System 47 (2014) 139–145

Contents lists available at ScienceDirect

System

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

Evaluating the comparability of two measures of lexical diversity

Fredrik deBoer*

Purdue University, United States

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form 18 August 2014
Accepted 12 October 2014
Available online 30 October 2014

Keywords:
Lexical diversity
vocd
HD-D
Lexicon
Computational linguistics
Applied linguistics
Textual processing

ABSTRACT

Language practitioners and others increasingly rely on computerized assessments of large samples of written texts. In order to provide teachers and researchers with useful knowledge, new, more accurate metrics must be developed to aid in these assessments. One common aspect of such assessments is lexical diversity, or the displayed range of diversity in vocabulary. The *vocd* program and the metric it develops, VOCD-D, have become popular options for researchers attempting to assess lexical diversity. However, researchers have argued that this metric is in fact a complex approximation of a more direct and less variable measure derived from probability sampling, known as HD-D. Using a data set of essays written by Chinese, Japanese, Korean, and native English-speakers drawn from the International Corpus Network of Asian Learners of English, this research investigates that approximation by comparing correlations across L1 and L2 writers. In all cases, the correlations between HD-D and VOCD-D are very high, suggesting that the similarity between these metrics is indeed a product of their statistical mechanisms.

1. Introduction

1.1. Statement of the problem

Despite considerable investigation, there remains significant scholarly disagreement about the nature of lexical diversity, or the displayed range of vocabulary in a language sample, and the best metric(s) for assessing it. The ability to accurately reflect the range of displayed vocabulary in a given text is seen as an essential element of evaluating that text quantitatively. This type of quantitative assessment has numerous uses for linguistic and educational research. Malvern and Richards (2012: 1) argue that lexical diversity has broad practical and theoretical applicability, with research and diagnostic use in fields such as language acquisition, linguistic interaction, demographic language performance, language impairment, and mental health research. Language teachers and testers, in particular, require valid and reliable metrics for measuring lexical diversity. For example, measures of lexical diversity can be used to evaluate how effectively language learners integrate vocabulary into their actual language production, which is potentially of greater interest to instructors and researchers than results on tests of passive vocabulary (Laufer & Paribakht, 1998; Nation, 2007). An understanding of how language learners utilize a diverse vocabulary in their language production can help instructors guide their teaching, particularly in contexts such as a college writing classroom where formal vocabulary instruction is rare. Additionally, assessment of lexical diversity has been utilized effectively in researching infant and childhood language development, such as by Duran, Malvern, Richards, and Chipere (2004).

* Tel.: +1 860 336 9931; fax: +1 765 494 3780.
E-mail address: fdeboer@purdue.edu.

http://dx.doi.org/10.1016/j.system.2014.10.008
0346-251X/

The process of assessing a writing sample's vocabulary diversity has proven to be far more complicated than assumed. This has resulted in a number of competing metrics that have been advanced as effective measures of lexical diversity, with different researchers advocating different metrics. Among the most popular and most discussed of these new metrics is D, a measure generated via a process of curve-fitting by the *vocd* computer program. D attempts to measure the diversity of vocabulary in writing by taking random samples of words and comparing the observed diversity to ideal curves (see below). D was theoretically explored in 1997 by David Malvern and Brian Richards and implemented computationally in 2000 by Gerald McKee, Malvern, and Richards. In order to avoid confusion, this research will refer to the D metric derived from *vocd* as "VOCD-D" throughout.

Since software capable of generating values of VOCD-D has become freely available in the last ten years, such as with the CHILDES project's CLAN and the University of Memphis's Coh-Metrix, much research has been conducted utilizing this metric for lexical diversity. But despite VOCD-D's popularity, it has also been subject to scrutiny, most consistently by Philip McCarthy and Scott Jarvis (2007, 2010). McCarthy and Jarvis have made two essential claims: one, that VOCD-D is affected by text length to a degree not acknowledged by its proponents, and two, that it is in fact a representation of another metric, HD-D, which is slightly more accurate and slightly more stable. HD-D, or hypergeometric diversity of D, is an alternative metric developed by McCarthy and Jarvis and described below. The first point has been addressed in several empirical studies (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007). This research is an attempt to assess the second.

1.2. Competing measures of lexical diversity

The simplest method for measuring lexical diversity lies in simply counting the number of different words (NDW) that appear in a given text. This figure is now typically referred to as *types*. This metric benefits from simplicity in both concept and in measurement, and can be easily generated from simple computer programs. However, the problems with NDW are obvious. The figure is entirely dependent on the length of a given text. It's impossible to meaningfully compare a text of 50 words to a text of 75 words, let alone to a text of 500 words or 3000 words. Problems with robustness of measures across differing sample size — the difficulty in making statistically reliable, interpretable measures across language samples of differing lengths — have been the most consistent issue with attempts to measure lexical diversity. As Malvern, Richards, Chipere, and Duran (2004) write, "how many different words appear in a language sample will in all probability depend on how many words there are in total and this is the heart of many problems in the measurement of lexical diversity" (p. 17).

The most popular method to address this problem has been Type-to-Token Ratio, or TTR. TTR is a ratio measurement where the number of different types is divided by the number of total words, or *tokens*. This calculation results in a proportion between 0 and 1, with a higher figure indicating a more diverse range of vocabulary in the given sample. A large amount of research has been conducted utilizing TTR over a number of decades. However, the robustness of TTR, and thus its value as a descriptive statistic, has been seriously disputed. These criticisms are both empirical and theoretical in nature. Empirically, TTR has been shown in multiple studies to steadily decrease with sample size, making it impossible, after a certain number of tokens, to use the statistic to discriminate between texts (Broeder, Extra, and van Hout, 1993; Chen & Leimkuhler, 1989; Malvern et al., 2004; Richards, 1987). Theoretically, TTR falls due to the nature of language and the repetition of functional vocabulary such as prepositions and articles. Malvern et al. (2004) explain the theoretical reason for this observed phenomenon: "Adding an extra word to a language sample always increases the token count (N) but will only increase the type count (V) if the word has not been used before…. Consequently, the type count (V) in the numerator increases at a slower rate than the token count (N) in the denominator and TTR inevitably falls" (p. 22). This loss of discriminatory power over sample size renders TTR an ineffective measure of lexical diversity.

Research involving lexical diversity has been applied in a number of educational fields and contexts. In order for consideration of lexical diversity to effectively contribute to such fields, however, it must be expressed in metrics that are valid and reliable. A variety of valuable extant studies have had their findings undermined by our contemporary understanding of TTR's unreliability over sample size. For example, Engber (1995) researched the relationship between the holistic scores of 66 student essays and several measures of lexical control, including lexical diversity. Engber found that there was a robust and significant correlation between a student's demonstrated lexical diversity and the rating of that student's essay. However, the research utilized the conventional TTR measure for lexical diversity, which is flawed for the reasons previously discussed. Li (2000) analyzed 132 emails written by 22 ESL students, which addressed a variety of tasks and contexts. These emails were subjected to linguistic feature analysis, including lexical diversity, as well as syntactic complexity and grammatical accuracy. Li found that there were slight but statistically significant differences in the lexical diversity of different email tasks (Narrative, Information, Persuasive, Expressive). She also found that lexical diversity was essentially identical between structured and non-structured writing tasks. However, she too used the flawed TTR measure for lexical diversity. In the context of the period of time in which these researchers conducted their studies, the use of TTR was appropriate, but its flaws have eroded the confidence we can place in such research.

Many mathematical transformations of TTR have been proposed to address these issues, such as Guiraud's Root TTR, which involves dividing types by the square root of the tokens, or Somer's S, which utilized logarithms. However, as Malvern and Richards (2012: 2) argue, none of these adjustments work. Among the problems with these mathematical transformations of TTR is a lack of construct-specific reasons to perform that particular transformation. For Guiraud's root, for example, taking the square root of the number of tokens does indeed reduce the speed with which TTR is lowered, making it easier to distinguish between different samples. But there is essentially no theory-internal reason to transform the figure in this way.