vocd: A theoretical and empirical evaluation

Philip M. McCarthy *University of Memphis, USA,* and
Scott Jarvis *Ohio University, USA*

A reliable index of lexical diversity (LD) has remained stubbornly elusive for over 60 years. Meanwhile, researchers in fields as varied as *stylistics*, *neuropathology*, *language acquisition*, and even *forensics* continue to use flawed LD indices – often ignorant that their results are questionable and in some cases potentially dangerous. Recently, an LD measurement instrument known as *vocd* has become the virtual tool of the LD trade. In this paper, we report both theoretical and empirical evidence that calls into question the rationale for *vocd* and also indicates that its reliability is not optimal. Although our evidence shows that *vocd*'s output (D) is a relatively robust indicator of the aggregate probabilities of word occurrences in a text, we show that these probabilities – and thus also D – are affected by text length. Malvern, Richards, Chipere and Durán (2004) acknowledge that D (as calculated by *vocd*'s default method) can be affected by text length, but claim that the effects are not significant for the ranges of text lengths with which they are concerned. In this paper, we explain why D is affected by text length, and demonstrate with an extensive empirical analysis that the effects of text length are significant over certain ranges, which we identify.

I Introduction

Lexical diversity (LD) can be described as the range and variety of vocabulary deployed in a text by either a speaker or a writer.[1] LD is used by researchers in many fields as it has been found to be indicative of a wide variety of variables, such as writing quality, vocabulary knowledge, general characteristics of speaker competence, and even a speaker's socioeconomic status (Avent and Austermann, 2003; Carrell and Monroe, 1993; Grela, 2002; Ransdell and Wengelin,

Address for correspondence: Phil McCarthy Ph.D., Institute for Intelligent Systems (IIS), 4th Floor FedEx Institute of Technology (rm 410), 365 Innovation Drive, University of Memphis, TN 38152, USA; email: pmmccrth@memphis.edu

[1]As opposed to the potential vocabulary that a speaker or writer may have available but is not currently using.

Language Testing 2007 24 (4) 459–488 10.1177/0265532207080767

2003). Often used interchangeably with the terms lexical variation (Read, 2000), lexical richness (Tweedie and Baayen, 1998), and vocabulary richness (Hoover, 2003), LD is a phenomenon that is measured by a wide variety of indices, each of which offers a specific, verifiable, and objective score of lexical deployment. Although LD indices are among the most applied tools in lexical analysis, the formulation of a fully valid and reliable LD measure has proven to be elusive (Jarvis, 2002; Tweedie and Baayen, 1998).

The most commonly discussed reason for LD's elusiveness can best be explained by reference to Heaps's law (Heaps, 1978). When applied to LD, Heaps's law predicts that the more words (tokens) a text has, the less likely it is that new words (types) will appear. Thus, the first few words of any given text are likely to be new types, whereas more subsequent tokens are likely to represent types that have been used before. The diminishing returns of new types flaw the most commonly used LD metric, the type-token ratio (TTR, Chotlos, 1944; Templin, 1957), which is formed simply from the division of the number of types by the number of tokens. Thus, when TTR is used to compare any two texts, the longer text generally *appears* to be less diverse (Arnaud, 1984; Kucera and Francis, 1967; Linnarud, 1986).

For over 60 years, researchers have tried to establish an index of LD that is not affected by text length – or, in other words, a measure of LD that accurately *projects* the LD trend of a text so that texts of differing lengths can be reliably compared (Honoré, 1979; Michéa, 1969; Yule, 1944). These attempts are nearly always some mathematical conversion of the relationship between types and tokens, such as using the square root or the log of the number of tokens (Dugast, 1979; Guiraud, 1960). These variations have been widely tested and all have failed to project adequately the LD trend of a text (Jarvis, 2002; Malvern *et al*., 2004; Tweedie and Baayen, 1998). Despite these clear and well-promulgated failings, however, the great need to analyze the lexical diversity of texts means that we continue to see a steady stream of publications using these unreliable indices (Ertmer *et al*., 2002; Miller, 1981; Ratner and Silverman, 2000; Smith and Kelly, 2002).

Recently, a new LD computational measuring instrument, known as *vocd* (MacWhinney, 2000), has offered hope that LD reliability is, after all, achievable (Malvern and Richards, 1997). The *vocd* program outputs an LD index that is calculated through a series of TTR samplings and curve fittings. Following the inclusion of *vocd* as the LD measuring tool on the widely used Computerised Language

Analysis (CLAN) suite of programs (freely available at http://childes.psy.cmu.edu) and a high-profile monograph dedicated to the index (Malvern *et al*., 2004), *vocd* appears to be steadily becoming the LD index of choice for researchers and students alike. And indeed, initial results of *vocd* appear promising (Owen and Leonard, 2002; Malvern *et al*., 2004) with some researchers (e.g. Harris Wright *et al*., 2003; Silverman and Bernstein Ratner, 2000) appearing already to be treating *vocd* as the industry standard. Closer inspection of *vocd*, however, reveals theoretical and empirical problems that we believe are deserving of discussion. It is the purpose of this paper to explicate those issues. To this end, we have organized our study into two main sections. In Section II, we use theoretical and empirical evidence to show that the sampling and curve-fitting procedures employed by *vocd* are not only redundant, but also negatively affect *vocd*'s precision of measurement. Our analysis in this section also challenges the primary claim of the creators of *vocd* (Malvern and Richards, 1997; Malvern *et al*., 2004) that their index is not significantly affected by text length. In Section III, we follow up on this finding with a much larger corpus and a more in-depth analysis, in order to evaluate *vocd*'s sensitivity to text length in relation to that of 13 other indices of LD. Our results show that *vocd* is indeed significantly affected by text length, but we also indentify 'stable ranges' within which texts of differing lengths can be reliably compared using *vocd*.

II *vocd* and the effects of probabilities

The purpose of this section is to discuss and demonstrate how *vocd* works, what it actually measures, how reliably it measures this, and whether it overcomes the text-length dependency problem that it was designed to solve. We will do this with empirical data comprising 266 narratives written in English by Finnish-speaking ($n = 132$) and Swedish-speaking ($n = 69$) adolescent learners of English living in Finland and by adolescent native English speakers ($n = 65$) living in the USA. The lengths of the texts range from 51 to 578 tokens ($M = 225.65$; $SD = 98.50$).

1 What vocd *does*

The *vocd* program is capable of calculating LD through either random sampling or sequential sampling. The random sampling option,

which the creators of *vocd* hold to be the superior option (Malvern *et al*., 2004: 63–75), is the default method used by *vocd*, and is also the method used in nearly all investigations of LD that use the *vocd* program. In this paper, therefore, we focus solely on *vocd*'s default method of *random sampling without replacement*.

The *vocd* program's default procedure operates as follows. First, *vocd* estimates a text's level of LD by taking 100 random samples[2] of 35 tokens drawn from the text and calculating a mean TTR for these samples. This procedure is repeated for samples of 36 tokens, 37 tokens, and so on, all the way to samples of 50 tokens. The program then plots the mean TTR values for each sample size in order to create a random-sampling TTR curve for the text. Figure 1 shows a random-sampling TTR curve for one of the texts from our database, which was written by an American fifth grader (125 types, 452 tokens). It is important to note that *vocd*'s output is in the form of numeric tables instead of graphs; for the benefit of the reader, we used *vocd*'s numeric output to create Figure 1.

After the mean TTR values for random samples of 35–50 tokens are plotted, *vocd* uses a formula known by its $\mathcal{D}$ coefficient (see

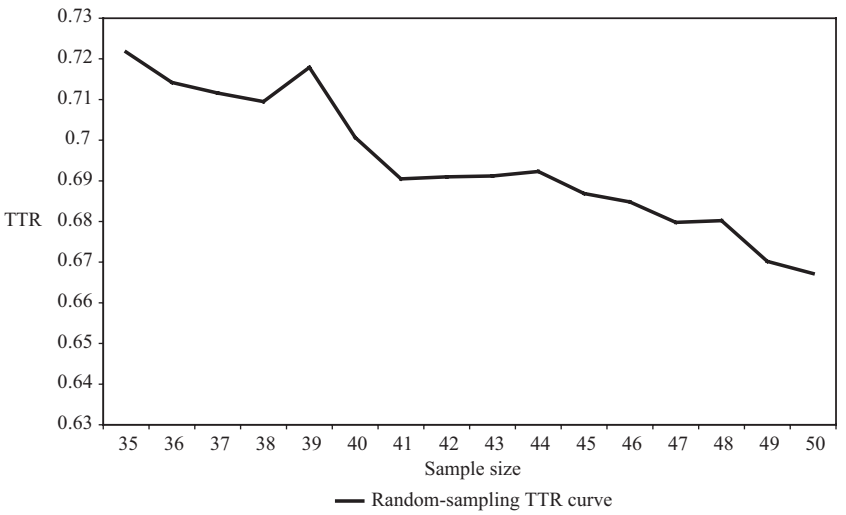

Figure 1 Mean TTR values for random samples of 35–50 tokens

[2]Malvern *et al*. (2004) state that the number of trials for each sample was 'fundamentally a pragmatic decision' (p. 55) where 100 trials proved to be 'sufficiently consistent' (p. 55).

Malvern *et al*., 2004: 51) to produce a theoretical curve that most closely fits the random-sampling TTR curve. The value of $\mathcal{D}$ that provides the best fit between the theoretical curve and the random-sampling TTR curve is referred to as optimal $\mathcal{D}$, $\mathcal{D}_{best\ fit}$, or more simply as D (2004: 56). Figure 2 shows the fit between the random-sampling TTR curve of the text mentioned earlier and the best-fitting theoretical $\mathcal{D}$ curve. In this case, D is 33.36, as can be seen in the following line from *vocd*'s output:

D_optimum < 33.36; min. least sq. val. = 0.000 >

This line also shows that *vocd* outputs a minimum least squared value (which is the sum of the squared distances between the two curves across all sample sizes) to help the user to assess the goodness of fit between the random-sampling TTR curve and the theoretical $\mathcal{D}$ curve. However, it is doubtful that most users will know how to make sense of this information. The line of output from *vocd* that says that the min-imum least squared value is 0.000 suggests that the fit between the two curves is perfect. Unfortunately, *vocd* does not output visual or numeric information comparing the two curves or allowing the user to see, as we can in Figure 2, that the fit is not perfect. Another important piece of information that is missing from *vocd*'s output is an inferential statistical test to indicate whether the fit between the two curves is close enough.

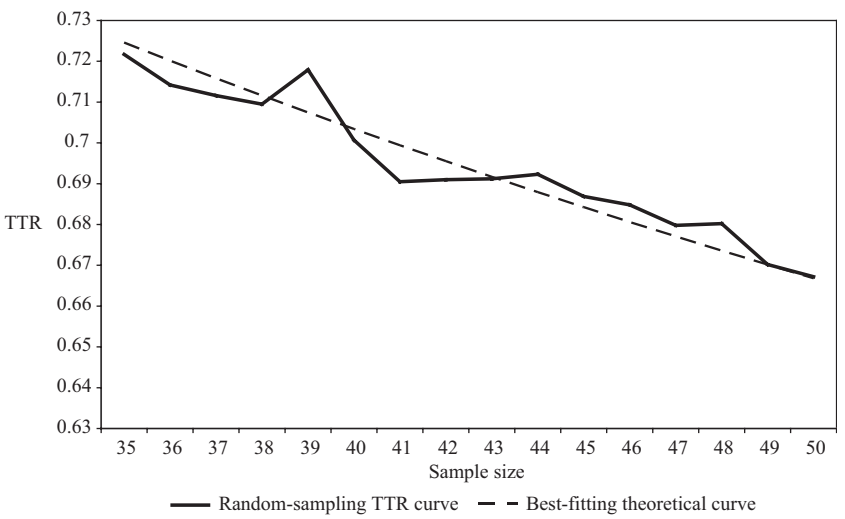

Figure 2 Curve fitting

The creators of *vocd* recognized that the nature of *vocd*'s random sampling method would cause it to arrive at a slightly different D value each time the program is run.[3] To address this problem, the program runs through the entire process of random sampling and curve fitting three times for each text, calculating D three times, and giving the average of the three Ds as the text's final index of LD. Using the average of three iterations does not completely solve the problem of obtaining a different final D score each time the program is run, but it does result in a higher level of consistency (Malvern *et al*., 2004: 56–57). The three preliminary D scores produced by *vocd* for the text mentioned earlier were 33.36, 33.04, and 33.78. The average of these three values is 33.39, and this is the final D score that *vocd* output as this text's index of LD. As D scores generally range between 10 and 100, with a higher D indicating a more diverse text (McKee *et al*., 2000), we can presume this particular text is relatively low in diversity.

2 *What* vocd *measures*

The fact that *vocd*'s final output is an index referred to as D suggests that the $\mathcal{D}$ formula is used to determine the value (i.e. the level) of LD in a text. However, it is important to distinguish between value and currency, and we will show in this section that the value of LD is determined by probabilities of word occurrence, and that the $\mathcal{D}$ formula in *vocd* simply serves to convert this value to a different currency (i.e. a different scale). We will also show that the random-sampling and curve-fitting procedures used by *vocd* introduce a certain degree of noise (or imprecision) into the currency conversion, which results in final LD indices (i.e. D scores) that are not fully precise.

To begin with the random-sampling procedure, recall that *vocd* calculates a mean TTR for 100 random samples of 35 tokens drawn from a text, then a mean TTR for 100 random samples of 36 tokens, and so on, all the way to random samples of 50 words. Now, using the mean TTR for 100 random samples of r number of words drawn from a text will result in an approximation of the mean TTR for *all possible combinations of r number of words* drawn from the text. This is precisely what the random-sampling procedure in *vocd* approximates, and the creators of *vocd* appear to have resorted to random sampling instead of 'exhaustive' sampling for reasons of

[3]Malvern *et al*. (2004) refer to the differences in output as 'stochastic'.

feasibility (cf. MacWhinney, 2000). In reality, however, it is just as feasible to calculate TTRs for all possible combinations of words as it is to calculate TTRs on the basis of several hundred random samples. We will show how this is done in order to illustrate the principles of probability that determine the LD values that *vocd* outputs.

For convenience, we will use the notation ATTR-35 to mean the average TTR for all possible combinations of 35 words drawn from a text. Likewise, ATTR-36 will mean the average TTR for all possible combinations of 36 words, and so forth. To calculate ATTR-35 for a text, it is necessary to consider the contribution that each word in the text makes to the TTR of a sample of 35 words. Whenever a new word occurs in a sample of 35 words, its contribution to the TTR of that sample is 1/35 (i.e. one type divided by 35 tokens), which equals 0.02857. This is true regardless of how many tokens of the word occur in the sample because multiple tokens of the same type still count as only one type. In order to calculate the contribution of a given word to ATTR-35, what we also need to know is the *probability* of the word's occurrence (at least once) in any given sample of 35 words.

Let us consider the contribution that the word *the* – occurring 10 times in a text of 100 words – would make to ATTR-35 for that text. The probability of finding x tokens of *the* in a sample of 35 words drawn from the text can be calculated with the hypergeometric distribution function. The formulas associated with this function would be somewhat cumbersome to reproduce and discuss here, so we refer the interested reader to Wu (1993). We will demonstrate how this works with the HYPGEOMDIST function in Microsoft Excel. The function is of the following form: = HYPGEOMDIST(x, r, n, N), where N is the text length (100, in this case), n is the number of tokens of a particular type in the full text (in this case, there are 10 *the*'s), r is the sample size (35, in this case), and x is the number of tokens of the particular type whose probability in the sample we want to find out about. In order to determine the probability of the occurrence of *the* at least once in the sample of 35 words, we would have to add its probability of occurring once (i.e. x = 1) with its probability of occurring twice (x = 2) with its probability of occurring three times (x = 3), and so forth, all the way to its probability of occurring n times (x = 10). An easier way, however, is simply to calculate the complement of its probability of non-occurrence (x = 0). Thus, when x = 0, r = 35, n = 10, and N = 100, the HYPGEOMDIST function outputs the value of 0.01034, and this is the probability of not finding a single *the* in a sample of 35 words drawn

(without replacement) from the text in question. The probability of finding at least one *the* in the sample of 35 is the complement of this number; i.e. 1 – 0.01034 = 0.98966. This value shows that there is roughly a 99% chance that *the* will occur in any randomly drawn sample of 35 words from the text. Another way of saying this is that *the* will occur (with at least one token) in 98.966% of all possible combinations of 35 words drawn from the text. The contribution of *the* to ATTR-35, therefore, is its contribution to the TTR of a single sample (i.e. 1/35) multiplied by the proportion of samples in which it will occur if all possible combinations of 35 words are taken into account: (1/35) × 0.98966 = 0.02828. The final ATTR-35 value for the whole text is determined by summing the contribution of *the* with that of all of the other words (types) in the text.

Using a simple program that we scripted ourselves using HyperTalk, we calculated ATTR-35, ATTR-36 … ATTR-50 for all 266 texts in our database. Figure 3 shows what the ATTR results of a single text look like when compared with those of the random-sampling output of *vocd*. As one can see, the ATTR curve – the curve that represents the average TTR for all possible combinations of 35, 36 … 50 words drawn from the text – is a perfectly smooth, gradually descending curve. The same probabilistic principles that determine the slope and curvature of the ATTR curve also underlie the random-sampling curve. The jaggedness (i.e. peaks and valleys) of the random-sampling curve is merely a

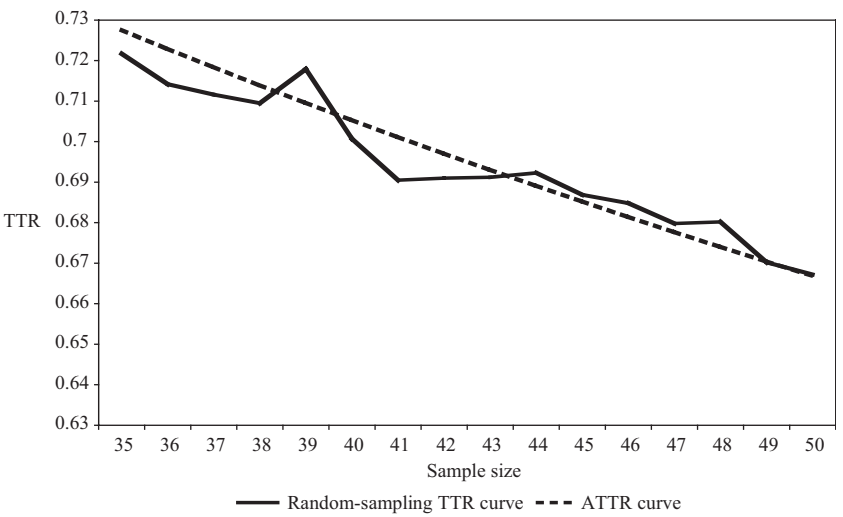

Figure 3 Random-sampling TTR curve versus ATTR curve

coincidence of incomplete and accidentally non-uniform sampling. Both curves measure the same thing; the random-sampling curve approximates what the ATTR curve measures with exactness. Figure 4 shows the same two curves, along with the best-fitting theoretical D curve that was shown earlier in Figure 2. Most striking about Figure 4 is the fact that the theoretical D curve and the ATTR curve are so close and so similar. It is relevant to point out, however, that the two curves do differ slightly in terms of both slope and curvature, with the result that they intersect rather than run parallel courses. The relevance of this observation will become clear.

We ran a series of Pearson bivariate correlation tests on the following indices for each of the 266 texts in our database: final D scores output by *vocd*, ATTR-35 … ATTR-50. The results of these tests are shown in Table 1, although to save space, we removed the results for all variables except D, ATTR-35, ATTR-40, ATTR-42 (i.e.

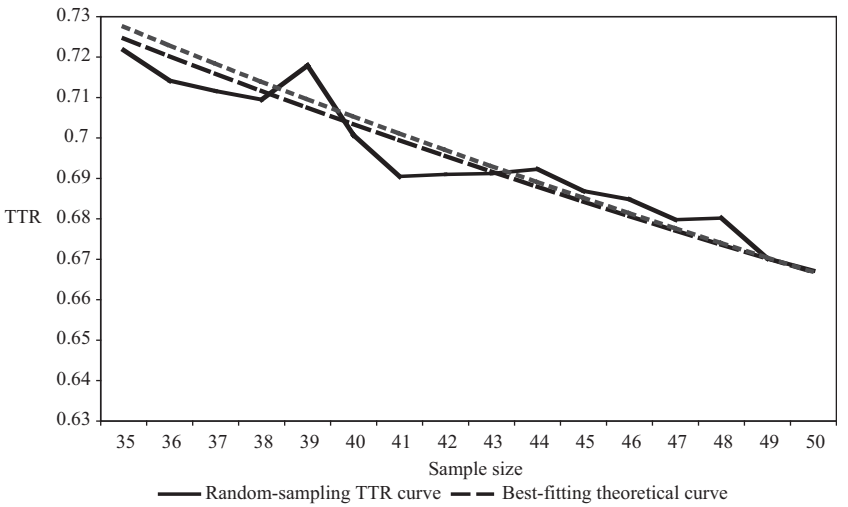

Figure 4 All three curves

Table 1 Pearson correlations for *vocd*, ATTR-35 to ATTR-50

	D	ATTR-35	ATTR-40	ATTR-42	ATTR-45
ATTR-35	.968				
ATTR-40	.970	.999			
ATTR-42	.971	.999	1.000		
ATTR-45	.971	.998	.999	1.000	
ATTR-50	.971	.995	.998	.999	1.000

the midpoint between 35 and 50), ATTR-45, and ATTR-50. The figures given in Table 1 are Pearson correlation coefficients (r). In each case, $p < .001$ and $N = 266$.

As is evident, all of the correlation coefficients in Table 1 are extremely high, which attests to the fact that all of these variables are measuring almost exactly the same thing. In fact, with correlations this high, where the shared variance (r^2) between variables is 0.94 and higher, the natural conclusion is that these variables themselves are almost exactly equivalent. By extension, we would also conclude that all but one of these variables are redundant. The variable that stands out as the most logical candidate for elimination is D because D is, in effect, simply a measurement of the height of the curve formed by the TTR values. In other words, TTR values are prior to and of a more fundamental nature than D is. Also, the fact that each one of the ATTR variables correlates almost perfectly with D shows that any one of them – i.e. the average TTR associated with any one sample size between 35 and 50 – could be used as the text's measure of LD. The *vocd* program's curve-related procedures are therefore redundant.

The fact that D's correlation with ATTR values is only *almost perfect* can be attributed to noise associated with the imprecision of random sampling and noise associated with the fact that $\mathcal{D}$ curves are not perfectly parallel to ATTR curves. As can be inferred from Figure 4 and Table 1, $\mathcal{D}$ curves tend to intersect ATTR curves somewhere between sample sizes of 42 and 50, and this is why D correlates best with ATTR values that represent sample sizes within this range. Our point is not to criticize *vocd* for not producing output that would correlate perfectly with ATTR values. In fact, we acknowledge that *vocd*'s output is remarkably consistent with ATTR values. Our point is that despite the fact that *vocd*'s creators have argued repeatedly and at length for the usefulness of the $\mathcal{D}$ formula and curve fitting in the measurement of lexical diversity (e.g. Malvern and Richards, 1997; Malvern *et al*., 2004; McKee *et al*., 2000), the role that $\mathcal{D}$ and curve fitting have been given in *vocd* turns out to be trivial. Although curve fitting does have the effect of smoothing (or averaging) the peaks and valleys of the random-sampling curves (see Figure 2), there are much more efficient ways of achieving this same effect or even making it completely unnecessary, such as when ATTR values are used in place of random-sampling values. Again, *vocd*'s use of the $\mathcal{D}$ formula effectively only converts random-sampling TTR values to D values. The values themselves are determined by probabilities, and the $\mathcal{D}$ formula simply changes these values to a different scale.

A few more words about the usefulness of the $\mathcal{D}$ formula are in order. First, this formula is an idealized mathematical model of the relationship between TTR and the number of tokens in a text as the text grows longer. Studies discussed by Malvern *et al*. (2004), including Jarvis (2002), have shown that the $\mathcal{D}$ formula produces idealized curves that are remarkably similar to (and statistically good fits for) the TTR growth curves of most texts (oral or written) that are relatively short (up to a few hundred words). Through curve fitting, a single value of D can be obtained that represents the general height of a TTR growth curve, and this allows the researcher to be able to compare the TTR levels of texts of different lengths – provided that a sufficiently good fit can be found between the best-fitting $\mathcal{D}$ curve and the TTR growth curve. Knowing what to do with texts for which sufficiently good fits cannot be found is a problem for researchers using this method of LD measurement, but the rationale itself for using curve fitting to compare the TTR growth curves of texts of different lengths seems solid (Jarvis, 2002). However, TTR growth curves are very different from random-sampling curves and ATTR curves. Besides the fact that they do not have the same slopes or curvatures, the main point for present purposes is that, at each point on a TTR growth curve, what is represented is simply the TTR of the words that have occurred in the text up to that point. In random-sampling and ATTR curves, on the other hand, each point on the curve reflects the probabilities of occurrence for *every word in the text*. Thus, with random sampling and ATTR, any point on the curve can be used to represent the LD of the entire text, which makes the use of curves and curve fitting redundant. This point is underscored by the results shown in Table 2, which indicate that ATTR-35, ATTR-50, and all points in between these values correlate almost perfectly (i.e. $r > .96$) with D.

So, what is it that *vocd* measures? As it turns out, neither D nor TTR is the essence of what it measures. Although it *is* true that the random-sampling procedure in *vocd* does technically calculate

Table 2 Bivariate correlation results between D and several ATTR and SOP values

	D	ATTR-35	SOP-35	ATTR-42	SOP-42	ATTR-50
ATTR-35	.968					
SOP-35	.968	1.000				
ATTR-42	.971	.999	.999			
SOP-42	.971	.999	.999	1.000		
ATTR-50	.971	.995	.995	.999	.999	
SOP-50	.971	.995	.995	.999	.999	1.000

type-token ratios, those TTRs are averaged over multiple samples, so what ultimately determines the final mean TTR for a set of samples is the sum of probabilities of occurrence for each word in a text. Recall our discussion toward the beginning of this section where we described how ATTR-r (ATTR-35, ATTR-36, etc.) is determined. To remind the reader, this is done by summing the probability of occurrence for each word multiplied by 1/r (or one divided by the sample size). Multiplying the probability of occurrence for each word by 1/r is what converts the probability into a TTR. This is an arbitrary conversion, however, and the real values underlying *vocd*'s output are sums of word probabilities. To show that this is the case, we ran a series of bivariate correlation tests between D values, ATTR values, and sum-of-probabilities (SOP) values. The correlation coefficients (r) from these tests are shown in Table 2. In each case, $p < .001$ and $N = 266$. Note that each pair of ATTR and SOP values at equivalent sample sizes (e.g. ATTR-35 and SOP-35) correlates perfectly, as the former is a function of the latter. Note also that the correlations between D and SOP values are the same as between D and the corresponding ATTR values. Again, these results show that D, ATTR, and SOP are equivalent measures. Given that D and ATTR ultimately derive their values from SOP and that they do so with perfect (in the case of ATTR) or near perfect (in the case of D) reliability, there is little question that SOP is the essence of what *vocd* measures.

To take stock, the four crucial points we have attempted to make in this section are (1) that SOP is the essence of what *vocd* measures, (2) that *vocd*'s D scores are ultimately determined by SOP, (3) that *vocd*'s random-sampling and curve-fitting procedures prevent the program's output from being perfectly precise and perfectly consistent and also make *vocd* an inefficient tool in terms of the economy of its algorithms (though we acknowledge that *vocd* is fast), but (4) its output is nevertheless a highly reliable indicator of SOP. This fourth point may seem to redeem *vocd*, but this depends on whether SOP is really what should be measured. We could, for example, just as easily measure mean probabilities (i.e. the mean probability of occurrence for each word in a text), and in fact our own preliminary tests have shown that mean probabilities correlate more highly with certain variables, such as writing quality[4] and the Shannon-Wiener

[4]The essays used in this analysis were given holistic quality ratings by two experienced ESL composition raters in an intensive English program at an American university. The raters' inter-rater reliability was .97.

diversity index (Shannon, 1948), than SOP does. These preliminary findings do not prove that mean probabilities are a more valid measure of LD than SOP is (cf. Jarvis, 2002: 81), but they do remind us that the construct validity of D and all other measures of LD will remain suspect until the field has arrived at a fully adequate theoretical construct definition of lexical diversity that could serve as the basis for any measure's validation. It is not enough to be able to show what an LD tool *does* measure; we also need a theoretically adequate explanation of what an LD tool *should* measure.

3 Probabilities and the text-length dependency problem

D was originally proposed by Malvern and Richards (1997) as a measure designed to overcome the text-length dependency problem that TTR and all other indices of LD appear to suffer. This has remained the goal with *vocd*'s instantiation of D, although Malvern *et al*. (2004) are cautious in their claims about how successful *vocd* is in this regard:

> We have claimed that by using a mathematical modeling procedure on a standard-sized window of 35 to 50 tokens of the TTR versus token curve, lexical diversity values will no longer be a *function* of text length. The use of the word 'function of' is crucial here, as we do not mean to imply that there is no *relationship* between lexical diversity values and the total number of words.
>
> (2004: 64, italics in the original)

Malvern and colleagues go on to show through empirical evidence that, although some of *vocd*'s optional methods for measuring LD are significantly affected by text length, its default method of random sampling without replacement (i.e. the method we are concerned with in this paper) does not vary significantly by text length for texts that are no longer than a few hundred words in length. Malvern *et al*. do nevertheless report that this method shows at least a slight sensitivity to text length ($\eta^2 = 0.045$) (2004: 66–67). As such, their claim appears to be that D actually is sensitive to text length, but not enough to matter for the lengths of texts that they are concerned with. In the present section we confirm the first part of their claim by showing that because *vocd*'s default method is fundamentally a measure of SOP, principles of probability make it inevitable that D will increase as text length increases. We will address the latter half of their claim (i.e., that D's sensitivity to text length is insignificant for short texts) in an in-depth analysis in Section 3.

We will illustrate the effects of text length on SOP (and consequently also on D) by looking first at a sample text from our database. For convenience, we have chosen a text that is 100 words (tokens) long. This particular text was written by a Finnish-speaking seventh-grader, and it is shown in its entirety below. (Note: Misspellings have been corrected):

> The girl was stolen a bread, but Chaplin take a fault. So, they are going to put a Chaplin in jail. But a woman saw that the girl stolen the bread, so they take Chaplin and the girl. When they was going to jail the girl and Chaplin ran away. They go to sit and talk about where they live. The girl say that she live no where. So they dreaming a home where they can take orange in the tree and eat a breakfast in a kitchen. The Chaplin say that do the work then we got a house.

Defining types as lexemes – as *vocd* does by default – we find 48 types in this 100-token text. The most frequently occurring types in the text are *the* (n = 9), *a* (n = 8), *they* (n = 7), *girl* (n = 5), and so forth. To save space, we will not list all of the types with their accompanying frequencies, but Table 3 below is a truncated list of these types that shows the probabilities and associated contributions to SOP-42 and ATTR-42 for each of these types. (We chose r = 42 because 42 falls in the middle of the range of sampling sizes used by *vocd*, and SOP-42 and ATTR-42 are also the two variables that correlate most highly with D and with all of the other values of SOP and ATTR.)

Table 3 Word types, their frequencies, and their contributions to SOP-42 and ATTR-42

Type	n	N	Contribution to SOP-42 (i.e. probability of occurrence)	Contribution to ATTR-42 (i.e. probability × 1/42)
the	9	100	0.9944	0.02368
a	8	100	0.9897	0.02356
they	7	100	0.98122	0.02336
girl	5	100	0.93914	0.02236
...				
take	3	100	0.80918	0.01927
...				
live	2	100	0.66606	0.01586
...				
talk	1	100	0.42	0.01
...				
			SUM (SOP-42) = 27.67895	SUM (ATTR-42) = 0.65902

Now, consider what would happen if the Finnish seventh-grader who wrote this text had written one word more. The first consequence would be that N would change to 101, and this would affect all of the probabilities and associated values shown in Table 3. If the added word represented the introduction of a new type into the text, such as the word *eventually*, then the numbers in Table 3 would change to what is shown in Table 4.

By comparing Tables 3 and 4, one can see that the probabilities of occurrence for all of the types already existing in the text drop slightly as the text becomes one word longer. However, the added probability of the new word (i.e. *eventually*) causes the text's SOP-42 to increase by a value of 0.2248, and causes the text's ATTR-42 to rise by a value of 0.0054. A rise in these two values is to be expected, of course, and a simple TTR would show the same trend. The original text has a simple TTR of 48/100 = .48, and the addition of the word *eventually* would cause the TTR to change to 49/101 = .4851, representing an increase of 0.0051, which is very close to the increase of 0.0054 in ATTR-42. The story is quite different, however, when we look at the effects that word repetition has on simple TTR versus ATTR-42. With simple TTR, the repetition of any already-existing word would result in a TTR of 48/101 = .4752, which represents a decrease from the original text's simple TTR by a value of 0.0048. The decrease in simple TTR with the repetition of

Table 4 Changes to SOP-42 and ATTR-42 with the introduction of a new type

Type	n	N	Contribution to SOP-42 (i.e. probability of occurrence)	Contribution to ATTR-42 (i.e. probability × 1/42)
the	9	101	0.99398	0.02367
a	8	101	0.98903	0.02355
they	7	101	0.98017	0.02334
girl	5	101	0.9368	0.0223
...				
take	3	101	0.80493	0.01916
...				
live	2	101	0.66119	0.01574
...				
talk	1	101	0.41584	0.0099
...				
**eventually*	1	101	0.41584	0.0099
			SUM (SOP-42) = 27.90377	SUM (ATTR-42) = 0.66438

an existing word is therefore almost as high as the increase in simple TTR with the introduction of a new word into the text. This is the fundamental problem with TTR because, after the first few tokens of a text, repetitions tend to occur more frequently than new words, which causes the simple TTR to fall more or less continually as the text grows longer. The method of measurement employed by *vocd* is intended to compensate for the overly strong effect of word repetition on TTR, but the end result is an overcompensation. Not only that, but because of the principles of probability, the effect of a word repetition on SOP and ATTR (and therefore also *vocd*'s D output) differs depending on how many tokens of that type already exist in the text. Table 5 shows what the effects would be if instead of adding the word *eventually* to his text, the Finnish speaker mentioned earlier were to add the word *the*, *a*, *they*, *girl*, *take*, *live*, or *talk* as the 101st word of his narrative.

As Table 5 shows, the repetition of any of these words would result in an increased probability of that word being selected at least once in a random sample of 42 words drawn from the text. However, the extent of the increase depends on how many tokens of that type exist in the text. The fewer the tokens (i.e. the smaller the n), the greater the increase in the word's probability of occurrence. Even though the probabilities of all of the other words in the text will decrease slightly when a repeated word's own probability increases, the overall change in the sum of probabilities (i.e. SOP-42) is quite small. The biggest change to SOP-42 occurs when the word with the highest number of tokens, *the*, goes from nine to ten tokens, with the result that the SOP-42 of the text decreases by 0.18828. This is the largest decrease in SOP-42 that can happen to this text with the addition of a 101st word, but even this decrease is smaller than

Table 5 Differential changes to probabilities, SOP, and ATTR depending on *n*

Type	n	Change to the word's probability of occurrence	Change to the SOP-42 of the text	Change to the ATTR-42 of the text
the	9→10	+ 0.00233	− 0.18828	− 0.00448
a	8→9	+ 0.00428	− 0.18607	− 0.00443
they	7→8	+ 0.00781	− 0.18216	− 0.00434
girl	5→6	+ 0.02531	− 0.16337	− 0.00389
take	3→4	+ 0.07935	− 0.10742	− 0.00256
live	2→3	+ 0.13887	− 0.04728	− 0.00113
talk	1→2	+ 0.24119	+ 0.05432	+ 0.00129

the increase of 0.2248 that would happen if a completely new type (e.g. *eventually*) were added to the text. The repetition of types that have fewer tokens results in smaller and smaller decreases in SOP-42, until we arrive at the most surprising and profound finding of all, which is that *the repetition of a one-token type actually increases SOP-42, and consequently also ATTR-42 and D*! The fact that D would increase when a word is repeated seems contrary to what we would expect an LD measure to do, and also quite clearly seems to suggest that D does indeed overcompensate for TTR's sensitivity to text length.

To confirm that this method of measurement truly does exhibit text-length effects in our own database, we calculated ATTR-42 for the first 50 through the first 200 tokens of each of the 141 texts in our database that are longer than 200 words in length. (We used ATTR instead of SOP because the ATTR scale, which ranges from 0 to 1.00, is easier to interpret, and we used ATTR instead of D because ATTR is more precise and consistent.) Given that it is not feasible to illustrate the ATTR-42 growth curves for all 141 texts, we have chosen instead to show a single curve representing the central tendency of these texts. This is shown in Figure 5. This figure shows that after a slight dip in ATTR-42 from about the 50th to the 100th token, the central tendency is for the ATTR-42 of texts to move gradually and steadily upward. The mean ATTR-42 for the

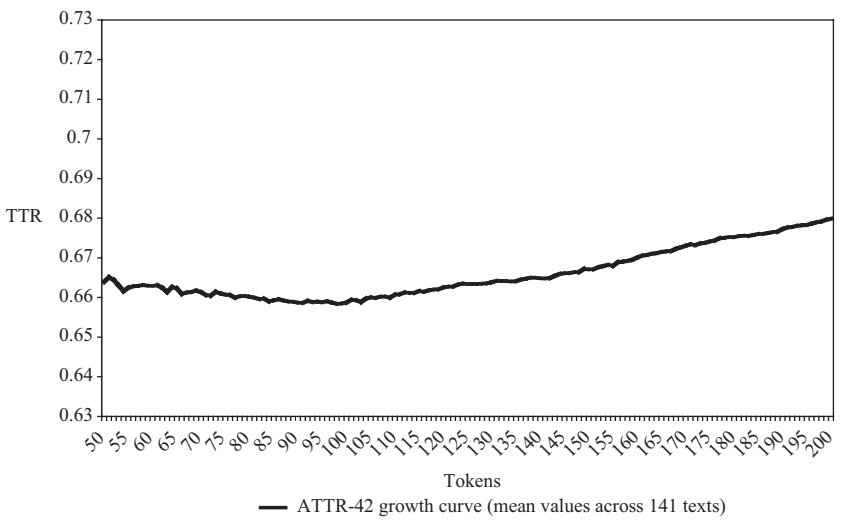

Figure 5 Mean ATTR-42 values as texts grow longer

first 50 tokens is 0.6640 (SD = 0.06052), and this climbs to 0.6799 (SD = 0.03395) at token 200. Although this increase is relatively small, a paired-samples t-test confirmed that the increase is significant ($t(140,1) = -3.783$, $p < .001$, $\eta^2 = .093$). This contrasts with the findings of Malvern *et al*. (2004), which were based on a smaller sample of texts ($N = 38$; $\eta^2 = .045$). Consequently, our results do not support the claim that the effects of text length on D are not significant (p. 67). Our own conclusion is that such a result clearly warrants further investigation. And indeed, this is the purpose of the following section.

III An in-depth analysis of *vocd*'s text-length dependency problem

1 The need for rigorous empirical testing

Few could doubt that a rigorous protocol of testing should precede acceptance of any tool or procedure as an industry standard. But it is reasonable to ask whether such rigor is necessary for evaluating the efficacy of something as inert as LD. To this we assert that such rigor is necessary, and we offer two important reasons for believing so.

First, LD is a phenomenon of great importance to numerous and widely varied fields of scientific inquiry. For example, LD is used to analyze texts in stylistics (Smith and Kelly, 2002), neuropathology (Bucks *et al*., 2000), language acquisition (Singh, 2001), and even forensics (Colwell *et al*., 2002). Without reliable tools, procedures, and analyses, researchers cannot be expected to produce accurate results, and if results are unreliable, then predictions based on such research will be unreliable. A thoroughly tested, reliable, and verifiable lexical analysis tool is the only way to ensure confidence in the conclusions drawn from experimental results and the future predictions such conclusions will generate.

Second, as evidenced above, LD is used not merely for rating learners' language abilities and differentiating genres of texts. LD can also be used to help determine the course of treatment for Alzheimer's patients (Bucks *et al*., 2000), for patients with specific language impairment (Thordardottir and Weismer, 2001), and for children with cochlear implants (Ertmer *et al*., 2002). Creators and testers of any LD index, therefore, need to accept the responsibility for how these tools might be implemented, for as Malvern *et al*. (2004) comment, 'these things matter' (p. 180).

2 *Historical norms for LD-testing corpora*

With such reasoning in mind, the rigor of recent studies that have tested existing LD indices may cause some concern if the degree of faith we can put in any LD index is to be considered. The evidence (see Table 6) suggests that corpora used for testing have perhaps raised more questions then they have answered. For example, none of the studies mentioned in Table 6 examined LD indices over more than one genre, none tested indices over both spoken and written modes, and while the Tweedie and Baayen (1998) corpus hosts a considerable 1 126 240 words, it consists of a mere 16 novels: hardly a representative sample of literature, let alone language.

There is, of course, no doubt that each of these studies contributed greatly to our understanding of LD. Each study was also effective in raising concerns over existing LD indices. Furthermore, we must recognize that none of the studies in Table 6 actually set out with the sole purpose of establishing absolute confidence in any particular index. But even allowing for each of these points, we feel that that lack of scale and scope in LD comparative studies is cause for concern. We also acknowledge that if such tests have been the standard by which LD indices have been historically tested, then it is not unreasonable for the new indices (such as *vocd*) to begin their testing in similar vein. However, with the ease that modern technology affords us to construct large corpora, we feel it reasonable that any LD index (and in this case, *vocd*) should be able to easily establish a high degree of validity by being tested under far sterner circumstances. In this section, we conduct just such testing.

It is not just the size of the LD-testing corpora that should be considered. When testing a commodity's worth, it may be beneficial to show that the new product is significantly better than the best available

Table 6 Major studies of LD reliability since 1993

Study	Texts	Words	Genres	Mode
Tweedie and Baayen (1998)	16	1 126 240	1	Written
Rietveld and Van Hout (1993)	3	89 307	1	Written
Bucks *et al.* (2000)	24	24 000	1	Spoken
Jarvis (2002)	276	82 800	1	Written
Owen and Leonard (2002)	91	45 500	1	Spoken
Harris Wright *et al.* (2003)	18	3600	1	Spoken
Daller *et al.* (2003)	42	8067	1	Spoken
Silverman and Bernstein Ratner (2002)	15	6404	1	Spoken
McKee *et al.* (2000)	38	12 008	1	Spoken

alternatives. With this in mind, it is worth noting that *vocd* has generally been tested against TTR. Silverman and Bernstein Ratner (2002), for example, tested *vocd* only against TTR, and the same applies to McKee *et al*. (2000) and Owen and Leonard (2002). As most researchers view TTR as the least reliable assessor of LD (e.g. Daller *et al*., 2003; Rietveld and van Hout, 1993), it is odd that more hardy varieties of LD are not more often used for comparison purposes. Jarvis (2002) tested the *D* formula against a more sophisticated measure, U (Dugast, 1978). The results suggested that *D* was no more effective than U at measuring LD. In this section, therefore, we report results of empirical testing where *vocd* is pitted against 13 alternative LD indices.

3 The corpus

For our corpus, we used 15 written genres from the Lancaster-Oslo-Bergen corpus (LOB, Johansson *et al*., 1978) and Brown corpus (Kucera and Francis, 1967), and six spoken genres from the London-Lund Corpus (LLC, Svartvik and Quirk, 1980) and the Wellington Corpus of Spoken New Zealand English (WSC, Holmes *et al*., 1998). We also added a further academic written genre (*Glencoe Science*, Biggs *et al*., 2003) and an academic spoken genre (*Michigan Corpus of Academic Spoken English*, Swales and Malczewski, 2001), making 23 genres in all. More details about the corpus can be found in McCarthy (2005). The bulk of the files from these corpora have previously been used in studies such as Biber (1988), Louwerse *et al*. (2004), and Dempsey *et al*. (2007). As in these studies, we selected a representative sample from the corpora (nine texts from each genre) and kept the overall upper text lengths relatively consistent (2000 tokens).

IV Method

To establish an LD index as an industry standard, we believe that the index has to be able to reconstitute (or *project*) the whole LD score from the various analyses of smaller parts of a text. To this end, we used the traditional standard of dividing an original text into smaller pieces to see whether the smaller elements can project the score of the whole text when reconstituted (Hess *et al*., 1986; McKee *et al*., 2000; Tweedie and Baayen, 1998). Thus if a 2000-token text is split into two 1000-token sections then the mean of the LDs produced

from the two smaller sections should approximate the LD of the original 2000-token text. By splitting the text into smaller and smaller sections, 500-tokens, 400-tokens, 250-tokens, and so on, the reconstituted LD scores begin to form a pattern that indicates the degree to which text length affects the LD measure. Specifically, the more the mean LD score of the section sizes correlates with the mean token size of the section size, the *less* the LD measure is able to satisfactorily project.

For dividing our texts, we followed the parallel sampling method of Hess *et al*. (1986). We made 11 equally sized sections of length 100, 154, 200, 250, 282, 333, 400, 500, 666, 1000, and 2000 tokens.[5] Thus, the same text was represented in 11 different ways. The 11 slices per text multiplied by the nine texts per genre, gave us 99 overall data points, three times as many as the sample size of 30 observations, which is generally acknowledged to be a minimum for correlations (van Genderen and Lock, 1977). In total, then, our corpus contained 16 written genres (WG = 16*9*11 = 1585 data points) and 7 spoken sections (SG = 7*9*11 = 693 data points), which together made 23 overall genres (OG = 23*9*11 = 2277 data points).

Unlike previous studies (e.g. McKee *et al*., 2000; Owen and Leonard, 2002) where *vocd* was tested against the weakest available indices (although unarguably the best known and most widely used ones), we added a number of rival LD indices for the purposes of comparison.[6] Thus, we added RTTR (Guiraud, 1960) and CTTR (Carroll, 1964) as square root correcting LD measures. We added log correcting measures represented by H (Herdan, 1960), U (Dugast, 1978), SS (Somers, 1966), Maas (Maas, 1972), and RK (Rubet's K, Dugast, 1979). We also added measures that adjust for frequency of occurrence of types: M (Michéa, 1969), S (Sichel, 1975), and K (Yule, 1944). And we added the commonly used measure known as W (Holmes and Sign, 1996; Bucks *et al*., 2000). We then added the original calculation of D – first promulgated by Malvern and

---

[5]We agree with an anonymous reviewer who pointed out that comparing parts of a text to a whole text can be problematic because the parts may not contain as much information as the whole. We also agree with the reviewer's concern that 'an unlikely repetition' within a text could actually cause diversity to increase.

[6]We acknowledge that each of these alternative indices has numerous theoretical and empirical flaws of their own. We further acknowledge that some are mathematical equivalents to each other. For an excellent report of the theoretical flaws of a wide variety of alternative indices, the reader is referred to Malvern *et al*. (2004). The purpose for including such indices is that while *all* LD indices are flawed, some are more obviously flawed than others. The point of this particular test is to see to what degree, if any, *vocd* has raised the bar.

Richards (1997) and analyzed in Jarvis (2002) – as a particularly interesting comparison to *vocd*. Finally, we added Raw TTR, the most traditional, simplest and generally criticized measure of LD.

V Results

We used a Pearson correlation to evaluate the effectiveness of projection across the three categories (WG, SG, OG). The results showed that *all* 14 LD measures significantly correlated with text length. That is, none, including *vocd*, overcame the projection problem of text length dependency (see Table 7).

Our results suggest that none of the measures explored in this study is able to satisfactorily project to the degree necessary to claim the title of industry standard. D *(vocd)* is undoubtedly a better performer than most alternative indices, although it is equally clear that it has not outperformed all its rivals.

Despite all indices appearing to be a function of text length, it is possible that some measures are more effective than others over particular lengths of texts. This possibility is supported by results of the variance explained by text length (see Table 8). For some measures, the variance is relatively small (Yule's K and Maas = 2%), while for others the variance is very large (TTR = 67%). Such results at least allow researchers to know which LD measures they would be wisest to avoid.

Table 7 Correlation with text length for all 14 LD measures (n = 2276). All results are r values with $p < .01$

	All genres (OG)	Written genres (WG)	Spoken genres (SG)
RK	0.82	0.86	0.83
CTTR	0.82	0.88	0.8
RTTR	0.82	0.88	0.80
TTR	− 0.77	− 0.81	− 0.77
W	0.66	0.66	0.78
H	− 0.65	− 0.68	− 0.78
S	− 0.57	− 0.63	− 0.73
M	0.56	0.63	0.74
D *(orig.)*	0.55	0.68	0.40
SS	0.44	0.47	0.62
D *(vocd)*	0.22	0.19	0.35
U	− 0.18	− 0.18	− 0.34
Maas	0.15	0.12	0.32
K	− 0.14	− 0.11	− 0.25

Table 8 Variance explained by text length (100–2000 tokens) for all LD measures ($F < 500$)

	n	R^2	F	P
K	2276	0.02	47.76	<0.01
M	2276	0.02	51.61	<0.01
U	2276	0.03	76.73	<0.01
D (*vocd*)	2276	0.05	115.23	<0.01

If measures such as K and Maas have as little as 2% of the variance explained by text length, then perhaps over shorter text comparisons the effect of text length may be negligible. We tested this hypothesis with a one-way ANOVA to track the linear trend of the best performing LD measures over increasing section sizes for the OG condition. This *a priori* test was supplemented with *post hoc* Bonferroni tests to establish where in the section size progression non-significant differences in output could be identified (see Table 9).

The *post hoc* Bonferroni test served to identify where in the 100–2000 word range a significant progression in LD score appeared (see Table 10). For even the best performing indices, the results suggest that reliability is limited to specific and quite short text lengths. K returns the broadest ranges, allowing texts as short as 100 tokens to be compared to texts as long as 500 tokens. D *(vocd)* is second in this analysis, suggesting texts between 100 and 400 tokens might be suitably compared, which is in line with the claims of Malvern *et al*. (2004). We believe that these 'stable ranges' – those ranges within which comparisons of texts can be made without a significant effect of text length – are areas where researchers may meaningfully compare texts.

The analysis described in this section is substantially larger than any other of its kind, and it supports many previous findings on LD projection (Jarvis, 2002; Tweedie and Baayen, 1998). We conclude that all existing LD measures should be used with caution as they are each significantly sensitive to text length. We further recommend that researchers should take note of the LD stability ranges provided in this study. Wright *et al*. (2003), for example, use *vocd* to compare texts produced by a group of patients with aphasia with a control group. The first group's texts range in length from 392 to 655 tokens. The second group's range is from 208 to 587 tokens. Results produced in this study suggest that if Wright and colleagues wished to use *vocd*, then the third range (250–666 tokens) would have provided the most stable results, meaning that all texts under 250 tokens should have been discarded.

Table 9 ANOVA results for contrasts by index across all 2277 section sizes of texts ($F < 79$)

Measure	F	P
K	5.99	<0.01
Maas	11.19	<0.01
D (*vocd*)	13.26	<0.01
U	20.5	<0.01

Table 10 Best ranges for the OG category provided from results of Abonferroni test on the five best performing LD measures

	Range 1	Range 2	Range 3	Range 4
K	100–500	154–666	250–1000	400–2000
D (*vocd*)	100–400	200–500	250–666	400–1000
U	154–250	200–500	254–1000	286–2000
Maas	100–154	154–333	200–666	250–2000
D (*orig.*)	100–200	154–286	200–333	250–400

One reasonable objection to our study is that the text lengths examined are too large for an index such as *vocd*. After all, the creators of *vocd* have always argued that the upper limit of the index is a non-specific 'few hundred' words (Malvern *et al*., 2004). On the other hand, Malvern *et al*. (2004) champion the validity of the *vocd* index by citing such research as Wright *et al*., (2003) where texts well in excess of 600 words were commonplace. We acknowledge that our study has certainly looked above and beyond the recommended limits suggested by *vocd*'s creators; however, as it seems clear that researchers are going to use *vocd* outside its limits (and that *vocd*'s creators are not averse to celebrating such use), it behooves us to inform the community as to the accuracy that researchers can expect when such quantities of text are analyzed.

VI Conclusion

In this paper, we have argued that the ubiquity of LD application establishes its importance to quantitative research. LD is used in fields as varied as *stylistics* (Smith and Kelly, 2002), *neuropathology* (Bucks *et al*., 2000), *language acquisition* (Singh, 2001), and even *forensics* (Colwell *et al*., 2002). Indeed, it is difficult to imagine a

textual analysis application whose scope is broader or whose application is more called upon. The great need to measure this LD phenomenon has led to the introduction of many indices, the latest of which is known as D – the name given to the output of the *vocd* computer program.

We have shown, in this paper, that while *vocd* has received encouraging reports in its early years, a number of theoretical and empirical problems are inherent in its measurement of LD. To this end, we have questioned its sampling and curve-fitting procedures, and we have proposed an alternative approach that could be used to improve the efficiency and precision of *vocd*'s measurement. We have also established a testing protocol for LD indices that is worthy of the technology we now have available to us: a corpus comprising the major modes of communication, multiple genre representation, and text segmentation well in excess of the minimum statistical standards.

While this study has raised a number of concerns about *vocd*'s reliability and about whether it truly measures what it should measure (i.e. its construct validity), we feel strongly that the *stable areas* identified here should be as useful to the creators of *vocd* as they are to researchers who rely on the evidence gleaned from LD. We accept that the limitations of the stable areas will be frustrating; however, we wish to it make clear that they do not ostensibly contradict the claims already made by *vocd*'s creators, and, if anything, they offer empirical evidence whose implications go beyond those claims. For researchers using LD indices, the stable areas should offer substantial help in building confidence in the results procured by such indices. Obviously, as the research of most studies cited in this paper suggests, texts will often have well in excess of a few hundred words and researchers will want to compare texts of very different sizes. For this reason, we must view measures such as *vocd* (at least its current version) as being merely a stop gap, and we encourage researchers to either develop a more robust version of *vocd* or develop new alternative indices that can significantly outperform it. To this end, we do not endorse the current version of *vocd* as an industry standard for measuring LD. Such a crown, we feel, would require an index that avoids text length dependency between at least 100 to 2000 tokens, functions across both major modes of communication, and gives reliable results across a wide range of registers. Ultimately, of course, it would also require an adequate construct definition of LD as well as evidence that the index truly represents the construct. We may even question whether a single index has the capacity to encompass the construct of lexical diversity. That is, a more exhaustive measure of the lexical diversity of a text might include such

properties as *the rarity* and the *semantic similarity* of the tokens used. Future research needs to address such issues.

LD has moved far beyond appraisals of language ability or vocabulary deployment. These days, LD is taken into account in socio-economic appraisals that help to determine entire curricular and methodological approaches to teaching in schools across the nation. LD is helping to develop approaches to forensics where vocabulary deployment may indicate guilt or innocence (Colwell *et al*., 2002). LD is used to help assess medical treatment such as cochlear implants (Ertmer *et al*., 2003) and neuropathological developments in such conditions as dementia (Bucks *et al*., 2000). With such a responsibility on the shoulders of LD, it is incumbent upon researchers in our field to provide the most rigorous of tests of our measures, and indicate clear lines of fallibility so as to provide our fellow scientists with only the best and most reliable indices available. We therefore conclude this paper by offering our agreement to Malvern and colleagues in reference to the need for wide and rigorous testing of measures of LD rather than the simplistic acceptance of the most conveniently available or currently fashionable LD measure:

> These things matter. Much of the research based on flawed measures has significant implications for theory, practice, and policy. It is important therefore that methodological issues of measuring vocabulary richness are understood and that these confusions are cleared up.
>
> (2004: 180)

VII References

Arnaud, P.J.L. 1984: The lexical richness of L2 written productions and the validity of vocabulary tests. In Culhane, T., Klein-Braley, C. and Stevenson, D.K., editors, *Practice and problems in language testing: Papers from the International Symposium on Language Testing*. Colchester: University of Essex, 14–28.

Avent, J.R. and **Austermann, S.** 2003: Reciprocal scaffolding: A context for communication treatment in aphasia. *Aphasiology* 17: 397–404.

Bernstein Ratner, N. 1988: Patterns of parental vocabulary selection in speech to very young children. *Journal of Child Language* 15: 481–92.

Bernstein Ratner, N. and **Silverman, S.** 2000: Parental perceptions of children's communicative development at stuttering onset. *Journal of Speech, Language, and Hearing Research* 43: 1252–63.

Biber, D. 1988: *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biggs, A., Daniel, L., Feather, R.M., Ortleb, E., Rillero, P., Snyder, S.L. and **Zike, D.** 2003*: Glencoe Science: Science level green*. New York: Glencoe/McGraw-Hill.

Bucks, R.S., Singh, S., Cuerden, J.M. and **Wilcock, G.K.** 2000: Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance. *Aphasiology* 14: 71–91.

Carrell, P.L. and **Monroe, L.B.** 1993: Learning styles and composition. *The Modern Language Journal* 77: 148–62.

Carroll, J.B. 1964: *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Chotlos, J.W. 1944: Studies in language behavior. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs* 56: 75–111.

Colwell, K., Hiscock, C.K. and **Memon, A.** 2002: Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology* 16: 287–300.

Daller, H., Van Hout, R. and **Treffers-Daller, J.** 2003: Lexical richness in the spontaneous speech of bilinguals, *Applied Linguistics* 24: 197–222.

Dempsey, K.B., McCarthy, P.M. and **McNamara, D.S.** Using phrasal verbs as an index to distinguish text genres. In D. Wilson and G. Sutcliffe (eds.), *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference* (pp. 217–222). Menlo Park, California: The AAAT Press.

Dempsey, K.B., McCarthy, P.M. and **McNamara, D.S.** 2006: Identifying text genres using phrasal verbs. *Proceedings of the 28th annual conference of the Cognitive Science Society*, Vancouver, Canada.

Dickens, C. 1995: *A tale of two cities*. Longman: London.

Dugast, D. 1978: Sur quoi se fonde la notion d'étendue théoretique du vocabulaire? *Le Français Moderne* 46: 25–32.

—— 1979: *Vocabulaire et stylistique. I Théâtre et dialogue. Travaux de linguistique quantitative*. Geneva: Slatkine-Champion.

Durán, P., Malvern, D., Richards, B. and **Chipere, N.** 2004: Developmental trends in lexical diversity. *Applied Linguistics* 25: 220–42.

Ertmer, D.J., Strong, L.M. and **Sadagopan, N.** 2002: Beginning to communicate after cochlear implantation: Oral language development in a young child, *Journal of Speech, Language, and Hearing Research* 46: 328–40.

Grela, B.G. 2002: Lexical verb diversity in children with Down syndrome. *Clinical Linguistics & Phonetics* 16: 251–63.

Guiraud, P. 1960: *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel.

Harris Wright, H., Silverman, S.W. and **Newhoff, M.** 2003: Measures of lexical diversity in aphasia, *Aphasiology* 17: 443–52.

Heaps, H.S. 1978: *Information retrieval: Computational and theoretical aspects*. New York: Academic Press.

Herdan, G. 1960: *Quantitative linguistics*. London: Butterworth.

Hess, C.W., Sefton, K.M. and **Landry, R.G.** 1986: Sample size and type-token ratios for oral language of preschool children, *Journal of Speech and Hearing Research* 29: 129–34.

Holmes, D.I. and **Singh, S.** 1996: A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing* 11: 133–40.

Holmes, J., Vine, B. and **Johnson, G.** 1998: *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Honoré, A. 1979: Some simple measures of richness of vocabulary, *Association for Literary and Linguistic Computing Bulletin* 7: 172–77.

Hoover, D. 2003: Another perspective on vocabulary richness. *Computers and Humanities* 37: 151–78.

International Computer Archive of Modern and Medieval English 2000: *Lancaster/Oslo/Bergen Corpus of British English* (CD-ROM).

—— 2000: *The London-Lund Corpus of Spoken English* (CD-ROM).

Jarvis, S. 2002: Short texts, best-fitting curves and new measures of lexical diversity, *Language Testing* 19: 57–84.

—— 2003. Measuring lexical diversity through "exhaustive sampling". Paper presented at the Second Language Research Forum (SLRF), Tucson, AZ.

Johansson, S., Leech, G. and **Goodluck, H.** 1978: *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo.

Kucera, H. and **Francis, W.N.** 1967: *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Linnarud, M. 1986: *Lexis in composition. A performance analysis of Swedish learners' written English*. (Lund Studies in English 74). Malmo: Liber Forlag (CWK Gleerup).

Louwerse, M.M., McCarthy, P.M., McNamara, D.S. and **Graesser, A.C.** 2004: Variation in language and cohesion across written and spoken registers. In Forbus, K., Gentner, D. and Regier, T. (eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society*. Cognitive Science Society, 843–48.

Maas, H.D. 1972. Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik* 8: 73–79.

MacWhinney B. 2000: *The CHILDES project: Tools for analyzing talk*, Vol. 2: *The database*, third edition. Mahwah, NJ: Lawrence Erlbaum.

McCarthy, P.M. 2005: An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Unpublished PhD dissertation, University of Memphis.

McKee, G., Malvern, D. and **Richards, B.** 2000: Measuring vocabulary diversity using dedicated software, *Literary and Linguistic Computing* 15: 323–37.

Malvern, D.D. and **Richards, B.J.** 1997: A new measure of lexical diversity. In Ryan, A. and Wray, A., editors, *Evolving models of language*. Clevedon: Multilingual Matters, 58–71.

—— 2000: Validation of a new measure of lexical diversity. In Beers, M., Bogaerde, B. v.d., Bol, G., de Jong, J. and Rooijmans, C., editors, *From sound to sentence: Studies on first language acquisition*. Groningen: Centre for Language and Cognition, University of Groningen.

Malvern, D.D. Richards, B.J., Chipere, N. and **Durán, P.** 2004: *Lexical diversity and language development: Quantification and assessment*. Houndmills, Hampshire: Palgrave Macmillan.

Meara, P. and **Bell, H.** 2001: P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect* 16: 5–19.

Michéa, R. 1969: Répétition et variété dans l'emploi des mots. *Bulletin de la Société de Linguistique de Paris*, 1–24.

Miller, J.F. 1981: Quantifying productive language disorders. In Miller, J.F., editor, *Research on child language disorders: A decade of progress*. Austin, TX: Pro-Ed, 211–20.

Orlov, Y.K. 1983: Ein model der häufigekeitsstruktur des vokabulars. In Guiter, H. and Arapov, M.V., editors, *Studies on Zipf's Law*. Bochum: Brockmeyer, 154–233.

Owen, A.J. and **Leonard, L.B.** 2002: Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech and Hearing Research* 45: 927–37.

Ransdell, S. and **Wengelin, Å.** 2003: Socioeconomic and sociolinguistic predictors of children's L2 and L1 writing quality. Arob@se, 1–2, 22–29 http://www.arobase.to/somm.html

Ratner, N. and **Silverman, S.** 2000: Parental perceptions of children's communicative development at stuttering onset, *Journal of Speech, Language, and Hearing Research* 43: 1252–63.

Read, J. 2000: *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, B.J. and **Malvern, D.D.** 1997. *Quantifying lexical diversity in the study of language development. New Bulmershe Papers*. Reading: University of Reading.

—— 1998. A new research tool: Mathematical modeling in the measurement of vocabulary diversity (Award reference no. Rooo221995). Final Report to the Economic and Social Research Council, Swindon, UK.

Rietveld, T. and **van Hout, R.** 1993: *Statistical techniques for the study of language and language behavior*. Berlin: Mouton de Gruyter.

Shannon, C.E. 1948: A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423, 623–56.

Sichel, H.S. 1975: On a distributive law for word frequencies. *Journal of the American Statistical Association* 70: 542–47.

Silverman, S. and **Bernstein Ratner, N.** 2000: Word frequency distributions and type-token characteristics. *Mathematical Scientist* 11: 45–72.

Singh, S. 2001: A pilot study on gender differences in conversational speech on lexical richness measures. Literary and Linguistic Computing 6: 251–64.

Smith, J.A. and **Kelly, C.** 2002: Stylistic constancy and change across literary corpora: using measures of lexical richness to date works, *Computers and the Humanities* 36: 411–30.

Somers, H.H. 1966: Statistical methods in literary analysis. In Leeds, J., editor, The computer and literary style. Kent, OH: Kent State University, 128–40.

Svartvik, J. and **Quirk, R.** 1980: *A Corpus of English Conversation*. Lund: CWK Gleerup.

Swales, J. and **Malczewski, B.** 2001: *Discourse management and new-episode flags in MICASE*. In Simpson, R.C. and Swales, J.M., editors, *Corpus linguistics in North America*. Ann Arbor, MI: University of Michigan Press, 145–64.

Templin, M. 1957: *Certain language skills in children*. Minneapolis, MN: University of Minneapolis Press.

Thordardottir, E.T. and **Ellis Weismer, S.** 2001: High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment, *International Journal of Language & Communication Disorders* 36: 221–44.

Tweedie, F.J. and **Baayen, R.H.** 1998: How variable may a constant be? Measures in lexical richness in perspective, *Computers and the Humanities* 32: 323–52.

Van Genderen, J.L. and **Lock, B.F.** 1977: Testing land-use map accuracy. *Photogrammetric Engineering and Remote Sensing* 43: 1135–37.

Wright, H.H., Silverman, S.S. and **Newhoff, M.** 2003: Measures of lexical diversity in aphasia, *Aphasiology* 17: 443–52.

Wu, T. 1993. An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software* 19: 33–43.

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