Behavior Research Methods
2010, 42 (2), 381-392
doi:10.3758/BRM.42.2.381

MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment

Philip M. McCarthy
University of Memphis, Memphis, Tennessee

and

Scott Jarvis
Ohio University, Athens, Ohio

The main purpose of this study was to examine the validity of the approach to lexical diversity assessment known as the measure of textual lexical diversity (MTLD). The index for this approach is calculated as the mean length of word strings that maintain a criterion level of lexical variation. To validate the MTLD approach, we compared it against the performances of the primary competing indices in the field, which include vocd-D, TTR, Maas, Yule's K, and an HD-D index derived directly from the hypergeometric distribution function. The comparisons involved assessments of convergent validity, divergent validity, internal validity, and incremental validity. The results of our assessments of these indices across two separate corpora suggest three major findings. First, MTLD performs well with respect to all four types of validity and is, in fact, the only index not found to vary as a function of text length. Second, HD-D is a viable alternative to the vocd-D standard. And third, three of the indices—MTLD, vocd-D (or HD-D), and Maas—appear to capture unique lexical information. We conclude by advising researchers to consider using MTLD, vocd-D (or HD-D), and Maas in their studies, rather than any single index, noting that lexical diversity can be assessed in many ways and each approach may be informative as to the construct under investigation.

Lexical diversity (LD) refers to the range of different words used in a text, with a greater range indicating a higher diversity. Easily applied to almost any type of text, LD has featured in a wide range of applications, producing a rich history of textual assessment. Thus, LD has been used by researchers in fields as varied as stylistics, neuropathology, language acquisition, data mining, and forensics; and LD indices have been found to be indicative of writing quality, vocabulary knowledge, speaker competence, Alzheimer's onset, hearing variation, and even speaker socioeconomic status (see Malvern, Richards, Chipere, & Durán, 2004; McCarthy & Jarvis, 2007). Although there can be little doubt as to the application of LD, the identification of a robust index to represent it has proven to be problematic. The problem is that LD indices have tended to demonstrate sensitivity to variations in text length. As a result, researchers who have been aware of this problem (e.g., Biber, 1989) have been forced to restrict their analysis to narrow bands of text length; and researchers who appear not to have been aware of this problem (e.g., Ertmer et al., 2002; Miller, 1981) have produced findings that may be misleading. This problem of text length sensitivity is particularly the case for the best known LD index, type–token ratio (TTR: Templin, 1957), the use of which and problems with which have been extensively documented (e.g., Malvern et al., 2004).

Recent years have seen the emergence of more robust approaches to LD assessment—particularly, vocd-D and the measure of textual lexical diversity (MTLD).[1] Each approach purports to assess LD, while having little or no effect of text length. But although vocd-D has been described frequently (Malvern et al., 2004; McCarthy & Jarvis, 2007; Owen & Leonard, 2002), MTLD has been used only as one variable among many in combinatorial engineering styled analyses (e.g., Crossley & McNamara, in press; McNamara, Crossley, & McCarthy, 2010) and, consequently, has yet to have its architecture described in detail. More important than a description, however, MTLD has yet to be validated or systematically compared with leading indices of LD. Thus, the foremost purpose of the present study is to fill these gaps in the research, particularly by addressing the following research question: To what degree can validity be attributed to MTLD as a form of LD assessment?

As was noted above, vocd-D has been used in numerous studies (e.g., Harris Wright, Silverman, & Newhoff, 2003; Malvern et al., 2004; Owen & Leonard, 2002; Silverman & Bernstein Ratner, 2000). However, McCarthy

P. M. McCarthy, pmmccrth@memphis.edu

© 2010 The Psychonomic Society, Inc.

and Jarvis (2007) demonstrated that the vocd-D approach suffered from two major shortcomings. First, it significantly varied as a function of text length, which contradicts the claims of Malvern et al. Second, McCarthy and Jarvis demonstrated that vocd-D merely replicates the hypergeometric distribution function (HD-D; see Wu, 1993, for a discussion of hypergeometric distribution). As we shall see, this means that vocd-D is an approximation of a well-established probability function. Indeed, when tested on a corpus of 266 texts (see Jarvis, 2002, for details of the texts), the correlation between vocd-D and the value of HD-D was $r = .971$. These concerns over vocd-D led us to a second major research question: How does vocd-D compare with HD-D when assessed across a wide variety of registers?

CONSIDERATIONS IN THE ASSESSMENT OF LEXICAL DIVERSITY

Text Length

A major problem for indices of LD is sensitivity to text length. To better understand this problem, it is essential to understand two key terms: *tokens* and *types*. As a text becomes longer (i.e., by adding words to it), the number of overall words in the text (or tokens) increases. But although the token increase is linear (one new word = one new token), the rate of increase of the number of different words in the text (or types) steadily slows. The slowing of the type increase occurs because, with each new instance of a token, there is a corresponding decrease in the likelihood of a new type, because no text of more than a handful of words can be meaningful without some kind of repetition of tokens. Thus, we can say that as a text increases in length, there is a corresponding decrease in the value of diversity that is calculated to represent it. But, importantly, the increasing rate of lexical repetition does not entail that a reader would perceive the text as changing in diversity levels. Instead, the change in values may be merely a calculation issue that is driving a misleading quantitative representation. Not surprisingly, then, this calculation issue has resulted in many researchers' reporting results that are highly confounded with text length (e.g., Ertmer et al., 2002; Miller, 1981), and their reporting of the diversity levels associated with those texts has further led to considerable confusion with regard to the characterization of the texts and the people who produced those texts.

Although generally problematic, the text length sensitivity problem does have some positive uses. For instance, the gradual decrease in type count can be an indication of the thematic saturation of a text or corpus (Glaser & Strauss, 1967; Lincoln & Guba, 1985; Morse, 1995). That is, when a text reaches the point at which no new types are being encountered, we can say that the text is (fully) representative of the word types that are indicative of that text's theme (see also McEnery, 2003, for a discussion of "representativeness and balance," p. 449). Such an approach is useful because it allows researchers greater confidence that their corpora comprise texts of a sufficient length to represent suitably their linguistic function. Indeed, as we shall see, the calculation of the MTLD index makes use of a notion closely related to thematic saturation (see the Rationale for MTLD section). But despite such uses, in most corpus analyses, researchers are comparing LD levels across a range of texts, and the LD text length problem means that the researchers are often unaware that their reported LD differences are (highly) confounded by the length of the individual texts.

Textual Homogeneity

A second major problem for LD indices can be described as the assumption of textual homogeneity. In LD terms, we can view this homogeneity assumption as the distribution of types across a text. That is, different rhetorical purposes and strategies may necessitate that different parts of a text have different diversity levels. Thus, if we consider Lincoln's *of the people, by the people, for the people*, we must acknowledge that no one around at the time appears to have claimed (at least, successfully) that Lincoln was being insufficiently diverse—this, despite a type–token ratio of only .556. Moreover, we must also acknowledge that no one complained of an adverse effect on working memory when the nine preceding words (*have a new birth of freedom and that government*) were mentioned—this despite the fact that these words boasted a type–token ratio of 1.00.[2]

Of course, examining texts over short intervals is more useful for demonstration than for practical purposes. However, the issue of the homogeneity assumption remains. For example, McCarthy, Myers, Briner, Graesser, and McNamara (2009) demonstrated that nearly one fifth of an average text is composed of guest genres. Specifically, narrative texts tend to have a high number of history-like sentences; history texts tend to have guest narrative sentences; and the guests for science texts appear to be an equal division of both narrative and history sentences. The point is that a text includes a structure. The structure is vital if readers/listeners are to form a coherent mental representation of the text (Van Dijk & Kintsch, 1983). That is, the structure serves a rhetorical purpose, and this purpose may manifest itself across the text in a variety of rhetorical forms, none of which need necessarily reflect the totality of the text. Thus, as Jarvis (2002) argued, the assessment of LD through procedures that ignore this structure "treats texts as if they were composed of a vocabulary substance that has identifiable particles but no structure (such as a bucketful of colored corn kernels)" (p. 62). Thus, the concern here is ecological validity, which, for nonsequential textual assessment approaches, means that words are being assessed in a way that poorly relates to the context in which they were written.

Sequential and Nonsequential Analysis Processing

For computationally derived textual indices, the tokens that make up the text can be processed sequentially, nonsequentially, or both ways. Sequential processing would appear to be the preferred system, because it maintains the integrity of a text. That is, the computational processing of the text is akin to the human processing of the text. After all, a text is more than mere words and sentences

strung together. A text includes a structure, and this structure binds together the textual components so as to allow a reader or listener to form a coherent mental representation (Van Dijk & Kintsch, 1983). On the other hand, nonsequential processing has a practical application. For example, it has the advantage of avoiding local clustering of content words, which Malvern et al. (2004) argued may lead to a distorted view of the overall text. Landauer, Laham, Rehder, and Schreiner (1997) went even further, claiming that there may be little benefit to word order when it comes to deriving meaning from texts. Ultimately, it may seem that designers of assessment systems defend the approach that works best for their own tool. After all, the designers of vocd-D admitted that their system's performance is weaker when used sequentially (Malvern et al., 2004, p. 72), and systems such as latent semantic analysis (Landauer et al., 1997) struggle to include any aspect of word order because of the computational expense involved in the calculation of multiword phrases (Olney, 2007). But this having been said, nonsequential processing is still a common approach, and the results from nonsequential analyses have undoubtedly been as useful as they have been ubiquitous. Consequently, we argue that textual analyses that incorporate both sequential and nonsequential processing may well offer researchers valuable insights into the properties of a text.

Summary

In sum, all textual analyses are fraught with difficulty and disagreement, and LD is no exception. There is no agreement in the field as to the form of processing (sequential or nonsequential) or the composition of lexical terms (e.g., words, lemmas, bigrams, etc.); and even a common position with regard to the distinction between the terms *lexical diversity*, *vocabulary diversity*, and *lexical richness* remains unclear (Malvern et al., 2004). In this study, we do not attempt to remedy these issues. Instead, we argue that the field is sufficiently young to be still in need of exploring its potential to inform substantially. Thus, we include in our analyses the most sophisticated indices of LD that are currently available.

INDICES OF LEXICAL DIVERSITY

vocd-D

Because vocd-D has been extensively described elsewhere, we offer here only a brief review of the approach and direct the interested reader to Malvern et al. (2004) and McCarthy and Jarvis (2007). Furthermore, we discuss vocd-D using its system's default options—that is, those options selected by vocd-D's creators as being the best working parameters and the parameters that are typically reported in vocd-D studies.

The calculation of vocd-D is the result of a series of random text samplings. The approach begins its calculation by taking from the text 100 random samples of 35 tokens. The TTR for each of these samples is calculated, and the mean TTR is stored. The same procedure is then repeated for samples from 36 to 50 tokens. An empirical TTR curve is then created from the means of each of these samples. The $\mathcal{D}$ coefficient (see Malvern et al., 2004, p. 51) is then used as part of a formula to produce a theoretical curve that most closely fits the empirical TTR curve formed from the random samples. The value of the best-fitting $\mathcal{D}$ is referred to as D. Because D is arrived at by random sampling, the value varies each time the assessment is run. Thus, to create a higher level of consistency, the procedure above is run three times, and an average D is the final output. Final values tend to range from 10 to 100, with higher values indicating greater diversity.

HD-D

The calculation of vocd-D suggests that the $\mathcal{D}$ coefficient is an essential ingredient in determining the LD value of the text. But McCarthy and Jarvis (2007) demonstrated that the vocd-D value is actually based on probabilities of word occurrence and that $\mathcal{D}$ serves no purposes except to convert the LD value into a new scale. More specifically, McCarthy and Jarvis demonstrated that vocd-D is merely a complex way of approximating the hypergeometric distribution, and to demonstrate this, they described an index that we refer to here as *HD-D*.

The hypergeometric distribution represents the probability of drawing (without replacement) a certain number of tokens of a particular type from a sample of a particular size. The way we have used this distribution for our own HD-D index is to calculate, for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text.[3] The probabilities for all lexical types in the text are then added together, and the sum is used as an index of the text's LD.

As was previously mentioned, HD-D is what vocd-D approximates. Although vocd-D relies on random sampling, instead of direct calculations of lexical probabilities, vocd-D's output is nevertheless determined by those probabilities. That is, when vocd-D calculates TTR for multiple random samples of 35–50 words drawn from a text, the results approximate what one would find if one calculated TTR for all possible combinations of 35–50 words drawn from the text. Of course, the number of possible combinations is so large that this would not be feasible to do, but it is possible to use the hypergeometric distribution to calculate these values directly without any sampling. This calculation is conducted by taking into consideration the mean contribution that each type makes to the TTR of all possible combinations of a sample of a certain size. If the sample size is 42, the mean contribution of any given type is 1/42 multiplied by the percentage of combinations in which the type would be found. This percentage is exactly the same as the probability of encountering the type in a sample of 42 words drawn from the text, and this can be determined directly from the hypergeometric distribution.

For this reason, it is not surprising that McCarthy and Jarvis (2007) found correlations of $r = .971$ between vocd-D and HD-D (i.e., sums of probabilities) for sample sizes from 35 to 50 (i.e., the sizes of samples that vocd-D uses in its random-sampling procedures). The correlation would have been perfect had it not been for the slight imprecision in vocd-D's output brought about by its reliance

on both random (nonexhaustive) sampling and curve fitting. It is worth pointing out that vocd-D and HD-D output are also on a different scale. The HD-D output is literally sums of probabilities, whereas vocd-D output is essentially sums of probabilities converted to type–token ratios and, then again, from type–token ratios to a *D* value.

Other LD Indices Used in This Study

Log correction. Because the text length problem of LD is related to frequency, log values have long been used as an LD corrective factor (e.g., Herdan, 1964). Over the years, a steady trickle of variations of log approaches were introduced, such as Somers (1966), Maas (1972), Dugast (1978), and Tuldava (1993). In turn, these approaches became the subject of a number of studies (e.g., Hess, Sefton, & Landry, 1986; Tweedie & Baayen, 1998), none of which reported positive evidence for the corrective approaches. However, such studies were quite limited in terms of the range of registers analyzed, and when McCarthy and Jarvis (2007) assessed log-corrected indices against 16 written registers, they found the effect of text length to be just 1.5%. Furthermore, McCarthy and Jarvis reported that Maas[4] showed no effect of text length if analyses were limited to the following ranges of text length: 100–154, 154–333, 200–666, and 250–2,000. As such, we include Maas in the present analysis as the most representative LD index from the log correction approach.

Frequency correction. A second approach to correcting for the text length effect is the frequency distribution of types. That is, Text A and Text B may contain exactly the same overall number of types and tokens; however, the number of tokens for each type may differ. For example, consider the sentence *The friendly man liked both the big dog and the little dog*, which contains nine types and 12 tokens, and then consider the sentence *The friendly man, whom the big dog liked, liked a little dog*, which also contains nine types and 12 tokens. Note that the first sentence contains 3 tokens of the type *the*, whereas the second sentence contains only 2 tokens of the type *the*; however, for the second sentence, the word *liked* has a frequency of 2, whereas it is just 1 in the first sentence. The concern for frequency variation resulted in the introduction of many new LD indices (e.g., Honore, 1979; Orlov, 1983; Yule, 1944), and, like the log corrections, each of these indices was extensively tested (e.g., Tweedie & Baayen, 1998) and was found to be confounded by text length. Yet, as with the Maas index, McCarthy and Jarvis (2007) showed that one frequency index, K, was reasonably effective. Indeed, K showed no effect of text length if analyses were limited to the following ranges: 100–500, 154–666, 250–1,000, and 400–2,000. Given these results, our study includes K as the best representative LD index of the frequency-correction type.

Regarding K, one further point of interest is highly relevant to the vocd-D calculation. Whereas vocd-D is determined by the sums of probabilities of encountering each type in the text in sample sizes from 35 to 50 tokens, K is determined by the sums of probabilities of encountering each type in the text when the sample size is set to just 2 words. Both indices also involve a scale conversion, but the important point is that they form a continuum, both relying on the hypergeometric distribution, where K's sample size is set to 2 words and vocd-D's sample size is set to 35–50 words.

MTLD

Processing MTLD

MTLD is an index of a text's LD, evaluated sequentially. It is calculated as the mean length of sequential word strings in a text that maintain a given TTR value (here, .720). During the calculation process, each word of the text is evaluated sequentially for its TTR. For example, . . . *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (.625) *people* (.556) . . . and so forth. However, when the default TTR factor size value (here, .720) is reached, the factor count increases by a value of 1, and the TTR evaluations are reset. Thus, given the previous example, MTLD would execute . . . *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) |||FACTORS = FACTORS + 1||| *for* (1.00) *the* (1.00) *people* (1.00) . . . and so forth.

Partial Factors

A partial factor value is calculated for the lexical remainders of a text (i.e., the final words that do not form a full factor). For example, a TTR of .887 forms 40.4% of the range between 1.00 and the full factor of .720. If a text contains 4 full factors and a remainder that has a TTR of .887, then the final factor count is 4.00 + 0.404 = 4.404.

Because a text rarely ends at exactly a completed factor, the question becomes how to evaluate an incomplete factor. Approaches such as *sequentially applied vocd-D* simply discard unused tokens, an action that reduces the integrity of the index. Apart from simplicity, the major reason for ignoring remaining data is that it is difficult to incorporate it into the index: Specifically, a smaller section of text (in terms of tokens) will always have a higher TTR. Thus, adding (or fitting) this higher TTR to the average TTR will misleadingly increase the apparent diversity of the text.

As was previously discussed, MTLD does not discard any remaining data. Instead, the factor size attributed to the incomplete factor is calculated on the basis of how far the TTR value has progressed toward the designated default factor size of .720. Although including a value for textual remainders increases the integrity of the index, remainder values have a number of problems. First, remainder values are only approximations, and therefore, their values are more open to question and to error. Second, the shorter the text, the greater will be the part of its value that is composed of a remainder. That is, the 0.25 of 1.25 is greater than the 0.25 of 2.25. Therefore, the shorter the text, the more likely it is to be composed of an approximation of the average number of words needed to form a factor. Consequently, shorter texts will be more difficult to evaluate with confidence. In testing the tool during the development process, we found that texts as short as 100 tokens can be used. Texts shorter than this

are often made up of only a partial factor (i.e., no full factors, but only a single remainder), and their accuracy is questionable.

Forward and Reverse Processing

The primary purpose of the reverse-processing phase is to provide a second MTLD value for the assessment. During development testing, we found that two processing runs, one forward and one reverse, were sufficient for producing MTLD values of the desired consistency and accuracy. That is, we found that a single processing run (e.g., forward) sometimes generated results with large variations, relative to the segmentation sizes (see the internal validation analysis), presumably resulting from the partial factors issues described above. Such variations had the potential to cause interpretational difficulties, and therefore, two processing runs were instantiated. Testing demonstrated that a dual-processing sequence was sufficient to counteract this occasional variational problem, because the remainder factor calculated in reverse sequencing was not the same as the remainder of the initial run and the average of the two processing runs was sufficient to smooth the result.

The choice of running the second sequencing as a reverse sequence is not without its problems, even though the action is computationally parsimonious and the result is favorable. To better understand this problem, recall that a factor is a sequential section of the text that has reached a TTR value of .720. However, which area of text actually forms a factor primarily depends on where in the text the sequencing begins, meaning that the MTLD approach could process the text repeatedly, starting its sequencing at each new token in the text, and subsequently produce as many MTLD evaluations as there are words in the text. In practice, conducting such a high number of evaluations is computationally expensive. But, more important, developmental testing suggested that it was not necessary, because the MTLD approach required only two processing runs (one forward and one reverse) to become what Malvern et al. (2004) termed "sufficiently consistent" (Malvern et al., 2004, p. 51).

The decision to make the second evaluation a reverse sequence might appear to reduce the integrity of the index. After all, people do not read a text backward, so computationally processing it backward seems problematic. However, whether the final *x* number of words of a text are computationally processed left to right or right to left does not affect the TTR value of that factor, that is, a set sequence of words will reach the same TTR value (and therefore, the same factor) whichever direction is taken, meaning that the computational parsimony of reverse processing does not affect the psychological consistency of word order. For example, the final five words of the first sentence of this paragraph are *the integrity of the index*, producing a TTR of .800. Reversing those words gives us *index the of integrity the*, still producing a TTR of .800. Thus, what is important is the identification of a sequential *within-the-text* segment of tokens of a target TTR factor size (i.e., .720), and not the right–left or left–right order with which those tokens are computationally processed. Note that a randomized version of the segment (e.g., *the the of index integrity*) also forms a final TTR of .800, seemingly calling into question the sequential nature of the MTLD approach. However, the randomized version is quite different from the left–right and right–left versions, because words of the same type are much more likely to be clumped together, allowing for quite different quantities of factors to be formed. As such, any randomization of the text has the potential to cause significant changes in the MTLD values.

Calculation of MTLD Value

The total number of words in the text is divided by the total factor count. For example, if the text = 340 words and the factor count = 4.404, then the MTLD value is 77.203. Two such MTLD values are calculated, one for forward processing and one for reverse processing. The mean of the two values is the final MTLD value.

Factor Size

MTLD uses a default factor size of .720. This default factor size is used in the MTLD prototype that features in the Coh-Metrix suite of textual analysis indices and is the factor size used in the engineering studies that have incorporated MTLD (e.g., Crossley & McNamara, in press; McNamara et al., 2010).

The value of .720 was reached following a series of tests, using narratives and expository texts from sources such as the Project Gutenburg Text Archives (www.archive.org/details/gutenberg). Evidence from this testing suggested that TTR trajectories tended to reach a point of stabilization at around .720 (±.03). At higher TTR values (e.g., >.760), sudden fluctuations in TTR values could still occur,[5] caused by what Malvern et al. (2004) described as "lexical clusters." At lower values (e.g., <.650), stabilization is strongly established in the TTR trajectory. Thus, potential index sensitivity risks being lost through the unnecessary use of token processing; instead, these tokens could be used to form a new factor, increasing the numbers of factors for the text and forming a potentially more accurate LD value.

Our initial testing suggested that there were no significant differences between TTRs within the range from .660 to .750. The value of .720 was selected because it fell on the higher side of the middle of this range. That is, the middle point in a safe range suggests the most conservative selection. However, the fewer the number of tokens needed, the greater the number of factors that can be generated, allowing for greater accuracy in the final index. Hence, .720 was selected.

The Rationale for MTLD

To better understand the rationale of the factors that form the MTLD value, we must first consider the LD index known as mean segmental TTR (MSTTR; Johnson, 1944). The MSTTR approach divides a text into segments of a set length (typically, 100 words). The remaining words are discarded. The TTR for each full segment is calculated, and the final TTR value is the mean of all full segments. Such an approach works well enough for

texts that are very long (e.g., over 1,000 words). However, MSTTR has several problems for shorter texts. First, discarding any amount of text reduces the text's integrity and, therefore, reduces the validity of the evaluation. Second, the choice of segment size is problematic: Using smaller segments (e.g., 10 words) results in a smaller size of discarded text. However, the smaller the segment, the lower the sensitivity of the index. That is, having relatively few words in a segment allows for relatively few word types, meaning that the TTR values are very high, regardless of the text analyzed. On the other hand, larger segments make for a more sensitive value, but including larger segments requires larger texts, which may not be available. Furthermore, the larger the segment, the larger the size of the discarded text is likely to be.

The MTLD approach replaces word segments with TTR factors. At first blush, such an approach would seem to be simply swapping out one version of a problem with another. However, the MTLD approach is different because it makes use of a notion closely related to thematic saturation, an aspect of text that we refer to as the *point of stabilization* (see Figure 1).

To understand the importance of the point of stabilization, recall that a sequentially assessed text first encounters almost nothing but new types. In Figure 1, this area can be seen as a continuous horizontal trajectory (marked as a). Upon encountering the first repeated type, the TTR value drops dramatically (b). Next, the sequence returns to entirely new types, resulting in a rise in the TTR trajectory, before a second repeated type causes another sharp drop (c). This area of sharp rises and falls continues until one reaches what we refer to as the *point of stabilization*. It is at this point that neither the introduction of repeated types nor even a considerable string of new types can markedly affect the TTR trajectory. Thus, we can consider the area of the sequence following the point of stabilization as type saturated, and we can view this manifestation in Figure 1 as a gradual and relatively smooth descent that the trajectory ultimately takes.

For MTLD, the key element is how many words it takes to reach the area prior to the point of stabilization. Clearly, the fewer words it takes, the less diverse is the text (in terms of sequential lexical deployment). Thus, the MTLD value is simply the average number of words required for the text to reach a point of stabilization.

The identification of the point of stabilization distinguishes the problem of MSTTR from the solution of MTLD. By dividing the text into factors (the number of words needed to reach a point of saturation), we are replacing an arbitrary segment size (typically, 100) with an empirically driven textual factor size. This replacement is useful because the MTLD factor sizes are generally much smaller than 100 words (as in our example in Figure 1), meaning that many more factors can be generated, and thus the derived MTLD value is more likely to be a sensitive evaluation of the diversity of the text. More specifically, the area of the text after the point of stabilization does not inform us any further as to the diversity of the text. In essence, whereof the TTR trajectory is smooth, thereof the tokens are being wasted. Thus, the argument is that a preplanned segment of 100 words might (and often does) form a TTR value that could be calculated with far fewer tokens. And similarly, a segment of 100 words might be too short, meaning that it has not yet reached a point where the diversity has worked itself out. The goal, then, is to make the segments optimally sized, and the approach taken by MTLD allows for just that. Of course, although the factor sizes are typically small with MTLD (i.e., fewer than 100 words), there is not the concern of lack of sensitivity associated with these small segment sizes (such as we have with MSTTR). The sensitivity problem for MSTTR occurs because small predetermined segments remain high in TTR values, making it hard to distinguish LD values across a range of texts. However, with MTLD, all factors, regardless of number of words,

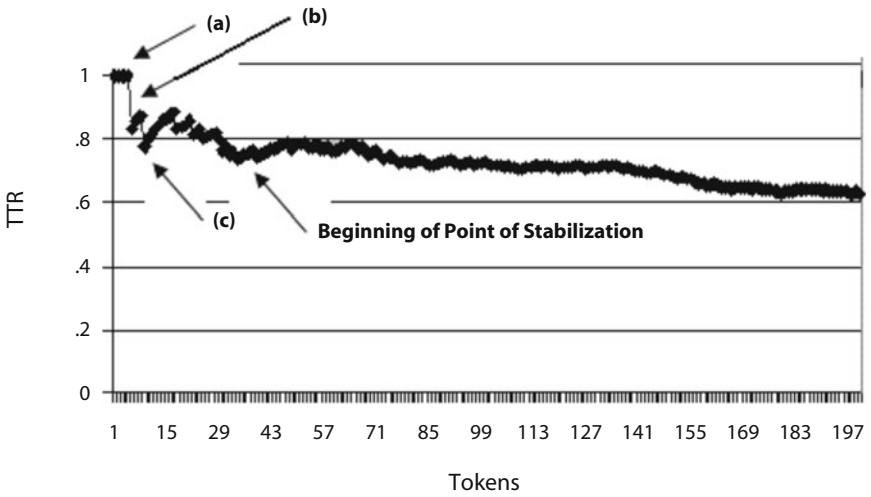

Figure 1. The type–token ratio (TTR) and point of stabilization as formed from the first 200 words of the first chapter of *The Red Badge of Courage.*

have reached a TTR of .720; therefore, we know that the values are sufficiently sensitive.

Various studies have shown MTLD to be at least as effective as the industry standard vocd-D index, and even one of the most informative and distinguishing variables in the entire arsenal of several hundred Coh-Metrix indices (see Crossley & McNamara, 2009; Crossley, Salsbury, & McNamara, 2009; McNamara et al., 2010). Such successes do not mean that all the settings of the MTLD approach are optimal. However, such results in conjunction with the present validation study do offer confidence that MTLD is a valuable resource for researchers using LD.

CONSTRUCT VALIDITY

Construct validation requires a well-developed theoretical framework that explains the nature of the target construct, its observable properties, and its interrelationship with other constructs and concepts. Although construct validation necessarily involves the presentation of logical arguments and even qualitative evidence linking the theoretical construct to the measure, it also requires comparing and contrasting the approach in question with approaches of both similar and dissimilar constructs in order to demonstrate that it assesses that which it is designed to assess (Ong & van Dulmen, 2006, p. 66). The types of comparison and contrast that are used can be described as ways of establishing various *types of validity* (see American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) that contribute to the overall construct validity of the measure. In this study, we assessed MTLD (along with the more established LD indices: vocd-D, HD-D, Maas, K, and TTR) in relation to the following four types of validity: convergent, divergent, internal, and incremental.

METHOD

Corpora

We used two corpora in this study. For the first three assessments, we used the same texts as McCarthy and Jarvis (2007) and will refer to this collection as the *MJ corpus*. The texts of the MJ corpus comprise 16 registers: 15 taken from the Lancaster–Oslo–Bergen corpus (LOB, Johansson, Leech, & Goodluck, 1978), and the remaining 1 extracted from Glencoe Science (Biggs et al., 2003). The registers are press reportage, editorials, press reviews, religion, skills and hobbies, popular lore, biographies, official documents, academic prose, general fiction, mystery fiction, science fiction, adventure fiction, romantic fiction, humor, and textbooks. Each register contains nine individual texts,[6] with each text originally composed of approximately 2,000 words to the nearest end of sentence. All texts over 2,000 words are trimmed to a total of 2,000 words. This collection of texts is well established and has featured in the studies of Biber (1988), Louwerse, McCarthy, McNamara, and Graesser (2004), Dempsey, McCarthy, and McNamara (2007), and McCarthy and Jarvis (2007).

The fourth assessment in this study was based on texts manipulated for cohesion The majority of the texts (19) are described in McNamara, Louwerse, McCarthy, and Graesser (in press). The remaining 4 texts are described in Best, Ozuru, Floyd, and McNamara (2006). We will refer to the entire 23-text collection from McNamara and colleagues as the M&C corpus. This additional corpus was needed to provide evidence of incremental validity (see the Results and Discussion section). Each of the 23 original texts was manipulated once; thus, we have two versions of 23 texts. As such, one version of the text is described as high in cohesion, and one version is described as low in cohesion. All the texts featured in independent experiments show learning gains from the higher cohesion version.

Whereas we argue that the MJ corpus offers the size and range of texts necessary to assess convergent, divergent, and internal validity type tests, the M&C corpus offers a task assessment not previously conducted in LD research: distinguishing low- and high-cohesion texts. The higher cohesion texts are likely to feature greater lexical overlap and, therefore, lower levels of LD. The point of importance here is that LD approaches that can distinguish the text types offer an incremental advantage over those that do not. Thus, the M&C corpus forms part of the incremental validity assessment.

Procedure

Following common practice in LD assessment (e.g., Hess et al., 1986; McCarthy & Jarvis, 2007; McKee, Malvern, & Richards, 2000; Tweedie & Baayen, 1998), each text of each register was divided into ever smaller sections (i.e., one section of 2,000, two sections of 1,000, etc.). This sectioning allowed each text to be represented in 11 different size forms for 99 data points per register; more specifically, the sectioning was necessary to establish the sensitivity of the LD indices when texts of varying lengths were assessed (see the discussion of internal validity, below). Thus, a total of 1,584 textual units were included in the study (see Table 1).

The final value for each of the section sizes was calculated by the mean value across each processing. For example, Text T of Register R, processed for 2,000 words, as evaluated by LD index LD-I, might receive a value of 78.88. However, the same text, evaluated by the same index, for 1,000 words would have two results, because there were two sections (i.e., the first 1,000 words and the second 1,000 words). As such, those two values (e.g., 79.43 and 73.03) were averaged; so, in this hypothetical example, the final value would be 76.23. In theory, the ideal index would show the same value for each division of the text; however, perfection was not the goal here. Instead, we sought to assess the degree to which there would be a significant change in the value of the index, because such a change would suggest a confound caused by the length of the text.

RESULTS AND DISCUSSION

Preliminary Analyses

The primary goals of this study were to assess the validity of sophisticated LD indices, particularly the MTLD and HD-D indices. However, as has been discussed, these indices have various possible settings. For example, MTLD can set its TTR factor value anywhere between .001 and

Table 1
Text Lengths (in Words) Used in the Analyses and Number of Sections for Each Length

Text Lengths	Number of Sections
2,000	1
1,000	2
666	3 (two texts of 667 tokens)
500	4
400	5
333	6 (two texts of 334 tokens)
286	7 (two texts of 285 tokens)
250	8
200	10
154	13 (two texts of 153 tokens)
100	20

.999, although the prototype tool is set at .720. Similarly, McCarthy and Jarvis (2007) set HD-D at 42, arguing that any level between 35 and 50 would be just as suitable. To better establish confidence in these prototype settings, two preliminary assessments were conducted.

In the first preliminary analysis, a range of potential MTLD factor default values were compared across the 16 registers of the MJ corpus. In total, 10 versions of MTLD were used, ranging from a factor size of .660 to .750. Interindex correlation averages for the range were very high (average, $r = .970$; minimum, factor size of .66 at $r = .953$; maximum, factor size of .70 at $r = .979$). On the basis of these results, we could presume that any factor size within the hypothesized range would be suitable and that averaging results across this range (or parts of this range) would not add to the usefulness of the index. We saw no benefit at that time for changing the MTLD value from that used in the previously published engineering studies (e.g., Crossley & McNamara, 2009) and opted to set our default value for MTLD at the factor size of .720 (average correlation, $r = .976$)

In the second preliminary analysis, HD-D values were compared with vocd-D values across the same 16 registers of the MJ corpus. Four versions of HD-D were used, with each index different only in the start and end sample sizes used in the calculation of the sum of probabilities that formed the index. Thus, HD-D (35–35) used 35 as the sample size, HD-D (42–42) used 42, HD-D (50–50), used 50, and HD-D (35–50) used sample sizes of 35, 36, 37, . . . up to 50 and then averaged the results. The results for each of the four HD-D indices correlated highly with vocd-D: HD-D (35–35), $r = .912$; HD-D (42–42), $r = .913$; HD-D (50–50), $r = .911$; and HD-D (35–50), $r = .913$. The high consistency supported the findings of McCarthy and Jarvis (2007), and thus, we followed McCarthy and Jarvis by using the median sample size value of HD-D (42–42) as the standard index for the remainder of the study.

Although the correlations between HD-D and vocd-D were high ($r > .910$), they were not as high as the correlations reported in McCarthy and Jarvis (2007) for 266 narrative texts ($r = .971$). Closer examination revealed that 10 of the 16 registers correlated at $r > .950$, 4 of the 16 registers correlated at $r > .910$, but 2 of the 16 registers correlated at $r < .600$. Examining the individual texts, we found that some low differences might be the result of a narrow band of extremely high LD. Using one such very high diversity text, we created 30 values for vocd-D ($M = 215.210$, $SD = 2.369$, minimum = 211.41, maximum = 219.5, range = 8.09). The result suggests a problem for vocd-D. Specifically, if a text is sufficiently long and has a much higher diversity level than is typical, randomly sampling as few as 35 tokens can lead to misleadingly high vocd-D values. It appears that even the multiple iterations of the vocd-D process does not always account for this issue. In contrast, the result for HD-D is the same no matter how many times the process is calculated, because, unlike vocd-D, HD-D is not stochastic. Thus, the results of this initial analysis suggest that using HD-D, instead of vocd-D, is a viable option for researchers, especially if the texts under examination are likely to be lengthy and/or high in diversity levels.

Validation Results

Convergent validity is the evaluation of how well an index agrees with other indices that are widely accepted as a standard against which to measure a given construct (here, LD); thus, our primary interest is the degree to which MTLD approximates vocd-D, HD-D, K, and Maas. Interindex correlations (see Table 2) demonstrate that MTLD correlates highly with all the established LD indices (minimum, K, $r = .694$; maximum, vocd-D, $r = .848$, effect size [r^2] = .719). The average MTLD correlation against a combination of vocd-D, K, and Maas is $r = .795$. And, if we replace vocd-D with HD-D, the average MTLD correlation against a combination of HD-D, K, and Maas is similar ($r = .768$). The result suggests that MTLD satisfies convergent validity to at least the same degree as other sophisticated and established LD indices.

Divergent validity is the evaluation of how well an index does not agree with indices that are considered to be flawed or misleading. In this case, we consider the flawed index to be TTR. The results (see Table 1) suggest that MTLD (like Maas, K, vocd-D, and HD-D) demonstrates greater differences from TTR than similarities (i.e., if $r < .710$, then the effect size [r^2] $< .500$ and, therefore, demonstrates greater difference than similarity). The MTLD and TTR correlation of $r = .322$ ($r^2 = .104$) is high, as compared with vocd-D, HD-D, and K, although Maas is higher ($r = .501$). One view of this result is to recall that although TTR is clearly a flawed index of LD, it is nonetheless still an index of LD. As such, divergent validity may best be achieved in this instance by avoiding a high correlation. To this end, Cohen (1988) suggested that $r = .10$ to $r = .29$ is a low correlation and $r = .30$ to $r = .49$ is a medium correlation. In these terms, none of the LD indices can be described as highly correlating with TTR. Taken as a whole, the results suggest that each of the LD indices satisfies divergent validity.

Internal validity is used here to evaluate the sensitivity of the LD indices to variations in text length (as evaluated by correlation analyses). In a textual analysis study, we can establish the degree of internal validity by manipulating one variable (i.e., text length) and assessing it against the measuring variable (i.e., each LD index). In this way, we can make a causal inference as to whether different lengths of text produce different outcomes. According to

Table 2
Correlations Between Lexical Diversity Indices, Including MTLD, vocd-D, HD-D, K, Maas, and TTR ($N = 1,584$)

	vocd-D	HD-D	K	Maas	TTR
MTLD	.848	.800	.694	.843	.322
vocd-D		.913	.833	.669	.088
HD-D			.825	.642	−.051*
K				.601	.086**
Maas					.501

Note—All correlations significant at $p < .001$, except $^*p = .043$ and $^{**}p = .001$.

Table 3
ANOVA Results for the Six Lexical Diversity Indices

	F	η_p^2
MTLD	124.804	.544
Maas	93.410	.472
vocd-D	91.418	.467
HD-D	80.360	.435
K	56.346	.350
TTR	9.046	.080

Note—All significant at $p < .001$.

this approach, the mean of the LD evaluations for the 11 size forms will approximate the LD of the original 2,000-word text if the index is not a function of text length. The more the evaluations of the size forms trend (upward or downward) with text length, the more we can say that the LD index varies as a function of text length and, consequently, the lower is its internal validity. The results suggest that MTLD has no correlation with text length; all the other indices do correlate with text length: TTR ($r = .811$, $p < .001$); HD-D ($r = .282$, $p < .001$); vocd-D ($r = .190$, $p < .001$); Maas ($r = .125$, $p < .001$); K ($r = .112$, $p < .001$); MTLD ($r = -.016$, $p = .530$). The results suggest that MTLD satisfies internal validity.

Incremental validity is used in this study to assess the degree to which a given index is informative above and beyond another (presumably similar) index. To assess incremental validity, we assessed the LD indices using the MJ corpus (a between-texts analysis) and also using the M&C corpus (a within-text analysis).

As has been discussed, the MJ corpus is composed of 16 individual registers. We used these registers as a grouping variable to assess which LD indices best distinguish the groups. More important for an assessment of incremental validity, we combined the variables into a model to ascertain which variables contributed most to the categorization. To begin, we conducted an ANOVA. Register was used as the grouping variable, and the individual LD indices were the dependent variables. The results of the ANOVA (see Table 3) suggest that all the variables distinguish the registers.

The ANOVA results are informative as to the potential power of the LD indices to discriminate registers. However, to better assess which variables contribute above and beyond other variables (i.e., incremental validity), we conducted a discriminant analysis. Because each index was significant in the ANOVA, we entered all the variables into the analysis; and because we did not want to bias the results, we used the stepwise method. The results of the discriminant analysis (see Table 4) suggest that only two LD indices significantly contribute to the prediction model (MTLD, followed by vocd-D; total accuracy = 36.9%; cross-validated accuracy = 36.1%; random chance = 6.25%; see Analysis 1 in Table 4). Although Maas produced the second highest effect size in the ANOVA, the second position in the model was taken by vocd-D. The result suggests that vocd-D contributes uniquely to the prediction model, whereas the contribution of Maas is subsumed by either MTLD or a combination of MTLD and vocd-D.

To better understand this issue, we removed vocd-D (although we retained HD-D) and reran the discriminant analysis (see Analysis 2 in Table 4). The result yielded a highly similar accuracy (total accuracy = 36.9%; cross-validated accuracy = 36.1%; random chance = 6.25%) and demonstrated that Maas becomes a significant contributor. The result suggests that MTLD does not completely subsume Maas. Note that by replacing vocd-D with HD-D, the total reported accuracy value was lower at 34.2% (cross-validated accuracy: 33.0%).

We then reintroduced vocd-D and removed MTLD (see Analysis 3 in Table 4). The result was weaker in terms of reported accuracy (total accuracy = 32.7%; cross-validated accuracy = 31.7%; random chance = 6.25%); however, it showed that Maas and vocd-D become significant contributors if MTLD is not present. The result suggests that Maas and vocd-D identify unique LD information as it applies to a diverse set of registers such as the MJ corpus.

Taken as a whole, our first incremental validity analysis suggests that at least three LD variables (MTLD, vocd-D, and Maas) contain valuable (i.e., unique) information. However, the results should not be interpreted to mean that the remaining LD indices do not contain unique information. Over different corpora, it is possible that the other variables would be significant contributors. As such, the main conclusion for this analysis is that LD indices cannot be assumed to be assessing the same latent trait, and each

Table 4
Results of Three Discriminant Analyses Featuring the Lexical Diversity Indices MTLD, vocd-D, and Maas

			Wilks's Lambda				
					Exact F		
Analysis	Step	Entered	Statistic	$df(1)$	Statistic	$df(1)$	$df(2)$
1	1	MTLD	0.456	1	124.804	15	1,568
	2	vocd-D	0.347	2	72.973	30	3,134
2	1	MTLD	0.456	1	124.804	15	1,568
	2	Maas	0.347	2	72.840	30	3,134
3	1	Maas	0.528	1	93.410	15	1,568
	2	vocd-D	0.357	2	70.282	30	3,134

Note—All significant at $p < .001$; $df(2) = 15$, $df(3) = 1{,}568$; lower Wilks's Lambda indicates a stronger model.

index might contribute to a better understanding of the characteristics of a text.

Our second assessment for incremental validity used the M&C corpus. Recall that the M&C corpus comprises two versions of 23 texts collected from 13 independent studies. All the texts featured in those experiments showed learning gains from the higher cohesion versions. In those experiments, cohesion referred to explicit lexical cues that linked one element of the text with another. The issue here is that higher cohesion texts will feature greater overlap than will their lower cohesion counterparts. Of course, some of that overlap will be in the form of semantic relatedness (e.g., *chairs* and *tables* approximate to *furniture*), and other examples will demonstrate grammatical differences (e.g., *help*, *helps*, *helped*, *helping*). Nevertheless, we still predict that a high-cohesion text will feature greater levels of simple word repetition than will a low-cohesion version, and it is this element that we assess here. To date, a variety a computational textual indices have assessed these texts, including string overlap, latent semantic analysis, and given/new span (see McCarthy et al., in press; McNamara et al., in press). The corpus is of interest to LD assessment because overlap can be regarded as the opposite of diversity. That is, because LD indices assess rates of diversity, rather than rates of overlap, we can predict that those LD indices that are able to assess cohesion differences will show significant decreases in their values. The contribution of LD assessment to cohesion evaluation is useful because validated cohesion indices (e.g., string overlap and LSA) typically assess cohesion on a sentence-to-sentence basis (see McCarthy et al., in press; McNamara et al., in press), and sentence-to-sentence assessment approaches have to consider variations in text length. Thus, a validated LD assessment of the M&C corpus offers an incremental advancement and a useful perspective for discourse psychologists conducting research in cohesion.

To assess the M&C corpus, we conducted a repeated measures ANOVA. The results (see Table 5) suggest that TTR, Maas, and MTLD significantly distinguish the high- and low-cohesion versions of the texts. TTR produced the highest effect size ($\eta_p^2 = .603$); however, the TTR result can be explained by differences in text length between the low- and high-cohesion versions of the texts [low, $M = 498.522$ words per text, $SD = 299.188$; high, $M = 659.783$ words per text, $SD = 389.903$; $F(1,22) = 23.122$, $p < .001$, $\eta_p^2 = .512$]. Given our previous internal validity results concerning TTR, we cannot argue that the index is a useful assessment of cohesion distinction.

Table 5
ANOVA Results for Lexical Diversity Indices Assessing Low- and High-Cohesion Texts

LD Index	*F*	*p*	η_p^2
TTR	33.432	<.001	.603
Maas	11.517	.003	.344
MTLD	5.775	.025	.208
K	0.744	.398	.033
HD-D	0.668	.423	.029
vocd-D	0.474	.498	.021

Maas produced the second highest effect size ($\eta_p^2 = .344$). Our divergent validity results suggest that Maas and TTR correlated relatively highly ($r = .501$), which is a concern for forming confidence in the Maas index. However, the lengths of the text in this particular analysis fall within McCarthy and Jarvis's (2007) guidelines for safe use of Maas. As such, the results suggest that Maas is a useful assessment of cohesion.

MTLD produced the third and final significant distinction ($\eta_p^2 = .208$); neither vocd-D nor HD-D was significant. The result suggests that Maas and MTLD satisfy incremental validity, because they are able to perform (at least one kind of) an assessment task that cannot be achieved by the industry standard index of vocd-D.

Finally, although we used the M&C corpus to provide evidence of incremental validity, the corpus can also provide further evidence of divergent validity. Recall that we argued that LD (differences in the composition of a text) is the opposite evaluation of cohesion (overlap in a text). Indeed, this was the basis for choosing the M&C corpus for validation material. To substantiate this claim, we conducted a correlation analysis of MTLD with the most powerful cohesion index used in McNamara and colleagues' studies (noun overlap; Best et al., 2006; McNamara et al., in press). The result ($r = -.366$, $p = .012$) suggests that there is an inverse relationship between cohesion and lexical diversity, and as a consequence, this analysis further supports MTLD in terms of divergent validity.

GENERAL DISCUSSION

The primary purpose of this study was to assess the validity of MTLD. Our results provide compelling evidence in its favor. In terms of convergent validity, MTLD values correlate highly with well-established sophisticated approaches. In terms of divergent validity, MTLD does not correlate highly with flawed LD approaches (e.g., TTR) and correlates negatively with approaches that assess the opposite of LD, such as cohesion. In terms of internal validity, MTLD results are consistent, regardless of the text length under analysis. And in terms of incremental validity, MTLD explains textual information that similar approaches do not account for. This study demonstrates that MTLD is a powerful index; however, further research is needed to fully explore the many possible variations in its settings. For example, variations in the default .720 may account for LD differences better in one kind of genre than in another. In addition, settings are needed for such features as LD for content words only and LD using lemmatized tokens.

Our study was also concerned with the LD index HD-D. Specifically, we sought to compare vocd-D and HD-D across a wide variety of registers. Our results suggest that the two indices measure the same latent trait. The major difference between the variables in practice may be only that the random sampling of the vocd-D approach leads to greater fluctuations when texts of very high diversity are evaluated.

A third finding, and arguably the most important, is that at least three of the sophisticated LD indices used in this

study do not appear to assess exactly the same latent trait. That is, MTLD, vocd-D (or HD-D), and Maas all appear to be able to capture unique LD information. Future work is needed to better determine the degree to which that information is captured and whether that information varies as a result of register.

The overarching construct of LD requires theoretical and empirical evidence (including qualitative) that goes beyond the kind of word range assessment provided here. As such, researchers need to be well informed as to LD indices, particularly their limitations, and must be responsible for the decisions they take in selecting them. With this caveat in mind, we advise those interested in LD to consider using MTLD, vocd-D (or HD-D), and Maas in their studies, rather than any single index, reminding researchers that LD can be assessed in many ways and that each approach may be informative as to the construct under investigation.

AUTHOR NOTE

This research was supported in part by the Institute for Education Sciences (IES; Grants R305GA080589, R305G020018-02, and R305G040046) and in part by the National Science Foundation (NSF; Grant IIS-0735682). The views expressed in this article do not necessarily reflect the views of the IES or the NSF. The authors acknowledge the contributions made to this project by Scott Crossley, Danielle McNamara, Max Louwerse, Zhiqiang Cai, Arthur Graesser, Diana Lam, Lisa Mintz, Emily Thrush, Teresa Dalle, and Charles Hall. Correspondence concerning this article should be addressed to P. M. McCarthy, Department of English, University of Memphis, 467 Patterson Hall, Memphis, TN 38152-3530 (e-mail: pmmccrth@memphis.edu).

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Best, R., Ozuru, Y., Floyd, R., & McNamara, D. S. (2006). Children's text comprehension: Effects of genre, knowledge, and text cohesion. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *Proceedings of the Seventh International Conference of the Learning Sciences* (pp. 37-42). Mahwah, NJ: Erlbaum.

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

Biber, D. (1989). A typology of English texts. *Linguistics*, **27**, 3-43.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, **18**, 119-135.

Crossley, S. A., & McNamara, D. S. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning*, **59**, 307-334.

Dempsey, K. B., McCarthy, P. M., & McNamara, D. S. (2007). Using phrasal verbs as an index to distinguish text genres. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference* (pp. 217-222). Menlo Park, CA: AAAI Press.

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

Ertmer, P. A., Bai, H., Dong, C., Khalil, M., Park, S. H., & Wang, L. (2002). Online professional development: Building administrators' capacity for technology leadership. *Journal in Computing Teacher Education*, **19**, 5-11.

Glaser, B. G., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Harris Wright, H., Silverman, S. W., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology*, **17**, 443-452.

Herdan, G. (1964). *Quantitative linguistics*. London: Butterworths.

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

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary & Linguistic Computing Bulletin*, **7**, 172-177.

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

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

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, **56**, 1-15.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

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

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

McCarthy, P. M., Dufty, D., Hempelman, C., Cai, Z., Graesser, A. C., & McNamara, D. S. (in press). Evaluating givenness/newness. *Discourse Processes*.

McCarthy, P. M., & Jarvis, S. (2007). A theoretical and empirical evaluation of vocd. *Language Testing*, **24**, 459-488.

McCarthy, P. M., Myers, J. C., Briner, S. W., Graesser, A. C., & McNamara, D. S. (2009). A psychological and computational study of genre recognition. *Journal for Language Technology & Computational Linguistics*, **24**, 23-55.

McEnery, T. (2003). Corpus linguistics. In R. Mitkov (Ed.), *Handbook of computational linguistics* (pp. 448-463). Oxford: Oxford University Press.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary & Linguistic Computing*, **15**, 323-337.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, **27**, 57-86.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (in press). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*.

Miller, D. P. (1981). The depth/breadth trade-off in hierarchical computer menus. In *Proceedings of the Human Factors Society 25th Annual Meeting* (pp. 296-300). Santa Monica, CA: HFES.

Morse, J. M. (1995). The significance of saturation. *Qualitative Health Research*, **5**, 147-149.

Olney, A. M. (2007). Latent semantic grammar induction: Context, projectivity, and prior distributions. In R. Dragomir & R. Mihalcea (Eds.), *Proceedings of TextGraphs-2: Graph-based algorithms for natural language processing* (pp. 45-52). Rochester, NY: Association for Computational Linguistics.

Ong, A. D., & van Dulmen, M. H. M. (2006). *Oxford handbook of methods in positive psychology*. Oxford: Oxford University Press.
Orlov, Y. K. (1983). Ein Model der Häufigekeitsstruktur des Vokabulars. In H. Guiter & M. V. Arapov (Eds.), *Studies on Zipf's law* (pp. 154-233). Bochum: Brockmeyer.
Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech & Hearing Research*, **45**, 927-937.
Silverman, S. W., & Bernstein Ratner, N. (2000). Word frequency distributions and type–token characteristics. *Mathematical Scientist*, **11**, 45-72.
Somers, H. H. (1966). Statistical methods in literary analysis. In J. Leeds (Ed.), *The computer and literary style* (pp. 128-140). Kent, OH: Kent State University.
Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.
Tuldava, J. (1993). The statistical structure of a text and its readability. In L. Hrebícek & G. Altmann (Eds.), *Quantitative text analysis* (pp. 215-227). Trier: Wissenschaftlicher Verlag.
Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers & the Humanities*, **32**, 323-352.
Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.
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.

NOTES

1. The program *vocd* produces an LD index called *D*. We use "vocd-D" to refer to the combination of the program and the output. MTLD (the measure of textual lexical diversity) is an LD assessment tool, producing an output index that also is referred to as MTLD.

2. The full text can be retrieved from http://showcase.netins.net/Web/creative/lincoln/speeches/gettysburg.htm (01/21/2010).

3. We used 42 because it is halfway between 35 and 50, which is the range used in vocd-D. McCarthy and Jarvis (2007) demonstrated that any number between 35 and 50 would be equally suitable.

4. In McCarthy and Jarvis (2007), the Maas index is referred to as *M*. More correctly, it is a^2, although we refer to it here simply as Maas to avoid confusion.

5. This issue is discussed in greater detail in the Rationale for MTLD section.

6. The LOB *science fiction* register includes only six texts. As such, three additional texts were added from the Brown corpus (Kučera & Francis, 1967).

(Manuscript received November 11, 2009;
revision accepted for publication February 7, 2010.)