AJSLP

Measuring Lexical Diversity in Narrative Discourse of People With Aphasia

Gerasimos Fergadiotis,[a] Heather H. Wright,[a] and Thomas M. West[a]

Purpose: A microlinguistic content analysis for assessing lexical semantics in people with aphasia (PWA) is lexical diversity (LD). Sophisticated techniques have been developed to measure LD. However, validity evidence for these methodologies when applied to the discourse of PWA is lacking. The purpose of this study was to evaluate four measures of LD to determine how effective they were at measuring LD in PWA.

Method: Four measures of LD were applied to short discourse samples produced by 101 PWA: (a) the Measure of Textual Lexical Diversity (MTLD; McCarthy, 2005), (b) the Moving-Average Type-Token Ratio (MATTR; Covington, 2007), (c) D (McKee, Malvern, & Richards, 2000), and (d) the Hypergeometric Distribution (HD-D; McCarthy & Jarvis, 2007). LD was estimated using each method, and the scores were subjected to a series of analyses (e.g., curve-fitting, analysis of variance, confirmatory factor analysis).

Results: Results from the confirmatory factor analysis suggested that MTLD and MATTR reflect LD and little of anything else. Further, two indices (HD-D and D) were found to be equivalent, suggesting that either one can be used when samples are >50 tokens.

Conclusion: MTLD and MATTR yielded the strongest evidence for producing unbiased LD scores, suggesting that they may be the best measures for capturing LD in PWA.

Key Words: aphasia, lexical diversity, validity

The cardinal deficit in aphasia is anomia, which is difficulty retrieving a word during discourse or in structured tasks (Goodglass & Wingfield, 1997). For this reason, addressing word-finding deficits has attracted considerable attention in aphasiology. One of the microlinguistic content analyses available for assessing lexical semantics in people with aphasia (PWA) is lexical diversity (LD). LD can be defined as "the range of vocabulary deployed in a text by a speaker that reflects his/her capacity to access and retrieve target words from a relatively intact knowledge base (i.e., lexicon) for the construction of higher linguistic units" (Fergadiotis & Wright, 2011, p. 1,415).

LD has been used in several applications in aphasiology. For example, it has been used to differentiate PWA from neurologically intact adults (Holmes & Singh, 1996), to measure selected aspects of semispontaneous discourse in PWA in order to capture clinically relevant aspects of noun and verb production (Lind, Kristoffersen, Moen, & Simonsen, 2009), and to examine whether and to what extent LD differs across adults with fluent and nonfluent aphasia (Wright, Silverman, & Newhoff, 2003). LD has also been used as an external criterion to investigate the concurrent validity of the story retell procedure (Doyle et al., 1998), which was designed to elicit language samples in PWA (McNeil et al., 2007). Besides clinical applications, LD has been used as a key variable for theory testing and development. Gordon (2008) studied the productive vocabulary of individuals with fluent and nonfluent aphasia in the context of the "division of labor" hypothesis (Gordon & Dell, 2003). Crepaldi et al. (2011) used LD to assess the predictions of neuropsychological models that may give rise to the characteristic differential noun–verb impairment in aphasia. Finally, LD has been used to investigate the efficacy and generalization to discourse of semantic feature analysis (Rider, Wright, Marshall, & Page, 2008).

LD has also been used in other areas of speech-language pathology, such as to track language development in children with cochlear implants (Ertmer, Strong, & Sadagopan, 2002), to study and diagnose specific language impairment (Owen & Leonard, 2002; Thordardottir & Namazi, 2007), and to differentiate bilingual children with and without specific language impairments (Kapantzoglou, 2012; Klee, Gavin & Stokes, 2007). Despite its widespread use in speech-language pathology, both in the field of aphasiology and other fields, identifying a robust measure to capture LD has been very challenging (Malvern, Richards, Chipere, & Durán, 2004). In this study, we collected and examined validity evidence for four computational tools that

[a] Arizona State University, Phoenix

Correspondence to Gerasimos Fergadiotis: gfergadiotis@pdx.edu

Editor: Swathi Kiran
Associate Editor: Jamie Reilly
Received July 28, 2012
Revision received October 15, 2012
Accepted January 28, 2013
DOI: 10.1044/1058-0360(2013/12-0083)

have been proposed for measuring the LD of a language sample, and we evaluate their appropriateness and applicability to aphasic discourse.

Measuring LD

One of the most commonly used approaches to measure LD is to use the ratio of unique lexical items divided by the total number of words in a sample (type-token ratio, TTR; Chotlos, 1944; Templin, 1957) after standardizing the length of the sample.[1] TTR is inherently flawed because it varies as a function of sample length. As the sample length increases, it is less probable that a speaker will produce new words because the number of lexical items that can be activated at any given time is considered finite (Heap, 1978).

When using TTR to gauge LD, shorter samples often appear to be richer, rendering comparisons across speakers who produce language samples of different lengths problematic. Researchers have attempted to address this issue by proposing a standardized sample size, but this approach has not produced satisfactory results. A major limitation is that in order for results to be comparable across studies, researchers have to agree on the number of tokens required to estimate the TTR. In aphasiology, some researchers have proposed 300 tokens as a standard length (Brookshire & Nicholas, 1994; Prins & Bastiaanse, 2004). However, consensus on this issue is primarily low because it is not always feasible to obtain a predetermined number of tokens. Individuals with aphasia often do not produce long samples; subsequently, researchers truncate samples based on the shortest sample in the study (e.g., Gordon, 2008).

Recently, a new generation of tools for measuring LD has emerged from the field of computational linguistics. These tools have been designed to produce length-invariant estimates of LD without discarding any data. With a few exceptions, these measures have been used in limited applications in the field of speech-language pathology.

The Measure of Textual Lexical Diversity (MTLD; McCarthy, 2005) employs a sequential analysis of a sample to estimate an LD score. Conceptually, MTLD reflects the average number of words in a row for which a certain TTR is maintained. To generate a score, MTLD calculates the TTR for increasingly longer parts of the sample. Every time the TTR drops below a predetermined value, a count (called the *factor count*) increases by 1, and the TTR evaluations are reset. The algorithm resumes from where it had stopped, and the same process is repeated until the last token of the language sample has been added and the TTR has been estimated. Then, the total number of words in the text is divided by the total factor count. Subsequently, the whole text in the language sample is reversed and another score of MTLD is estimated. The forward and the reversed MTLD scores are averaged to provide the final MTLD estimate.

The Moving Average Type Token Ratio (MATTR; Covington, 2007; Covington & McFall, 2010) measures LD by calculating TTRs for successive nonoverlapping segments of a sample. The algorithm selects a window length of x tokens, and the TTR for tokens 1 to x is estimated. Then, the TTR is estimated for tokens 2 to (x +1), then 3 to (x + 2), and so on for the entire sample. The final score is the average of the estimated TTRs.

The D (Malvern & Richards, 1997; McKee, Malvern, & Richards, 2000) generates LD scores that conceptually reflect how fast TTR decreases in a sample. If a language sample consists of types that are being used repeatedly, TTR would decrease faster as a function of the sample size. The D performs a series of random text samplings to plot an empirical TTR versus number-of-tokens curve for a sample. Thirty-five tokens are randomly drawn from the sample without replacement, and the TTR is estimated. This process is repeated 100 times, and the average TTR for 35 tokens is estimated and plotted. The same routine is then repeated for subsamples of 36 to 50 tokens. The average TTR for each subsample of increasing token size is subsequently plotted to form the empirical curve. Then, the least squares approach is used to obtain an estimate of D that produces a theoretical curve that maximizes the fit to the empirical TTR curve. Lower D values result in steeper theoretical curves that fit the empirical curves of samples with poorer LD. The whole process is repeated three times, and the final D value is the average of the three runs.

Recently, McCarthy and Jarvis (2007) argued that D might be related to probabilities of word occurrence that can be modeled using the hypergeometric distribution (HD). The HD is a discrete probability distribution that expresses the probability of k successes after drawing n items from a finite population of size N containing m successes *without* replacement. For example, if a container contains m white marbles and $N - m$ black marbles (total number of marbles $= N$), and drawing a white marble is defined as a success, the HD gives the probability of drawing k white marbles after n draws without replacement.

McCarthy and Jarvis (2007) used the HD to create a new measure of LD called the HD-D. The assumption underlying HD-D is that if a sample consists of many tokens of a specific word, then there is a high probability of drawing a sample that will contain at least one token of that word. McCarthy and Jarvis reported strong linear correlations between HD-D and D scores in two studies (McCarthy & Jarvis, 2007, $r = .97$; McCarthy & Jarvis, 2010, average $r = .91$ across several types of discourse evaluated in the study). Based on these findings, McCarthy and Jarvis argued that D is an approximation of HD-D expressed in a different metric. Further, they attributed the less than perfect correlations between the two measures to the main difference in the nature of the two measures—the fact that D is based on random sampling and curve fitting, which introduces error in the estimation process, as opposed to HD-D, which

[1] Various transformations of TTR have also been attempted (Carroll, 1964; Engber, 1995; Guiraud, 1960; Herdan, 1960), some of which have been applied to aphasic discourse (e.g., Prins, Snow, & Wagenaar, 1978; Wachal & Spreen, 1973). However, these attempts have been unsuccessful (e.g., Tweedie & Baayen, 1998; Vermeer, 2000).

is directly estimated based on probabilities of word occurrence in a language sample.

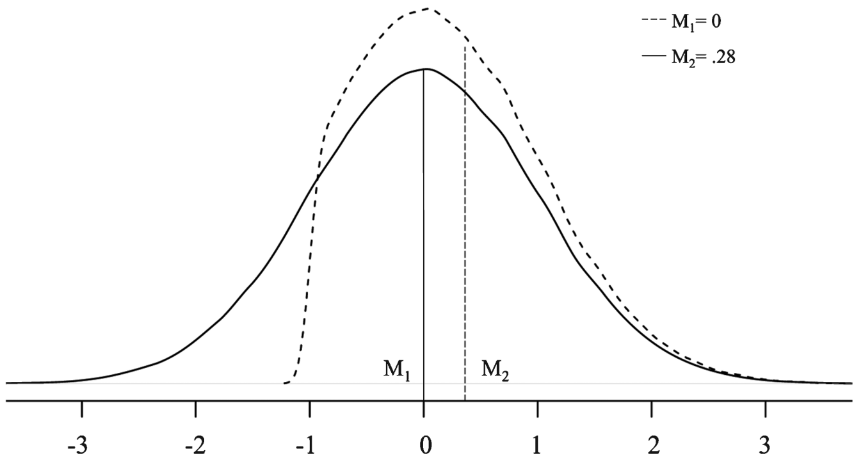
A feature of HD-D is that it does not require a minimum of 50 tokens to be estimated. By default, D is required to estimate the average TTR for 50-token subsamples in order to establish the empirical curve that is modeled. If there are <50 tokens in the sample, the program terminates without providing a score for the specific sample. This is problematic for researchers who work with PWA (and other clinical populations), who often produce limited verbal output. The reason is twofold. First, language samples with <50 tokens that are discarded may lead to a loss of valuable information. Typically, we reach more robust conclusions about a client's language skills the more data are available. Second, from a missing data theory perspective, only when data are *missing at random* or *completely at random* (both terms introduced by Rubin in 1976) is the missing mechanism ignorable. Conversely, if the data are *missing not at random,* and this fact is ignored and the data are analyzed, statistical parameter estimates may include substantial bias that may lead to invalid inferences (Enders, 2010; Graham, 2009; Little & Rubin, 2002; Rubin, 1976).

To illustrate this point, Figure 1 shows two probability density functions for a standardized variable X (e.g., height). The first curve corresponds to a simulated complete data set of 1,000 observations, and M_1 is its mean. When data are missing not at random, the probability that data are missing depends on the unseen scores themselves. The second curve corresponds to such a truncated data set for which data are missing not at random and missingness is related to an observation's value: Values <–1 are not observed systematically. This scenario could occur for instance if our sample consisted of U.S. fighter pilots who have a minimum standing height requirement of 64 in. In this case, M_2 would be a biased estimator of the mean height of the general population. Similarly, if LD scores are not missing in a haphazard fashion when estimated with D, but instead are missing in a systematic way, parameter estimates may also be biased. Further, even though Figure 1 demonstrates the effect of systematic missingness on a mean, the same logic applies to the estimation of any statistical model parameter (e.g., variances, covariances, factor loadings, regression coefficients). Finally, notice that if the data points were missing at random (e.g., some from the lower end of the distribution and some from the higher end of the distribution with equal probability), the estimated mean would not have been biased despite the missing values.

Validity evidence for MTLD, MATTR, and D. McCarthy and Jarvis (2010) explored the convergent and divergent validity of MTLD scores. Using indices that have been used in the past to estimate LD, such as Maas (i.e., a TTR transformation; Maas, 1972), D, and Yule's K (a probabilistic index; Yule, 1944), McCarthy and Jarvis found that MTLD correlated moderately to strongly with all three indices: Correlations were –.84, .69, and .85, respectively. McCarthy and Jarvis argued that based on these results, convergent validity was supported. MTLD did not correlate strongly with TTR, which is influenced by length (r = .32; r^2 = .10). This finding was interpreted as evidence of divergent validity.

Fergadiotis (2011) examined a corpus of four types of discourse (i.e., procedural discourse, eventcasts, storytelling, recounts) from 442 neurologically intact adults in order to evaluate four LD techniques: D, the Maas index, MTLD, and MATTR. LD scores were estimated for each type of discourse and were modeled using structural equation modeling to uncover their latent structure. Across all four types of discourse, the highest loadings were associated with the observed variables that were estimated using the MATTR and MTLD variables (median λ_{MATTR} = .96, median λ_{MTLD} = .94), followed by D (median λ_D = .85) and Maas (median λ_{Maas} = –.55). Results suggested that MATTR, MTLD, and D were strong indicators of LD. However, for the D- and Maas-generated variables, the

Figure 1. Two probability density functions for (a) a complete data set and (b) the same data set for which data are missing not at random from the lower end of the distribution. M_2 is a biased estimate of the population mean.



results were consistent with the presence of method factors that represented the influence of construct-irrelevant sources. By experimentally manipulating the samples through truncation, Fergadiotis demonstrated that the irrelevant source (i.e., method factors) influencing the measures was associated with length.

The validity of D-score interpretations has been explored in several studies (Malvern & Richards, 1997; Malvern et al., 2004; Richards & Malvern, 1997). Estimates of D correlated strongly with well-validated measures of language and also developmental and demographic variables. However, the validity of D score interpretations has been questioned. For example, Owen and Leonard (2002) found that mean LD, as estimated by D, was found to vary as a function of sample length: Samples that were truncated to 250 words had a significantly lower mean D score compared to samples that were truncated to 500 words. Owen and Leonard concluded that "it appears that D does not entirely avoid the problem of sample size influence" (p. 935).

Statement of the Problem

Several sophisticated techniques have been developed recently to address the limitations of flawed LD measures such as TTR. Although these methods assert to measure LD, each one is based on its own theoretical assumptions, which are reflected in the computational machinery they employ. Therefore, it is not clear whether these techniques measure the same construct and to what extent they produce valid and reliable scores. Further, D's minimum-50-tokens-per-sample requirement raises questions regarding the measure's applicability to PWA. The purpose of this study was to examine the validity of score interpretations from four computational measures of LD when they are applied to aphasic discourse. Our specific aims were to:

- Investigate the relationship between D and HD-D to determine whether clinicians and researchers can use them interchangeably.
- Assess if the minimum requirement of 50 tokens to estimate D may lead to biased estimates that may result in invalid conclusions about patients' language skills.
- Assess whether all techniques (MTLD, MATTR, D, HD-D) measure the same latent variable and to what extent.
- Examine whether there is a single latent variable determining performance for each estimation technique or whether there is evidence consistent with the presence of systematic residual covariance that jointly determines the scores that could undermine validity.

Method and Results

Participants

Data from 101 monolingual PWA were included in this study. Their data were retrieved from AphasiaBank (MacWhinney, Forbes, Fromm, & Holland, 2011), which is an online, shared database that collects and analyzes digital recordings of discourse from PWA across a series of tasks. All of the participants had acquired aphasia secondary to a single left-hemisphere stroke. Inclusion criteria and the sample's characteristics including gender, age, race/ethnicity, years of education, average performance on the Boston Naming Test (Goodglass, Kaplan, & Barresi, 2001), and Western Aphasia Battery—Revised (WAB–R; Kertesz, 2007) aphasia quotient classification, are presented in Table 1.

Discourse Elicitation and Data Preparation

Stimuli and instructions. Discourse samples were collected in a single session, and several tasks for eliciting discourse were employed, including personal narratives, sequential pictures, single pictures, and telling of the story of *Cinderella* (Grimes, 2005). For this study, only the samples based on the Cinderella story were analyzed. To elicit the story, participants were presented with the wordless stimulus book *Cinderella* (Grimes, 2005). They were told to look through the book to remember how the story goes, and they were allowed as much time as desired to view it. Then, the book was taken away, and the participants were asked to tell as much of the story as they could. The examiners used standard written scripts to keep verbal instruction and prompts consistent across testing sites. Further, the examiners were instructed to remain silent as much as possible during administration of the task while also providing as much nonverbal encouragement as possible.

Transcription and language sample preparation. Samples were digitally recorded and orthographically transcribed in the CHAT format that is compatible with a set of programs called CLAN (MacWhinney, 2000). Words were

Table 1. Study participants' demographic information.

Characteristic	Participants (*N* = 101)
Gender ratio[a]	56M:43F
Age in years	63.09 (11.32)
Ethnicity[a]	
African American	7
Asian	3
Hispanic	3
Other	4
White	80
Education level completed[a]	
Some high school	4
12th grade	21
Some college	17
Bachelor's or higher	55
Aphasia duration in years	6.04 (5.45)
BNT	8.45 (4.30)
WAB–R AQ	76.20 (15.40)

Note. *SD*s are shown in parentheses. BNT = Boston Naming Test (Goodglass, Kaplan, & Barresi, 2001; WAB–R AQ = Western Aphasia Battery—Revised (Kertesz, 2006) aphasia quotient.

[a]Gender information was unavailable for two individuals, ethnicity information was unavailable for four individuals, and education information was unavailable for four individuals.

tagged morphosyntactically, and function words were removed from the samples because they have little or ambiguous meanings and convey predominantly grammatical relationships. As a result, only content words (i.e., nouns, verbs, adjectives and *–ly* adverbs) were analyzed. Further, to avoid conflating LD with grammaticality, a lemma-based analysis was performed. Based on Kiparsky's (1982) levels of morphological derivation, level three inflections were disregarded (e.g., *eat, eats, ate* = *eat*). Repetitions, repairs, fillers, and paraphasias were coded and were subsequently excluded from analysis. The average number of content words of the analyzed samples was 84.94 (*SD* = 53.28).

Estimating LD. Five measures were applied to the language samples to estimate LD. They included MTLD, MATTR, D, HD-D, and TTR (the last measure was added to address the goals of specific Aims 3 & 4; see next section). MTLD, HD-D, and TTR were estimated using a stand-alone application tool, the Gramulator 5.0 (McCarthy, Watanabe, & Lamkin, 2012). D was estimated using the voc-D program in CLAN. Finally, MATTR was estimated using computer software that was developed by Covington (2007). To avoid missing data, the length of the MATTR window was set to 17 tokens, which was the minimum number of tokens in the language samples.

Relationship of D and HD-D and the Impact of Missing Data

Statistical approach: Curve fitting and analysis of variance (ANOVA). To determine whether D and HD-D can be used interchangeably to measure LD, we analyzed these scores using the R statistical language (R Development Core Team, 2011) and SPSS 20. The nonlinear least square fit function (Bates & Watts, 1988) into the "stats" package of R was used to fit a linear and an exponential curve. Model fit was based on visual and algebraic information. To determine if the minimum requirement of 50 tokens to estimate D will lead to biased estimates, we conducted a one-way ANOVA in SPSS 20.

Preliminary analysis. Language samples from 101 PWA were included in the study. Data were prepared for statistical analysis following Kline (2010) and Tabachnick and Fidell (2007). Descriptive statistics for the number of types, tokens, and TTRs, as well as the estimated LD indices, are provided in Table 2. After being imported into SPSS, the data were screened for missing values. All variables had complete data except for D, for which ~27% of the data were missing due to an insufficient number of tokens in the samples. Distributions were visually inspected and assessed in terms of the normality assumption; skewness and kurtosis statistics were estimated. Several distributions were noted to be skewed and with various degrees of kurtosis. For this reason, the maximum likelihood ratio (MLR) estimator was used, which estimates parameters using maximum likelihood with standard errors and a χ^2 test statistic that are robust to nonnormality.

Relationship between D and HD-D: Results. To investigate the relationship between D and HD-D, SPSS 20 and R were used to fit a linear and an exponential curve to the data. First, a linear regression analysis was performed to predict D from HD-D for the participants with complete data. The linear regression equation was significant, R^2 = .85, F(1, 72) = 362.69, p < .001. Then, an exponential model was fit to the same data. The exponential equation was also significant, R^2 = .99, F(1, 72) = 15972.94, p < .001. However, the difference in variance explained using the exponential model was substantial (ΔR^2 = .14), which suggests that the exponential model better captured the relationship between D and HD-D than the linear model. The two models were further compared using the residual sum of squares, the residual standard error, and the standard errors of the model parameters. Based on the algebraic information, the exponential model demonstrated considerably better fit (see Table 3). Finally, the data were plotted along with the linear and exponential curves suggested by the models, and fit was evaluated visually (Figure 2). Overall, the exponential model demonstrated better fit compared to the linear model. Substantively, the excellent fit of the exponential model to the bivariate data (graphically and in terms of the R^2) provides strong evidence that HD-D and D are essentially isomorphic; that is, using the exponential function, one can estimate D scores based on observed HD-D with excellent accuracy.

voc-D and missing data: Results. The curve estimation analysis was followed up by an ANOVA in SPSS to explore

Table 2. Descriptive statistics of the major study variables.

Variable	*N*	*M*	*SD*	Range	Skewness	Kurtosis
Types	101	42.44	21.15	11–108	–0.80	–0.48
Tokens	101	83.08	53.15	17–273	–1.17	–2.85
TTR	101	00.55	00.11	.27–.80	–0.01	–0.09
D	74	31.55	14.88	7.20–90.28	–1.17	–2.85
HD-D	101	–7.78	04.46	–23.83–.76	–0.94	–1.22
MTLD	101	25.11	09.81	10.45–65.89	–1.32	–2.96
MATTR	101	0.76	00.08	.49–.92	–0.56	–0.77

Note. TTR = type-token ratio; D (McKee, Malvern, & Richards, 2000); HD-D = Hypergeometric Distribution D (McCarthy & Jarvis, 2007); MTLD = Measure of Textual Lexical Diversity (McCarthy, 2005); MATTR = Moving-Average Type-Token Ratio (Covington, 2007).

Table 3. Model summaries and parameter estimates.

Model	*R*	R^2	*F*	RSS	Residual *SE*	*a* (*SE*)	*b* (*SE*)
Linear	0.92*	0.85	392.69	2504	5.9	56.87 (1.45)	3.82 (0.19)
Exponential	0.99*	0.99	15972.94	129.1	1.34	75.08 (0.59)	0.14 (<0.01)

Note. RSS = residual sum of squares; SE = standard error.
*$p < .01$.

whether the missing D data were missing not at random. Using the exponential function, D scores were computed based on HD-D for the language samples with <50 tokens, for which D in CLAN was not able to produce estimates (solid data points in Figure 2). Henceforth, voc-D generated scores and imputed D scores based on HD-D will be referred to as D_{VOC} and D_{IMP}, respectively, and the combined variable $D_{VOC} + D_{IMP}$ will be referred to as D_{COM}. Also, a binary missingness variable was created that denoted whether D scores were voc-D generated or were missing and estimated from HD-D. To assess the assumption of homogeneity of variance, Levene's test of equality of error variances was performed, and it was not statistically significant, $F(1, 99) = .62$, $p = .43$. A between-subjects one-way ANOVA was conducted with D_{COM} as the dependent variable and missingness (voc-D values present or not) as the independent variable. There was a significant effect of missingness, $F(1, 99) = 14.45$, $p = <.001$, $\eta^2 = .13$, Cohen's $d = .88$.

The box plot in Figure 3 shows graphic summaries for D_{VOC}, D_{IMP}, and D_{COM}. Results suggested that the probability of D scores missing was related to the values of D; that is, the missing D scores were more likely to be lower than the scores that were estimated with voc-D. Figure 3 shows the impact of this finding in the current data set. If the number of tokens in a sample did not limit calculating D, the estimate of the mean would have been approximately equal to 28.24 (D_{COM}). However, the estimation process was more likely to fail for language samples that had lower LD. As a result, values from the lower end of the distribution were eliminated systematically, and the mean estimate of D in our sample using the voc-D program alone was biased upward ($D_{VOC} = 31.55$).

Investigating Construct Validity

After addressing the first two questions, we wanted to determine whether all of the techniques (i.e., MTLD,

Figure 2. Bivariate scatter plot of the scores generated by vocD and HD-D with fitted linear and exponential curves. Solid dots represent scores that were imputed using the exponential model.

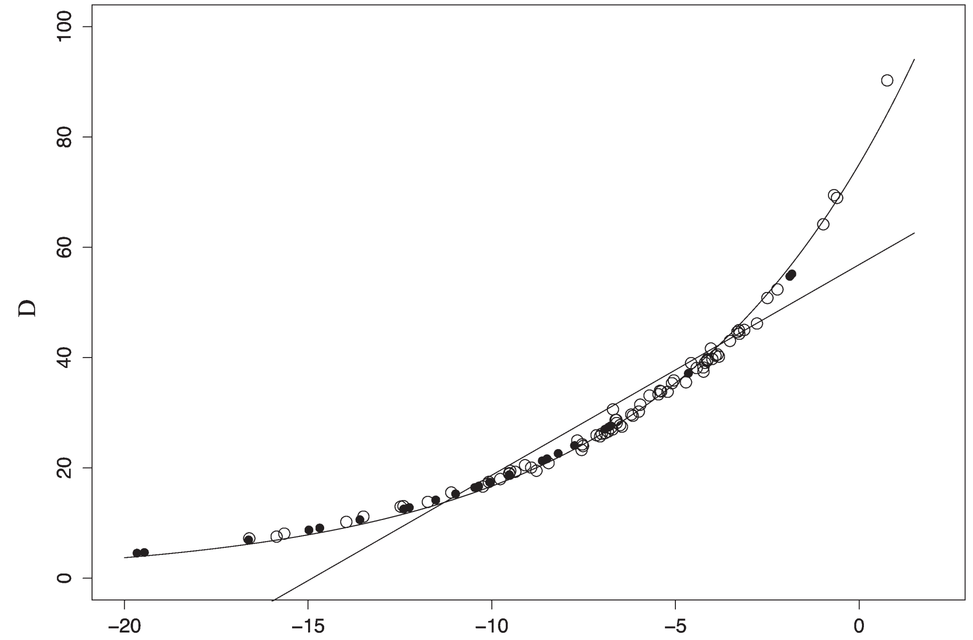
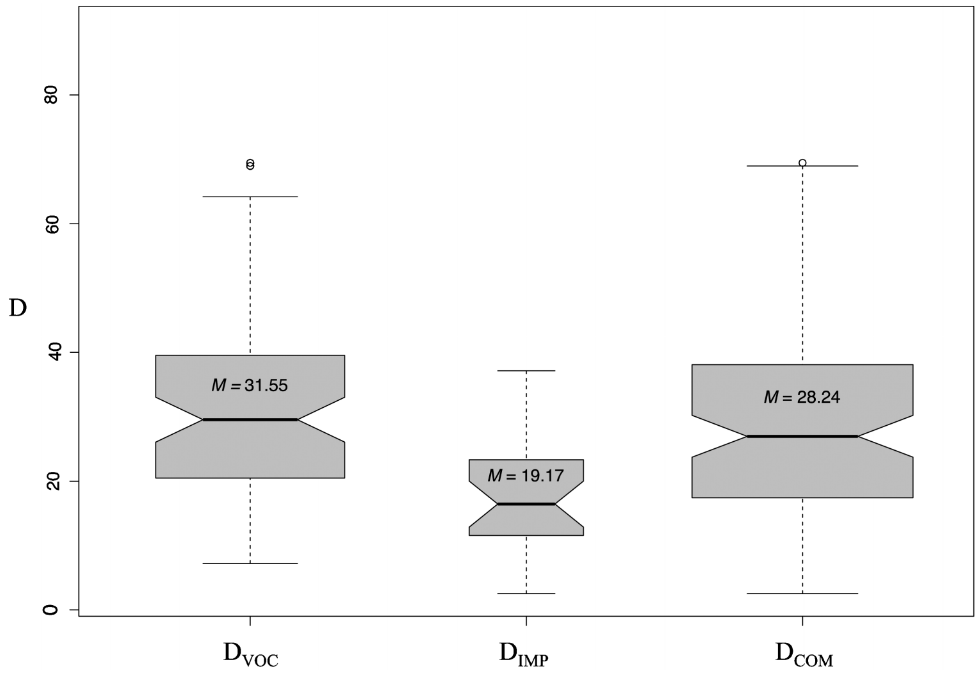


Figure 3. Boxplots of D scores generated with vocD, imputed with HD-D, and combined. The width of the plots is proportional to the square root of the samples sizes. The black line in each box indicates the median for each set of scores. If the notches of two plots do not overlap, this is ''strong evidence'' that the two medians differ (Chambers, Cleveland, Kleiner, & Tukey, 1983).



MATTR, D, HD-D) measure the same construct and whether there was evidence consistent with the presence of systematic residual covariance that would indicate length effects. Because the data generated by voc-D were not missing at random (and would therefore bias the model parameter estimates), we imputed D values for the missing data points from HD-D scores and the exponential function used earlier.

Statistical approach: Confirmatory factor analysis (CFA). The LD of a sample was conceptualized as an unobserved latent variable, and its relationship with four observed variables (MTLD, MATTR, D, and TTR) was modeled using CFA in Mplus 6. Two models that reflected competing hypotheses were specified a priori; they were evaluated in terms of global fit and localized areas of strain and were compared directly using a χ^2 difference test. Following Bollen (1989), the magnitude of the standardized loadings and error covariances from the best fitting model were used to compare the relative influence of the factor on the manifest variables and to answer the substantive questions of the paper.

The first model assumed a single latent common factor ξ that was interpreted as the mathematical instantiation of LD in a sample. The loadings, λ_{MTLD}, λ_{MATTR}, λ_D, and λ_{TTR} indicated how strongly the latent variable influenced the observed variables, or, alternatively, how strongly a score generated by a given technique reflected the LD of a sample. Finally, the model stipulated that once the effect of the common factor was taken into account, there was no systematic covariance among the residual terms of the observed indicators. Given that TTR *is* influenced by sample length, the lack of residual term covariances reflected the hypothesis that MTLD, MATTR, and D did not share TTR's susceptibility to length effects. In other words, TTR was used not *in spite of* but *because of* its known length dependency to help us uncover similar behavior in other indices. The second model was identical to the first except that the error covariance between the TTR and D error terms was allowed to be freely estimated. This specification was consistent with the findings of Owen and Leonard (2002) and Fergadiotis (2011), who suggested that D might be influenced by length. Therefore, allowing the error covariance to be estimated modeled the specific hypothesis that D, similar to TTR, is influenced by a construct unrelated to LD, most possibly, length.

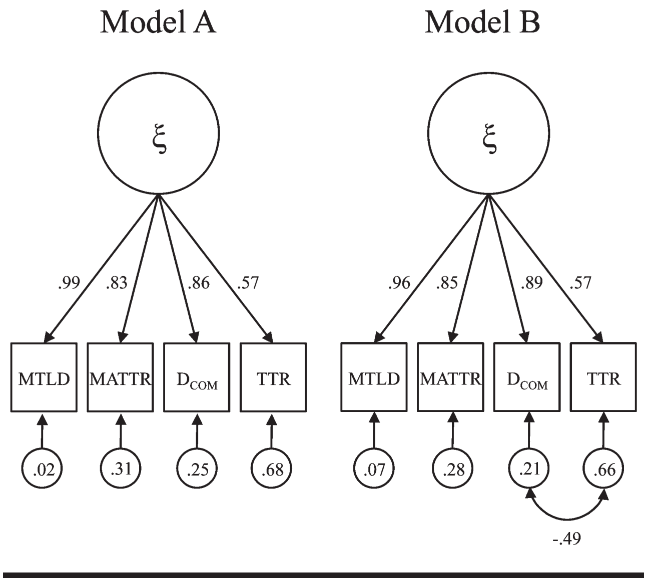
Investigating construct validity: Results. The models were estimated in Mplus 6.1 using the MLR estimator. Four fit indices were taken into account to examine global model fit. Fit indices included the Satorra-Bentler scaled χ^2 statistic (Satorra & Bentler, 1994) to take into account the nonnormality of the data, the comparative fit index (CFI; Bentler, 1990), the root-mean square error of approximation (RMSEA; Steiger & Lind, 1980), and the standard root mean residual (SRMR; Hu & Bentler, 1998). A good fitting model was expected to have a nonsignificant χ^2 at the .05 level; a CFI value >.95; an RMSEA value <.08, with the upper bound of the 90% confidence interval <.10; and an SRMR value <.08 (Brown, 2006; Hu & Bentler, 1999; Kline, 2010;

Steiger, 2007). To assess for local strains in the solution, modification indices and normalized residuals were considered. Two hypotheses using nested model comparisons were tested followed by a substantive evaluation of model parameters. To perform the nested model comparisons, the scaled difference χ^2 test statistic (Satorra & Bentler, 2001) was used.

First, a unidimensional CFA model with four indicators (MTLD, MATTR, D_{COM}, TTR) was fit to the data (see Model A in Figure 4). The covariances among the residual terms of the indicators were fixed to 0. To identify the model, the variance of the latent variable was set equal to 1. The model converged to a solution with no out-of-range parameter values for which the fit indices provided mixed evidence of adequate model fit: $\chi^2(2, N = 101) = 17.74$, $p < .001$; CFI = .93; RMSEA = .30 (90% confidence bands = .19 and .42); and SRMR = .04. The largest normalized residual was associated with the covariance of TTR and D_{COM} (–1.23). Also, based on the modification indices, allowing the residual variances of TTR and D_{COM} to covary would improve the model fit significantly (approximate $\Delta\chi^2 = 14.44$; expected parameter change = –.45).

In the next step, a model with a single factor defined by the four indicators (MTLD, MATTR, D_{COM}, TTR) was fit to the same data (see Model B in Figure 4). The model was identical to Model A with one exception: The TTR and D_{COM} residual terms were allowed to covary. The model converged to a solution for which all fit indices suggested excellent fit to the data, $\chi^2(1, N = 101) =.38$, $p = .54$; CFI = 1.00; RMSEA = .00 (90% confidence bands = .00 and .22); and SRMR = .007. Similarly, no local model strain was noted (highest normalized residual = –.2; no modification indices with values >3.84). Also, none of the standardized parameter estimates took on out-of-range values, and all of the estimates were statistically significant. The two models were further compared using the scaled difference χ^2 test statistic to explore whether fixing the TTR and D_{COM} covariance to 0 had a statistically significant impact on global fit. Based on the results, the null hypothesis that Model A did not fit significantly worse than Model B was rejected, $\Delta\chi^2(1) = 23.02$, $p < .001$. Figure 4 shows the parameter estimates for both models (Mplus output is available upon request). Based on the results, Model B demonstrated excellent global fit and a lack of localized misfit and statistically outperformed Model A.

Figure 4. Two alternative confirmatory factor analysis models of the lexical diversity scores based on language samples from 101 people with aphasia. Completely standardized robust maximum likelihood parameter estimates. For each observed variable, the variance accounted for by the common factor, R^2 = (1 – residual variance) = λ^2. For all parameter estimates, $p < .001$, except for MTLD's residual variance (p = .004).



Discussion

The main purpose of this paper was to collect validity evidence regarding techniques for measuring LD for the study of aphasic discourse. The specific aims were to investigate the relationship between D and HD-D; explore whether using D may lead to biased estimates when studying aphasic discourse; determine whether MTLD, MATTR, and D measure the same latent variable and to what extent; and examine whether there is evidence of length effects for the aforementioned LD indices. In a series of analyses, HD-D and D were found to correlate highly using an exponential function. Also, statistically significant mean differences were uncovered between scores that were estimated using vocD and the scores for which the vocD algorithm failed to produce scores due to inadequate number of tokens in the samples. Finally, a unidimensional CFA model of MTLD, MATTR, D, and TTR that allowed for the residual terms of TTR and D to correlate freely exhibited excellent fit to the data. To the best of our knowledge, this was the first time that missing data theory was employed to study the performance of the D index. Further, it was the first time that CFA was used to study the validity of the MTLD, MATTR, and D score interpretations.

The Relationship of HD-D and D and the Impact of Missing Data

One of the findings from the study was that an exponential function could be used to model the relationship between HD-D and D with very high accuracy. This finding was consistent with the hypothesis that HD-D and D may be equivalent indices. McCarthy and Jarvis (2007, 2010) reported a strong relationship between HD-D and D based on which they argued that D was an approximation of HD-D expressed in a different metric. They further argued that the correlations were less than perfect because the estimation of D using vocD, which involves random sampling and curve fitting, introduced error in the measurement. Findings from the current study lend more support to McCarthy and Jarvis' conclusion. However, as opposed to previous studies that used a linear model, we used nonlinear regression and modeled the data using an exponential curve that demonstrated substantially better and nearly perfect fit to the data. These findings indicated that the random error element in the estimation of D was considerably less than what had been suggested previously in the literature.

The finding of the close-to-perfect relationship between HD-D and D has significant implications for practical applications. Solely on the basis of computing a score, the correlation of the two indices suggested that the choice of which index to use may be arbitrary when measuring LD in language samples that have >50 tokens. One significant advantage of D for speech-language pathologists and researchers is that D is integrated in CLAN, which allows for great flexibility in coding, manipulation, and preparation of transcribed data and further offers a wide range of automatic analyses (MacWhinney, Fromm, Forbes, & Holland, 2011; MacWhinney, Fromm, Holland, Forbes, & Wright, 2010).

However, we also presented evidence that under certain conditions, the use of D may introduce bias in statistical analyses and so may lead to invalid conclusions. Specifically, we found that the probability of a missing D datum was related to its value: Lower D scores were more likely to be missing, thus biasing the mean upward. It could be argued that this difference may not be significant enough to warrant attention when using D. Two things should be noted that may be of particular importance to researchers who are interested in estimating and analyzing LD scores. First, results may be more biased under certain experimental designs. In this study, the difference between D_{VOC} and D_{COM} was small because ~70% of the observations in D_{COM} were included in both sets. The difference could be considered a lower bound of bias because the data have been aggregated. If an experimental design called for two groups, the group with the lower mean LD would have more values missing because the D estimation process is more likely to fail for language samples that have lower LD. As a result, the mean of the group with the lower LD would be biased upward (similar to the D_{VOC} data in Figure 3). The group with the higher mean LD would have less missing values, and therefore, its estimated mean would not differ as much. Overall, however, the mean difference between the two groups would shrink artificially, perhaps masking the real difference. Second, it is not clear how this bias would propagate if data were to be used in more complicated multivariate statistical techniques such as variants of canonical correlation or structural equation modeling approaches.

Evidence for Construct Validity

The unidimensional configuration of Model B supported the hypothesis that MTLD, MATTR, D, and TTR reflect the same latent variable despite their computational differences. The scores that were generated by the different techniques were found to be strong indicators of the latent variable. However, there was variation in the magnitude of the relationship between the factor and each of the indicators (see Model B, in Figure 4). Overall, based on the results of this study and holding everything else constant, MTLD appeared to provide the most accurate reflection of the LD of a sample. In contrast, ~⅔ of the variance in TTR reflected variance that was unrelated to the latent variable that represented the LD of a sample.

The findings from this study regarding MTLD confirmed and expanded previous results that had been reported in the literature. For example, in earlier studies, MTLD was found to correlate strongly with a number of LD indices, including D and Maas, leading researchers to argue in favor of MTLD's validity (e.g., McCarthy, 2005). However, the methodology that had been used previously to collect validity evidence regarding LD indices had relied primarily on the examination of correlational relationships with language samples from neurologically intact adults. In the current study, MTLD was entered in a model with a single factor that was formed by the common variance across four LD indices. The excellent fit of Model B to the data and its structure and parameters provided a more coherent and accurate representation of how observed scores from LD indices such as the MTLD may relate in this and other studies.

D_{COM} was also found to be a strong indicator of LD, but the interpretation of its scores may not be as straightforward as for MTLD. The proportion of variance in D_{COM} attributed to LD was ~79% (Figure 4), suggesting a strong relationship with the latent variable. Nevertheless, this estimate was considerably lower than the respective parameter value of MTLD (i.e., 93%) that was found in this and a previous study (Fergadiotis, 2011). Further, the residual variance of D_{COM} that represented the combination of random error and systematic variance that was irrelevant to LD was high relative to MTLD (21% and 7%, respectively). Therefore, in probabilistic terms, an MTLD score may convey more information about the LD of a sample than D in the sense that an MTLD score reflects primarily LD (93%) and little of anything else (7%). A score generated by D would reflect LD (79%) and a mixture of sample length effects and random noise (21%).

Specifically, the current study showed that D scores may have been determined by two sources. First, the scores were determined by a factor that influenced the scores across all four indices. Arguably, this factor represented the LD of a sample. But, unlike MTLD and MATTR, D_{COM} correlated with TTR, which is known to decrease as a function of sample length. Therefore, similar to previous studies (e.g. Owen & Leonard, 2002; McCarthy & Jarvis, 2007), Model B was consistent with the hypothesis that effects related to length may influence D. Further, based on the direction of the error covariance parameter ($cov_{DCOM*TTR} = -.49$), and given that TTR decreases as a function of length, Model B predicts that as sample length increases, D scores will also increase. Even though this systematic variance is only a fraction of the residual variance of D_{COM}, and therefore has an upper bound of 21%, it constitutes a second dimension along which D scores vary systematically.

Regarding MATTR, the CFA results indicated that this index was also a strong indicator of LD and therefore it may be useful for analyzing the LD of discourse produced by

PWA. A great advantage of MATTR is its face validity because it is equivalent to TTR and fairly straightforward to grasp and explain. Face validity is a very desirable property, especially for professionals who work with individuals with speech and language disorders in clinical settings (e.g., Gordon, 2008; Lind et al., 2009). MATTR does not require an understanding of frequency distributions, curve fitting, or the nature of stochastic processes in order to convey its meaning. Therefore, it may enable more meaningful communication between clinicians, patients, and their families.

The magnitude of MATTR's relationship with the latent factor was smaller compared to MTLD's and comparable to D_{COM}'s (λ_{MATTR} = .85; see Figure 4). This finding contrasted with the results in Fergadiotis (2011), where MATTR was the strongest indicator of LD when assessing language samples from neurologically intact adults (average λ_{MATTR} across four types of discourse = .95). This discordant result may have been due to methodological differences between the two studies. Unlike Fergadiotis, the current study employed a lemma-based analysis of content words from language samples from PWA. To avoid missing data, the length of the window that MATTR used to estimate LD was set equal to the number of tokens in the shortest language sample included in the study (17). In contrast, the size of the window in the Fergadiotis study was equal to 50 tokens. It is possible that using a smaller window in the current study forced MATTR to estimate LD scores based on less information, thus becoming more susceptible to random fluctuations. One way to test this hypothesis in future studies would be to experiment with the size of the window in MATTR and observe whether using smaller window sizes would cause shrinkage of the loading of MATTR on the LD factor.

Finally, the residual variance of MATTR, similar to MTLD's, did not correlate with TTR's residual variance. Once the variance accounted for by the common factor in MATTR and TTR was partialled out, the two manifest variables were conditionally independent. In other words, the CFA solution assumed that TTR and MATTR shared one and only one common cause—the LD of the sample. Given TTR's known flaw to vary with sample length, the lack of covariance between the residual terms of these two variables constitutes evidence that MATTR may be a length-invariant measure.

Taken together, the results from the CFA carry important practical implications. The findings that (a) MTLD and MATTR were strong indicators of LD, (b) they did not show evidence of systematic length effects, and (c) their residual variances were very small, constitute evidence in favor of the validity of their score interpretations. It has been argued that the measurement of psychological constructs is a process of evidentiary reasoning (Mislevy & Yin, 2009; Toulmin, 1969). When measuring LD, which is the focus of this paper, researchers and clinicians are interested in drawing inferences about the LD of a sample based on the numerical values they estimate. If the value is high, clinicians want to be able to draw the conclusion that the LD of the sample is high. However, such a conclusion would be valid if and only if there is evidence to support that the numerical estimate reflects primarily LD and little of anything else.

The current study contributes to our understanding of how to measure LD because it provides the evidence necessary to justify this reasoning step from the numerical estimate to an inference about the LD of a sample. Specifically, the results from the CFA suggest that MTLD and MATTR reflect LD and little of anything else. On the other hand, the findings of this study suggest that when a language sample is evaluated using voc-D, scores may not be interpreted unequivocally as reflections of the language sample's LD without taking into account the sample's length and considering voc-D's propensity to generate missing values not at random. This methodological aspect of the measurement process constitutes a necessary (but not sufficient) condition to reach conclusions that are meaningful, appropriate, and useful. And its importance is even greater considering the purposes for which LD data are often collected—to understand the underlying deficits of people with communication disorders, to test theories, to evaluate our treatment approaches, or to change policies.

Given the need to quantify outcomes at the functional level (i.e., discourse), future studies should consider evaluating and developing a broader range of tools to assess lexical-semantic deficits in discourse produced by PWA. Further, it may be fruitful to consider the validity of score interpretations of other microlinguistic measures (e.g., informativeness); then, valid measures should be considered collectively to better conceptualize and quantify the microlinguistic level of discourse in PWA.

Acknowledgments

This research was partially supported by the National Institute on Aging Grant R01AG029476. Much of the work was completed when the authors were affiliated with Arizona State University in Tempe, AZ. We are especially grateful to the study participants. We also thank the volunteers in the Aging and Adult Language Lab at Arizona State University for assistance with language analyses and Franklin Chen from Carnegie Mellon University for assisting with CLAN coding.

References

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York, NY: Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bollen, K. A. (1989). *Structural equation modeling with latent variables*. New York, NY: Wiley.

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test–retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37*(2), 399–407.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

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.

Covington, M. A. (2007). *MATTR user manual* (CASPR research report 2007–05). Atheus, GA.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17,* 94–100.

Crepaldi, D., Ingignoli, C., Verga, R., Contardi, A., Semenza, C., & Luzzatti, C. (2011). On nouns, verbs, lexemes, and lemmas: Evidence from the spontaneous speech of seven aphasic patients. *Aphasiology, 25*(1), 71–92.

Doyle, P. J., McNeil, M. R., Spencer, K. A., Goda, A. J., Cottrell, K., & Lustig, A. P. (1998). The effects of concurrent picture presentations on retelling of orally presented stories by adults with aphasia. *Aphasiology, 12,* 561–574.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139–155.

Ertmer, D. J., Strong, L. M., & Sadagopan, N. (2002). Beginning to communicate after cochlear implantation: Oral language development in a young child. *Journal of Speech, Language, and Hearing Research, 46,* 328-340.

Fergadiotis, G. (2011). *Modeling lexical diversityacross language sampling and estimation techniques* (Doctoral dissertation). Available from Proquest Dissertations and Theses. (UMI No. 3486935)

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430.

Goodglass, H., Kaplan, E., & & Barresi, B. (2001). *The Boston Diagnostic Aphasia Examination; The assessment of aphasia and related disorders* (3rd ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Goodglass, H., & Wingfield, A. (1997). *Anomia: Neuroanatomical and cognitive correlates*. San Diego, CA: Academic Press.

Gordon, J. K. (2008). Measuring the lexical semantics of picture description in aphasia. *Aphasiology, 22*(7–8), 839–852. doi:10.1080/02687030701820063

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science: A Multidisciplinary Journal, 27*(1), 1-40. doi:10.1016/S0364-0213(02)001118

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Grimes, N. (2005). *Walt Disney's Cinderella*. New York, NY: Random House.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique* [Problems and methods of statistical linguistics]. Dordrecht, The Netherlands: D. Reidel.

Heap, H. S. (1978). *Information retrieval—Computational and theoretical aspects*. New York, NY: Academic Press.

Herdan, G. (1960). *Quantatative linguistics*. London, England: Butterworth.

Holmes, D. I., & Singh, S. (1996). A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing, 11,* 133–140.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Kapantzoglou, M. (2012). *Latent language ability groups in bilingual children across three methods of assessment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (3522169)

Kertesz, A. (2007). *Western Aphasia Battery—Revised*. New York, NY: Grune and Stratton.

Kiparsky, P. (1982). From cyclic phonology to lexical phonology. In H. G. van der Hulst & N. Smith (Eds.), *The structure of phonological representations (Part 1)*. Dordrecht, The Netherlands: Foris Publications.

Klee, T., Gavin, W. J., & Stokes, S. F. (2007). Utterance length and lexical diversity in American– and British–English speaking children: What is the evidence for a clinical marker of SLI? In R. Paul (Ed.), *Language disorders from a developmental perspective: Essays in honor of Robin S. Chapman* (pp. 103-140). Mahwah, NJ: Erlbaum.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Lind, M., Kristoffersen, K. E., Moen, I., & Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics, 23*(12), 872-886. doi:10.3109/02699200903040051

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed., pp. 131–175). New York, NY: John Wiley.

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes [The relationship between lexical diversity and the length of a sample]. *Zeitschrift für Literaturwissenschaft und Linguistik, 8,* 73–79.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs* (3rd ed.). Mahwah, NJ: Erlbaum.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25* (11), 1286–1307. doi:10.1080/02687038.2011.589893

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. H. (2010). Automated analysis of the Cinderella story. *Aphasiology. 24*(6–8), 856–868. doi:10.1080/02687030903452632

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

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

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* (Doctoral dissertation). Available from Proquest Dissertations and Theses. (UMI No. 3199485)

McCarthy, P. M., & Jarvis, S. (2007). Voc-D: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488. doi:10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.

McCarthy, P. M., Watanabe S., & Lamkin, T. A. (2012). The gramulator: A tool to identify differential linguistic features of correlative text types. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 312–333). Hershey, PA: IGI Global.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–337.

McNeil, M. R., Sung, J. E., Yang, D., Pratt, S. R., Fossett, T. R. D., Doyle, P. J., & Pavelko, S. (2007). Comparing connected language elicitation procedures in persons with aphasia: Concurrent validation of the story retell procedure. *Aphasiology, 21*(6–8), 775–790. doi:10.1080/02687030701189980

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment? *Language Learning, 59*(1), 249–267.

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, Language, and Hearing Research*, *45*(5), 927–937.

Prins, R., & Bastiaanse, R. (2004). Analyzing the spontaneous speech of aphasic speakers. *Aphasiology, 18*(12), 1075–1091. doi:10.1080/02687030444000534

Prins, R. S., Snow, C. E., & Wagenaar, E. (1978). Recovery from aphasia: Spontaneous speech versus language comprehension. *Brain and Language, 6*(2), 192–211. doi:10.1016/0093-934X(78)90058-5

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Richards, B. J., & Malvern, D. D. (1997). *Quantifying lexical diversity in the study of language development*. Reading, UK: The University of Reading.

Rider, J. D., Wright, H. H., Marshall, R. C., & Page, J. L. (2008). Using semantic feature analysis to improve contextual discourse in adults with aphasia. *American Journal of Speech-Language Pathology, 17*(2), 161–172. doi:10.1044/1058-0360(2008/016)

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. van Eye & C. C. Clogg (Eds.), *Latent variable analysis in developmental research* (pp. 285–305). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66,* 507–514.

Steiger, J. H. (1990), Structural model evaluation and modification. *Multivariate Behavioral Research, 25,* 173–180.

Steiger, J. H. (2007), Understanding the limitations of global fit assessment in structural equation modeling, *Personality and Individual Differences*, *42*(5), 893–898.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. New York, NY: Harper Collins.

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

Thordardottir, E. T., & Namazi, M. (2007). Specific language impairment in French-speaking children: Beyond grammatical morphology. *Journal of Speech, Language, and Hearing Research*, *50*, 698-714.

Toulmin, S. (1969). *The uses of argument*. Cambridge, England: Cambridge University Press.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323–352.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83. doi:10.1191/026553200676636328

Wachal, R. S., & Spreen, O. (1973). Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech, 16*(2), 169–181.

Wright, H. H., Silverman, S. W., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology, 17*(5), 443–452. doi:10.1080/02687030344000166

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