, 2021) that investigates three LD measures (MTLD, MTLD-W, and MATTR) using five different word unit operationalizations.

## A SUMMARY OF JARVIS AND HASHIMOTO (2021)

### Overview

Jarvis and Hashimoto (2021) investigate five different operationalizations of word types within three lexical diversity (LD) measures. The aim is to determine the most helpful LD measures and to demonstrate potential influences of the different word units on each LD index. Their three LD measures consist of the measure of textual lexical diversity (MTLD), moving average MTLD with wrap-around measurement (MTLD-W), and moving average type-token ratio (MATTR). They employ five different definitions of word types: orthographic forms, lemmas with automated part-of-speech (POS) tags (lemmas-A), lemmas with manually corrected POS tags (lemmas-C), lemmas, and word families. Jarvis and Hashimoto utilize the three LD measures and five types of word units to examine 60 narrative essays written by English, Finnish, and Swedish first-language speakers. Fifty-five human raters evaluated each writing sample, with raters comprising 20 graduate (first-language users of English) and 35 undergraduate students (15 first-language speakers of English; 20 second-language speakers of English with TOEFL iBT scores over 100) studying linguistics in a university in America. The results for the three LD measures (MTLD, MATTR-50, and MTLD-W) were found to be similar, meaning that it was not possible to determine whether individual

measures outperformed others. Mixed results were reported for two of the word units (orthographic forms and lemmas-A) across the LD measures; in contrast, the other three word units (word families, lemmas, and lemmas-C) yielded very similar results across the three LD measures.

### Issues Concerning the Assessment of Lexical Diversity

Jarvis and Hashimoto present three main issues in assessing LD: differing operationalization of types, text length, and human LD ratings. They explain that, in essence, LD relates to word variety in writing and speaking, and word variety can be measured by the number of different words found in written or spoken texts. Tokens are instances of each word occurring in a text, and multiple tokens are repeated items of those found earlier in the text. Conversely, types represent the number of unique words in the text without repetition. Jarvis and Hashimoto state that most LD measures are variety-repetition (VR) measures, dependent on the counting of types. Since types are so important in LD as measured by VR, Jarvis and Hashimoto believe that it is crucial to reach a theory-and-evidence-based principle of how types should be determined and described in this field.

Text length has long been a major issue for LD measurement, and one that many studies have questioned. Jarvis and Hashimoto suggest that there are several VR measures that can potentially address this concern. They cite Carroll (1938) as being the first to devise a means to solve the problem of text-length variation; many other researchers have since followed (e.g., Carroll, 1964; Covington & Maas, 1972; Dugast, 1978; Guiraud, 1960; Herdan, 1960; Johnson, 1939; McFall, 2010; McKee, Malvern, & Richards, 2000; McCarthy, 2005; McCarthy & Jarvis, 2007; Vidal & Jarvis, 2020; Yule, 1944).

### Lexical Diversity Measures

The different measures of lexical diversity are a function of the type-token ratio (TTR, Johnson, 1939), which computes the total number of types (unique words) divided by the total number of tokens (all words) in a text. The Mean Segmental Type-Token Ratio (MSTTR, Johnson, 1944) divides the text into equal-sized parts and takes the mean of the TTRs of several consecutive samples as the final LD score. The problem with MSTTR is that not all the text is used during the calculation process, and this discarding of data has a significant impact on the LD measurements in short texts (McCarthy & Jarvis, 2010).

*D* (Malvern & Richards, 1997) appears to be the most widely used LD measure. *D* is calculated using CLAN (Computerized Language ANalysis) software, developed by Brian MacWhinney (2000). The calculation is made through a series of random sampling and curve-fitting procedures by the *vocd* program within CLAN. Jarvis and Hashimoto (2021) affirm that *D* increases along with text length, as recorded by both Fergadiotis, Wright, and West (2013) and McCarthy and Jarvis (2007, 2010).

The Measure of Textual Lexical Diversity (MTLD), developed by McCarthy (2005), uses sequential analysis of a sample. A constant TTR value (e.g., under 0.72) is maintained for increasingly longer parts of the sample. For instance, MTLD computes TTR from the first word, the first two words, and so on, until it drops below 0.72. If a TTR value falls below 0.72 at 55 tokens, then the first segment length is 54. The MTLD program then calculates the second segment from token 55, and the final MTLD value is a measure of the mean length of all such segments in which the TTR remains above 0.72.

The Moving Average Type-Token Ratio (MATTR), introduced by Covington and McFall (2010), is

also a VR measure. MATTR employs a ‘moving window’, which estimates TTR for each successive window (a fixed length of text, e.g., 50 tokens) until the end of the text; the resultant final MATTR is the mean TTR value of all segments of the text. One advantage of MATTR is that it includes all the words in each text.

MTLD-W, introduced by Kyle, Crossley, and Jarvis (2020) and Vidal and Jarvis (2020), adopts the moving window approach of MATTR, while including a ‘wrap-around’ process that calculates the final segment length by adding words to the initial segment of a text until a TTR of 0.72 is reached.

Since MATTR and MTLD appear to be more accurate than other LD indices, and MTLD-W offers improvements on MTLD, Jarvis and Hashimoto (2021) chose to use these three LD measures in their study.

### **Defining Word Units**

A key issue addressed by Jarvis and Hashimoto (2021) relates to the different ways in which word units can be defined, and how best to operationalize these units in LD studies. Existing possible categories consist of word families, flemmas, lemmas, and orthographic forms. Word families include all derivations and inflections of the same root (Bauer & Nation, 1993). Flemmas cover all inflections of words with the exact spelling irrespective of the meaning, or part of speech (Pinchbeck, 2014). Lemmas are all the inflections of a word with the same part of speech. With orthographic forms, all inflections are regarded as different types. In Jarvis and Hashimoto’s paper, all four different word units are employed.

### **Subjective vs. Objective Constructs of Lexical Diversity**

Jarvis and Hashimoto also consider conceptual elements, both subjective and objective, that comprise the construct of lexical diversity. For this purpose, they refer to Zipf’s (1935) study, which regarded lexical diversity as a phenomenon existing fundamentally in the mind, relating more to redundancy in language use (a subjective construct) than to repetition (an objective construct). Similarly, Yule (1944) treated ‘lexical richness’ as a reflection of the number of types in a learner’s mental lexicon. Jarvis (2017) presented a study investigating Zipf’s suggestion that human perception of lexical diversity could be superior to other LD measures. In this earlier study, Jarvis observed that human judges were consistent in their ratings without receiving any training. In Jarvis and Hashimoto’s (2021) study, which employs the same methodology as Jarvis (2017), the raters appeared to offer high inter-rater reliability (Cronbach’s  $\alpha > 0.90$ ), suggesting again that human raters show excellent agreement without any training or LD rubric. While few studies have explored the relationships between human ratings of LD and LD measures, Jarvis and Hashimoto posited that VR measures of LD would be able to account to a large degree for the variation in human judgments.

### **Research Objectives**

Jarvis and Hashimoto’s primary aim was to ascertain the most effective of the three VR-based LD measurements (MTLD, MATTR, and MTLD-W), when compared with LD as determined by human ratings. An additional goal was to discover which word unit definitions most closely reflect human ratings. Also, since the application of different word units requires part-of-speech (POS) tagging, Jarvis and Hashimoto wanted to compare the accuracy between automated POS tags and human corrected POS tags. In their study, they ran the cleaned texts through the TreeTagger program automatically, with TreeTagger adding the POS tags and the base form lemma/flemma to the orthographic forms. To establish the accuracy of TreeTagger,

they also included the human corrected POS tag process. Accordingly, their two major research objectives were to determine: (1) which VR measures (MTLD, MTLD-W, MATTR) mirror human ratings; and (2) which word units worked best amongst the three LD measures.

The five categories of word units used in the study are orthographic forms, lemmas-A (lemmas with the automated POS tagger), lemmas-C (lemmas with manually corrected POS), flemmas, and word families.

## The Corpus

The corpus data for the study came from Jarvis (2017), with participant English essays written by first-language users of English ( $n=13$ ), Finnish ( $n=31$ ), and Swedish ( $n=16$ ). Participants were required to write a narrative descriptive essay about an eight-minute-long portion of the Chaplin film *Modern Times*. All writing samples were rated for CEFR writing proficiency by 41 college students majoring in linguistics. Fifty-five human raters judged the LD of each essay, with all raters being undergraduate or graduate students of linguistics. These two groups of human raters did not receive any training, but reportedly had high inter-rater reliability (Cronbach's  $\alpha = .977$  and  $.983$ ). Jarvis and Hashimoto's data computing process differed from many earlier LD studies. Instead of using existing programs, they created their own Python scripts using the three LD measures (MTLD, MTLD-W, and MATTR). They utilized the Tree-Tagger program to produce POS tags automatically, and treated lemmas and flemmas as lemmas-A. They also created a file with corrected POS tags (lemmas-C). Root forms of all words based on Bauer and Nation's (1993) classification of level-six word families were listed.

## Results

The results of the research demonstrated a high degree of accuracy for TreeTagger (accuracy statistics above 0.90) across major POS divisions except for expletives. Pearson correlations (see Table 1) indicate that MTLD had the highest correlations with lemmas-C. The word family worked better with MTLD-W; lemmas-A performed better with MATTR-50.

**TABLE 1. Pearson Correlations between Automated Measures and Mean LD Ratings**

	MTLD	MTLD-W	MATTR-50
Orthographic form	0.490	0.411	0.499
Lemma-C	0.528	0.474	0.478
Lemma-A	0.384	0.363	0.501
Flemma	0.516	0.466	0.476
Word family	0.525	0.485	0.485

*Note.* All coefficients in this table have a  $p$ -value less than 0.00133

Jarvis and Hashimoto also compared five different operationalizations of word types within the three LD measures through pairwise comparison. Their results indicate that word family, flemma, lemma-C, and orthographic form outperformed lemma-A in MTLD, and that word family outperformed lemma-A in MTLD-W. Regarding MATTR-50, lemma-A outperformed orthographic form, flemma, and word family, and orthographic form outperformed flemma. Moreover, MATTR-50 outperformed MTLD-W when using lemma-A word types. Linear regression analysis indicates that the highest values were obtained with MTLD by using lemmas-C and word families. In addition, using Cook's distance, Jarvis and Hashimoto investigated the texts that did not meet their designated criteria with the automated LD measures or different types of operationalization; five texts were found to be outlier texts.

## Conclusions

Jarvis and Hashimoto conclude their paper with a discussion of three main points. First, they retrace their research questions and note that MTLD correlates most highly with human ratings, followed by MATTR-50 and MTLD-W. They report no significant differences between LD measures, and their confidence intervals revealed few substantive differences between operationalizations of types following as many as thirty comparisons. However, they suggest that their results do not imply that all the measures or types provide the same function. Their findings also indicate that using the uncorrected POS tags (lemma-A) might lead to unreliable results; MATTR-50 should not be expected to produce better results with less favourable data. In addition, orthographic forms produced the second strongest correlations with human ratings for MATTR-50 ( $r=.499$ ), but the second weakest correlations with MTLD ( $r=.490$ ) and MTLD-W ( $r=.411$ ). They attribute the reasons to window size, noting that further investigation is necessary into the relationships between measures, window size, and types. In their study, Jarvis and Hashimoto also observe that among types, word families played a constant and significant role, yielding the second highest correlation with human judgments and MTLD ( $r=.525$ ), the strongest for MTLD-W ( $r=.485$ ), and the third highest for MATTR-50 ( $r=.485$ ). They suggest that this finding is quite unexpected, citing recent papers which claim that lemmas (Kremmel, 2016; Kremmel & Schmitt, 2016) and flemmas (McLean, 2018) are more appropriate than word families to assess vocabulary knowledge. Jarvis and Hashimoto believe that highly professional human raters would judge words with the same root as being less diverse than words belonging to a variety of word families. They conclude that word families, flemmas, and lemmas-C were the three most stable types in their studies.

The second point relates to the accuracy of TreeTagger in their study, which was an unexpectedly high 97.2%. In their research, Jarvis and Hashimoto investigated accuracy in connection with three prominent POS tags – noun, verb, and adjective – and they maintain that human examinations of POS mainly focus on these macro-level tags. However, they also suggest that even a small number of POS mistakes can have contrary effects on LD measurements. Therefore, they consider that POS accuracy checking is essential, and is something that needs to be implemented in future natural language processing and applied linguistics studies.

Third, Jarvis and Hashimoto suggest that a degree of construct validity is demonstrated in their paper, since each of the three measures accounted for no more than 27.6% of the variance relating to the LD of human judgments, indicating there were factors other than VR measures that could influence LD. Jarvis

(2013a, 2013b, 2017) has suggested that there are as many as seven variables that might explain differences in human ratings, and the VR measures of LD under discussion here might only be a small part of the LD construct.

## **A CRITIQUE OF JARVIS AND HASHIMOTO (2021)**

Jarvis and Hashimoto's (2021) paper represents a pilot study that attempts to validate three VR (variety-repetition) measures of lexical diversity with human rater LD scores according to five operationalizations of word units. Their study contributes to current LD studies both methodologically and theoretically. The research, however, is not without its shortcomings, and we turn our attention to these in the following sections.

### **Problems with the Corpus**

First, the corpus used in the study is problematic. As the authors themselves point out, many texts in their corpus are short, with some comprising fewer than 150 words. Widely observed within LD studies, if texts are too short, no differences between the different operationalizations of types among texts can be observed. According to Kyle et al. (2020), Jarvis and Hashimoto (2021), and Vidal and Jarvis (2020), human judgments tend to be influenced by text length, and usually, long texts receive higher lexical diversity scores simply because longer texts include a greater range of ideas. The texts in Jarvis and Hashimoto's corpus are all narrative writing samples describing a movie clip, meaning only one genre is covered; also, there are only sixty essays in total. As for the participants, only those at four proficiency levels from CEFR A1 to B2 level are included in their study, with just two students at B2 level, and nine students at A1. If the intention of the research is to build a standard for current LD studies, then a much larger corpus, which includes more genres and writing samples from participants at different proficiency levels, will be necessary.

### **Determining the Most Appropriate Word Units**

Further concerns are that the most appropriate types across all three LDs have not been determined in the research, and neither have the types that best fit specific measures. Through an examination of three similar LD measures using different definitions of word types, the authors report mixed results, suggesting that the choice of word unit influences the measurement of LD. Decisions regarding which types are most appropriate for use in future studies need further explication. Jarvis and Hashimoto claim that the most stable word counting units employed in their study are word families, flemmas, and lemmas-C (lemmas with human corrections). One element that needs considering, however, is that they take word families as being at level six of Bauer and Nation's (1993) levels of word family. Counting word families in this way reduces learner LD scores during the calculation, because it treats all the words which share the same root as the same type, and so will not distinguish between participants with different levels of word knowledge. In Jarvis and Hashimoto's study, most participants belong to A2 ( $n=23$ ) and B1 ( $n=26$ ) CEFR levels. Lower-level learners will know fewer derivations and inflections, so it is necessary to choose the level of word family carefully, or to consider using other lexical units, in order to measure their productive vocabulary knowledge more accurately.

### **Accuracy of Human LD Rating Scores**

The third main reservation we have with Jarvis and Hashimoto's paper relates to the human rating of LD scores, and the fact that using a large number of human raters to measure LD is hard to implement in practice. In their study, 41 human raters rated both the writing qualities and LD scores, indicating that the same raters had been used twice to rate the same essays. Human raters scored the writing samples using the CEFR Overall Written Production rubric, and they also rated LD after being told that it is not the same thing as writing quality. Because the raters did not receive any training in the rating of LD, it is unclear to what extent the CEFR writing rubric might have influenced them. In other words, the accuracy of their LD rating is questionable.

Regarding the number of human raters in the study, there were 55 reliable raters remaining after four non-native raters were removed. To find as many reliable raters as this to rate all the writing samples in a study seems impractical. Kyle et al. (2020) also adopt direct human judgments in rating all writing samples, while in their paper they use the adjustment scores from two trained human raters until the raters reach an agreement on the same essay. In Kyle et al.'s research, abundance (number of different types) was found to reflect the LD rating most. It should also be pointed out that Jarvis and Hashimoto's study includes both native and non-native raters, and the potential influence of the different first language of the raters has not been considered.

### **CONCLUSION**

Jarvis & Hashimoto's paper is important because it employs three widely used LD measures (MTLD, MTLD-W, and MATTR) to evaluate language learner proficiency levels from CEFR A1 to B2. Their study also includes five different word units for each LD measurement to investigate how these might influence the LD results in distinguishing different proficiency levels. Their findings indicate that the three LD measures produce different results with five types of word units. These mixed results suggest that lemmas, flemmas, and word families work well with all three LD measurements, with further research required in this area.

Our evaluation of the paper has highlighted three main weaknesses with Jarvis and Hashimoto's approach. The first relates to the corpus used in their paper: Some of the texts are short (fewer than 150 English words), which undoubtedly influenced the LD scores as judged by the human raters. In addition, the corpus does not include texts written by high proficiency level participants (C1 and C2 learners), and there are only two B2 proficiency level participants. The second shortcoming concerns the failure of the study to determine which word units work better than the others across the three LD measurements. The third weakness of the research relates to the human raters used in the study: Fifty-five raters scored the LD, of whom 41 also rated the writing quality of the essays; whether the raters have been influenced by the CEFR writing rubric remains unclear.

Further research is necessary to address the questions emerging from Jarvis and Hashimoto's study. Future studies might look more closely at the construct of lexical diversity, the POS taggers used, and the issues relating to lexical diversity measurements (e.g., the interaction between measures, window size, and operationalization of types). Research investigating corpora with a wider range of texts is also needed.



## REFERENCES

- Bauer, L., & Nation, P. (1993). Word families. *International journal of Lexicography*, 6 (4), 253–279, <https://doi.org/10.1075/ijcl.16.1.02bar>
- Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, amaa061, <https://doi.org/10.1093/applin/amaa061>
- Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, 2, 379–386.
- Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.
- Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théorique du vocabulaire. *Français (Le) Moderne Paris*, 46 (1), 25–32.
- 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>
- Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, 22, 397–408, [https://doi.org/10.1044/1058-0360\(2013/12-0083](https://doi.org/10.1044/1058-0360(2013/12-0083)
- Guiraud, P. (1960). *Problemes et méthodes de la statistique linguistique*. D. Reidel.
- Herdan, G. (1960). *Quantitative linguistics*. Butterworth.
- Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning*, 63 (S1), 87–106, <https://doi.org/10.1111/j.1467-9922.2012.00739.x>
- Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–44). John Benjamins Publishing.
- 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. *Natural Language Processing for Learner Corpus Research (NLP for LCR)*, 7 (1), <https://doi.org/10.1075/ijlcr.20004.jar>
- Johnson, W. (1939). *Language and speech hygiene: An application of general semantics*. Edwards Brothers.
- Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs*, 56 (2), 1–15.
- Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL QUARTERLY*, 50 (4), 976–987, <https://doi.org/10.1002/tesq.329>.
- Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, 13 (4), 377–392.
- Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 1–17, <https://doi.org/10.1080/15434303.2020.1844205>
- MacWhinney, B. (2000). Volume 1: Transcription format and programs Volume 2: *The Database. The CHILDES project: Tools for analyzing talk*.
- Mass, H. D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift*



*für Literaturwissenschaft und Linguistik*, 2 (8), 73–79.

- 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, The University of Memphis.
- 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>
- McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15 (3), 323–338, <https://doi.org/10.1093/lc/15.3.323>
- McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, 39 (6), 823–845, <https://doi.org/10.1093/applin/amw050>
- Pinchbeck, G. G. (2014). Lexical frequencies profiling of Canadian high school diploma exam expository writing: L1 and L2 academic English. *Roundtable presentation at American Association of Applied Linguistics, Toronto, Ontario*.
- Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*, 41 (4), 601–606, <https://doi.org/10.1093/applin/amy059>
- 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>
- Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.
- Zipf, G. K. (1935). *The psycho-biology of language*. Houghton Mifflin.

## ABSTRACT

### **Evaluating Jarvis and Hashimoto's Operationalizations of Word Types and Their Influence on Lexical Diversity Measures**

Yajie LI

Graduate School of Integrated Arts and Sciences

Hiroshima University

Simon FRASER

Institute for Foreign Language Research and Education

Hiroshima University

Jon CLENTON

Graduate School of Integrated Arts and Sciences

Hiroshima University

This review article examines a recent study (Jarvis and Hashimoto, 2021) investigating three lexical diversity (LD) measures, each using five different word unit operationalizations. Jarvis and Hashimoto's aim is to determine the most effective LD measures and demonstrate the potential influences of the different word units on each LD index. The LD measures include the measure of textual lexical diversity (MTLD), moving average MTLD with wrap-around measurement (MTLD-W), and moving-average type-token ratio (MATTR). Each measure is investigated with types operationalized as orthographic forms, lemmas using automated part-of-speech (POS) tags, lemmas with manually corrected POS tags, flemmas, and word families. These measures are used to examine 60 narrative essays written by English, Finnish, and Swedish first-language speakers; correlations with the LD ratings of 55 human raters are investigated. Jarvis and Hashimoto conclude that while the three LD measures are comparable, two of the word unit operationalizations produce mixed results.

In the review, following a summary of the paper and explanation of important concepts, the strengths and weaknesses of the study are evaluated with a view to assessing its importance in the field. Jarvis and Hashimoto's research undoubtedly advances understanding of lexical diversity and its measurement. However, there are problems concerning the corpus of texts and the human raters used in the study. Also, the question of which word units work best remains unanswered.

## 要 約

### Jarvis と Hashimoto による単語タイプの運用と その語彙多様性指標への影響の評価

ヤジェ・リー

広島大学大学院総合科学研究科

サイモン・フレイザー

広島大学外国語教育研究センター

ジョン・クレントン

広島大学大学院総合科学研究科

本レビュー論文では、3種類の語彙多様性 (LD) 指標を調査した最近の研究 (Jarvis and Hashimoto, 2021) を検証する。これら3種類の LD 調査では各々 5つの異なる単語単位の運用が用いられている。Jarvis と Hashimoto の目的は、最も有効な LD 指標を決定し、異なる単語単位が各 LD 指標に及ぼす潜在的な影響を実証することである。LD 指標には、テキストの語彙多様性指標 (MTLD)、ラップアラウンド測定による移動平均 MTLD (MTLD-W)、及び移動平均タイプ-トークン比 (MATTR) が含まれる。それぞれの指標は、正書法、自動品詞タグを用いたレンマ、人手で品詞タグを修正したレンマ、フレマ、ワードファミリーとして運用されるタイプで調査されている。これらの指標を用いて、英語、フィンランド語、スウェーデン語の第一言語話者によって書かれた60の物語エッセイを調査し、55人の評価者による LD 評価との相関を調査した。Jarvis と Hashimoto は、3つの LD 指標は同等であるが、2つの単語単位の運用においては結果が分かると結論付けている。

本レビューでは、論文の要約と重要な概念の説明を行った後、この分野における重要性を評価する観点から、この研究の長所と短所を評価する。Jarvis と Hashimoto の研究は、語彙の多様性とその測定に関する我々の理解を間違いなく前進させるものである。しかしながら、この研究で使用されたテキストのコーパスと人間の評価者についての懸念があり、どの単語単位が最も効果的かという問題には答えがないままである。