JSLHR

Research Article

Measuring Lexical Diversity for Discourse Analysis in Aphasia: Moving-Average Type–Token Ratio and Word Information Measure

Kevin T. Cunningham[a] and Katarina L. Haley[a]

Purpose: The purpose of this study was to compare the utility of two automated indices of lexical diversity, the Moving-Average Type–Token Ratio (MATTR) and the Word Information Measure (WIM), in predicting aphasia diagnosis and responding to differences in severity and aphasia subtype.
Method: Transcripts of a single discourse task were analyzed for 478 speakers, 225 of whom had aphasia per an aphasia battery. We calculated the MATTR and the WIM for each participant. We compared the group means among speakers with aphasia, neurotypical controls, and left-hemisphere stroke survivors with mild aphasia not detected by an aphasia battery. We examined whether each measure distinguished levels of aphasia severity and subtypes of aphasia. We used each measure to classify aphasia versus neurotypical control and compared the areas under the curve.
Results: The WIM and the MATTR differentiated among people with aphasia, neurotypical controls, and people with mild aphasia. Both measures demonstrated moderately high predictive accuracy in classifying aphasia. The WIM demonstrated greater sensitivity to aphasia severity and subtype compared to the MATTR.
Conclusions: The WIM and the MATTR are promising measures that quantify lexical diversity in different and complementary ways. The WIM may be more useful for quantifying the effect of treatment or disease progression, whereas the MATTR may be more useful for discriminating discourse produced by people with very mild aphasia from discourse produced by neurotypical controls. Further validation is required.

Participation in discourse is integral to human flourishing. This dynamic domain of interpersonal communication is involved in important moments such as sharing a memory, telling a story, and giving advice. Changes in discourse often occur after left-hemisphere stroke and neurodegenerative disease and pose significant sources of activity restriction (Kagan et al., 2008; Worrall et al., 2011). Performance in discourse can be a sensitive diagnostic marker for individuals whose aphasia is so mild that it is not captured by formal language testing (Fromm et al., 2017). Discourse deficits are also prominent in neurodegenerative disease (Ash et al., 2006; Fraser et al., 2016; Wilson et al., 2010) and show promise for predicting the onset of dementia (Mueller et al., 2016). From a qualitative perspective, clinicians value information provided by discourse analysis as they understandably associate this domain most closely with real-life social and communication participation (Davidson et al., 2008; Laliberté et al., 2016; Maddy et al., 2015).

Lexical diversity is one discourse feature of particular interest because it is essential to communication effectiveness. The construct includes both the range of speakers' vocabularies and the relative occurrences of word types they produce. Speakers with high lexical diversity use different types of words, while speakers with lower lexical diversity use few or repetitive word types. While many other discourse measures require labor-intensive manual coding (Pritchard et al., 2018), lexical diversity can be calculated automatically using a variety of algorithms developed in the field of computational linguistics. Thus, this amenability to automation makes it a promising candidate to improve the implementation of discourse analysis, which has been very limited due to time constraints and lack of rater knowledge (Bryant et al., 2017; Simmons-Mackie et al., 2005; Verna et al., 2009).

[a]Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina, Chapel Hill

Correspondence to Kevin T. Cunningham: kevin_cunningham@med.unc.edu

Editor-in-Chief: Sean M. Redmond
Editor: Christos Salis

Received June 8, 2019
Revision received September 19, 2019
Accepted November 26, 2019
https://doi.org/10.1044/2019_JSLHR-19-00226

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

The ease of calculation also promises to increase sample sizes for studies involving discourse analysis, which have been generally small when studying aphasia (Bryant et al., 2016).

Evolving Methods of Calculating Lexical Diversity

One early measure of lexical diversity, the type–token ratio (TTR; Templin, 1957), is obtained simply by dividing the number of different words (word types) by the total number of words (word tokens). While very influential, the TTR was found to be mediated by the length of a sample (Heaps, 1978). As the sample length increases, the TTR decreases even when there is no change in lexical diversity. In response to this undesirable effect, the field of quantitative linguistics has developed a number of measures to capture lexical diversity with algorithms that do not vary according to sample length. They include D (McCarthy & Jarvis, 2007), the Measure of Textual Lexical Diversity (McCarthy & Jarvis, 2010), the Hypergeometric Distribution (McCarthy & Jarvis, 2010), and the Moving-Average Type–Token Ratio (MATTR; Covington & McFall, 2010). As demonstrated by Fergadiotis et al. (2013), D requires speech samples of more than 50 words. Given that many individuals with nonfluent aphasia say fewer words than this threshold for a narrative task, it is not an optimal measure for the population with aphasia. The authors concluded that the Hypergeometric Distribution and the MATTR appeared particularly suitable for measuring lexical diversity in the population with aphasia.

Though the MATTR was originally designed as part of a project to analyze discourse in mental health disorders (Covington et al., 2005), it has been widely applied to individuals with neurogenic communication disorders (Elbourn et al., 2019; Fraser et al., 2014; Fromm et al., 2017; Masrani et al., 2017; Mueller et al., 2016). Like the TTR, this algorithm calculates the ratio of unique words in a sample (types) to the number of words produced (tokens). However, these calculations are based on a moving analysis window that is customized to the number of words one wishes to consider at a time. The length-adjusted analysis window moves sequentially through the text, progressing one word at a time until a token ratio has been completed for all sequential windows that can be obtained in the sample. The mean token ratio from all these windows is the MATTR metric. The fixed window size makes the MATTR a length-invariant measure, meaning that it should not vary depending on the number of words in a sample. The length of the MATTR analysis window does influence the metric, just like the length of the speech sample influences the TTR. Covington and McFall (2010) advised that the decision about which sampling window to use depends on the construct of interest. If one's interest is calculating a speaker's overall vocabulary in a sample that permits, then a very large window (10,000+) should be set. A short window (e.g., 10) is more suitable if one's goal is to detect repetition of words in very close proximity to each other. The authors note that standard window sizes should be set to ensure reproducibility, since results of different windows cannot be compared between studies. The minimum window is five words (Covington & McFall, 2010).

Fergadiotis et al. (2013) applied the MATTR using a window size of 17 words to discourse produced by people with aphasia. They demonstrated that the MATTR-17 differs between neurotypical controls (TYPICAL) and people with aphasia and is mediated by the discourse task. Fromm et al. (2017) calculated the MATTR using a window size of 20 words for individuals with anomic aphasia, left-hemisphere stroke survivors without aphasia per the Western Aphasia Battery (Kertesz, 2006), and TYPICAL. They found that the MATTR-20 was different for each of the three groups, indicating that the algorithm was sensitive to linguistic impairment not captured by the Western Aphasia Battery. The MATTR has also been demonstrated to improve the prediction of the presence and subtype of primary progressive aphasia. Fraser et al. (2014) found that the MATTR at window sizes of 10, 20, 30, 40, and 50 helped differentiate between nonfluent agrammatic primary progressive aphasia and TYPICAL and, to a lesser extent, between semantic-variant primary progressive aphasia and nonfluent agrammatic primary progressive aphasia. It was not helpful in improving the prediction of semantic-variant primary progressive aphasia from TYPICAL. The MATTR has also been considered as a candidate treatment outcome measure. Bunker et al. (2018) used the MATTR with a varying window length depending on the particular participant to evaluate outcomes of an integrated treatment for aphasia and apraxia of speech. Results showed that it responded differently from other outcomes such as novel correct information units and morphosyntactic complexity.

Limitations to the MATTR

As illustrated by the varied window sizes in previous applications, one challenge for applying the MATTR to discourse in aphasia is the range of language fluency in this population (Kertesz, 2006). Varying fluency is a multidimensional characteristic of aphasia. Verbal productivity, the number of words produced by a speaker, is one aspect of fluency and strongly predicts a rater's judgement about whether a speaker is fluent or nonfluent (Park et al., 2011). People with nonfluent aphasia often produce very few words in a discourse task, and people with fluent aphasia typically produce many more words. To avoid excluding samples with few tokens when using the MATTR, investigators typically select an analysis window equal to the smallest number of words produced in the participant sample (Fergadiotis et al., 2015, 2013; Fromm et al., 2017). While this decision maximizes data inclusion, results are incommensurable among studies that use different window lengths. In considering the MATTR as an outcome measure for intervention, Bunker et al. (2018) noted that it was difficult to judge the relative severity of deficits and treatment response because the previous literature reported norms for different windows. One goal for discourse analysis is to establish core outcomes for communication disorders, which involves the selection of key measures that can be used across studies to

improve reproducibility and meta-analyses (Dietz & Boyle, 2018; Wallace et al., 2016). To date, the variable analysis window of the MATTR makes it difficult to compare studies, rendering the measure unsuitable for inclusion in a core outcome set. While standardizing the analysis window solves this difficulty, increasing the analysis window beyond five to 20 words would seriously limit application to nonfluent aphasia.

Finally, since changing the sampling window may change the construct of the measure (Covington & McFall, 2010), use of the algorithm in the population with aphasia poses particular challenges due to differences in output productivity. For some speakers, an MATTR with a sampling window of five may reflect local repetition due to self-corrections or self-cueing, whereas for others, it may capture the size of the available lexicon due to sparse output approximating five words. Unlike the TTR, the MATTR should not vary for a particular speaker depending on the length of the sample, and it achieves this length invariance by allowing the user to specify an ad hoc sampling window. However, this robustness to sample length comes at a cost of decreased reproducibility and possibly changing measurement constructs in a population where verbal productivity varies.

Reconsideration of Length Variance: Shannon Entropy

Given these limitations, additional length-variant measures may be needed to quantify discourse deficits in aphasia, where verbal productivity can crucially impact discourse success. In the current study, we consider use of Shannon entropy (Shannon, 1948) to quantify lexical diversity. Instead of expressing lexical diversity merely as a collection of word units, entropy expresses diversity by also taking into consideration the serial relationship among words. The probability that a speaker would produce any given word depends on what has been said before, both locally and more distant. Discourse, viewed as a simple string of words, has a degree of predictability, and entropy conveys the degree to which the temporally unfolding strings of words are novel or unexpected. It is a foundational algorithm in information theory. Information in this sense refers to the degree of unexpectedness or novelty among a string of symbols. In our case, the symbols are the words, and the string is the speech sample. Consider the following example from the "Cinderella" task in which a speaker retells a children's story:

1. I do remember that Cinderella now goes to live with her evil stepmother and two stepsisters.
2. I don't remember much but I remember they are evil.With the addition of Phrase 2 to Phrase 1, little information is added to the sample. Next, the speaker says:
3. There is one named Esmeralda. She gives chores to torment Cinderella.

The addition of unexpected words in Phrase 3 is not predicted by the preceding phrases, and the information magnitude therefore increases substantially. Entropy expresses this information in the number of bits—units that denote the memory required to compress this information optimally. Mathematically, Shannon entropy can be expressed as

$$H(X) = -\sum_{i=1}^{n} P(X_i)\log_2 P(X_i) \quad (1)$$

where P is the probability of symbol X occurring. For quantifying lexical level information, each token word is considered a symbol. We call this lexical-level entropy the *Word Information Measure* (WIM). The WIM measures the novelty among individual words in a discourse sample. In the example above, the WIM of Phrase 1 is 2.77 bits; with the inclusion of Phrase 2 to the sample, it increases to 2.97 bits; and with the addition of Phrase 3, it increases to 3.33 bits.

Unlike the MATTR, the WIM is length variant: As more words are communicated by a speaker, typically more information is exchanged. It is expressed on a logarithmic scale ($\log_2$). We chose the WIM as a possible index of lexical diversity since, while it measures diversity among word types, it is sensitive to sample length. However, unlike TTR, which obtains an unexpected result as sample size increases, the WIM increases as more tokens are added to a sample. Exploratory work has used the WIM to measure the diversity of nonverbal and verbal communication of individuals with autism spectrum disorder (ASD). Parish-Morris et al. (2018) found that individuals with ASD demonstrate a lower entropy of both forms of communication than neurotypical peers during social interactions. Entropy has also been used by Malyutina et al. (2016) in aphasia, quantifying the variation among verb use types.

Study Purpose

The purpose of the current study was to examine the utility of the MATTR and the WIM for analyzing discourse production in aphasia. We are interested in determining whether these measures can serve as indices of discourse impairment and candidate outcome metrics. For this practical purpose, a measure should predict the presence of aphasia accurately and demonstrate sensitivity to global language variations among people with aphasia. Therefore, we asked the following:

1. Do the WIM and the MATTR differ on average among individuals with aphasia per a language battery, left-hemisphere stroke survivors without aphasia per the same battery, and TYPICAL?
2. How well do the WIM and the MATTR predict individual cases of aphasia compared to TYPICAL?
3. Do the WIM and the MATTR demonstrate expected variations in severity among people with aphasia?
4. How does the number of words produced by a speaker predict the WIM and the MATTR for people with aphasia?

Method

Participants

Discourse samples were obtained from the AphasiaBank (MacWhinney et al., 2011). Aphasia diagnosis was made using the cutoff suggested by Kertesz (2006) on the Western Aphasia Battery–Revised. Participants were included if they were a TYPICAL, a left-hemisphere stroke survivor not aphasic per the Western Aphasia Battery due to an Aphasia Quotient above 93.8 (NABW), or had Broca, Wernicke, conduction, or anomic-type aphasia (APHASIA) per the same aphasia battery. We excluded individuals with global aphasia because data had been submitted to the AphasiaBank for only four individuals with this subtype at the time of data collection (January 2019). This decision about particular aphasia subtypes was made to maximize variation of aphasia profiles for our validation study. We extracted only those transcripts from speakers who had the first narration of the Cinderella task demarcated in the Computerized Language Analysis program (MacWhinney, 2000). Two participants were excluded due to producing fewer than five words since these transcripts could not be analyzed with the MATTR. In all, the discourse samples from 478 participants were analyzed, including 225 APHASIA, 225 TYPICAL, and 28 NABW. Of APHASIA, the distribution of subtypes was as follows: 91 anomic aphasia, 63 Broca, 50 conduction, and 21 Wernicke. Brief demographic information is available in Table 1, with full language testing and demographic results for these particular participants available at the AphasiaBank.

Transcription Analysis

For each participant, we analyzed transcriptions of the Cinderella task that were available in chat format (MacWhinney, 2000) on the AphasiaBank website. Using the Computerized Language Analysis program (MacWhinney, 2000), we extracted an orthographic transcript that included no chat codes. We excluded unintelligible words, but all other verbal productions were included, such as whole-word repetitions, filler words, and so forth. The Cinderella task involves retelling a children's story (MacWhinney et al., 2010). We restricted analysis to this particular task due to variability in available samples on the AphasiaBank and previous findings of task effects on MATTR scores (Stark, 2019).

Word count and the WIM were calculated using the diversity function in the qdap package in R (Rinker, 2013) and a custom script to apply the analysis to all the transcripts in a folder directory. The MATTR was calculated using the mattr function in the R package koRpus (Michalke et al., 2018). We demonstrate reliability of this package for calculating the MATTR (see Appendix A) and report the performance norms for the neurotypical speakers at varying analysis windows for the MATTR (see Appendix B). The analysis window was set at the smallest number of words produced by our participants (five), which is also the smallest possible window of the MATTR (Covington & McFall, 2010). When discussing our particular analysis and findings, we will refer to this measure as the MATTR-5 to reflect that our findings are not necessarily generalizable to other analysis windows. When discussing the theoretical or practical applications of the general measure, we will continue to use the term *MATTR*. All data including the measures of lexical diversity and word count are available on the AphasiaBank.

Statistical Analysis Plan

RQ1: Do the WIM and the MATTR-5 Differ Among APHASIA, TYPICAL, and NABW Groups?

All diagnostic groups were dummy-coded across analyses. For both the WIM and the MATTR-5, an analysis of variance (ANOVA) was planned using post hoc pairwise two-tailed *t* tests with a Holm correction for multiple comparisons and an alpha of .05 determined prior to analysis. The omega-squared (ω^2) effect size was calculated for models with significant differences.

RQ2: Do the WIM and the MATTR-5 Predict APHASIA and TYPICAL Group Membership Correctly?

For each discourse measure, we generated a receiver operating characteristic (ROC; the plot of the true-positive rate on the false-positive rate) for classifying the APHASIA from TYPICAL participants. We then calculated the area under the curve (AUC; the probability that a measure would rank a randomly selected APHASIA higher than a randomly selected TYPICAL). Based on inspection of the results, we generated a combined model by performing a multiple linear binomial logistic regression of the MATTR-5 and the WIM as predictors of diagnostic group (Kurt et al., 2008) and then using this model as the classification criterion for the ROC. For the combined model, each predictor

Table 1. Sex ratio, mean education, age, and aphasia testing of participants by diagnostic group.

Group	Sex (female:male)	Education in years, *M* (*SD*)	Age in years, *M* (*SD*)	WAB-AQ, *M* (*SD*)
TYPICAL	125:100	15.2 (2.3)	55.0 (23.4)	
NABW	19:9	16.1 (2.8)	61.2 (13.5)	96.30 (1.70)
APHASIA	92:133	15.0 (3.0)	62.0 (12.0)	69.7 (17.6)

Note. WAB-AQ = Western Aphasia Battery–Aphasia Quotient; TYPICAL = neurotypical speakers; NABW = speakers who were not aphasic per the Western Aphasia Battery; APHASIA = speakers with aphasia.

only included the unique variance explained by the WIM or the MATTR, controlling for variance shared between the predictors. Analyses were completed with the R package pROC (Robin et al., 2011). Next, we performed Delong's test (DeLong et al., 1988) comparing the AUC for the MATTR-5 and the AUC for the WIM. We descriptively report the AUC for the combined model due to limitations in the use of Delong's test for such nested regression models (Seshan et al., 2013). Optimal sensitivity and specificity thresholds for each method were determined using the method of Youden (1950). We chose to compare only APHASIA and TYPICAL in this analysis because, based on inspection of the narrative samples, it was not obvious whether all NABW participants had quantifiable language impairment or whether symptoms in some cases could be strictly subjective in nature. We were interested in predicting clinically salient aphasia for this initial validation study.

RQ3: Do the WIM and the MATTR-5 Demonstrate Expected Variation by Aphasia Severity and Subtype?

An ANOVA was performed for each measure with groups of APHASIA stratified by the severity cutoffs on the Aphasia Quotient suggested by Kertesz (2006): mild (75–93.8), moderate (50–75), severe (25–50), and profound (< 25). The omega-squared (ω^2) effect size was calculated for models with significant differences. Post hoc pairwise two-tailed *t* tests were performed using a Holm correction for multiple comparisons. The same analysis procedure was also performed for APHASIA stratified by aphasia subtype per the Western Aphasia Battery–Revised (Kertesz, 2006).

RQ4: How Does Word Count Predict the MATTR-5 and the WIM?

Two simple linear regressions were performed, each with word count as a predictor of the MATTR-5 or the WIM. Diagnostic group (APHASIA, TYPICAL, or NABW) was included as an interaction term. An ANOVA was calculated for each model. Pairwise comparisons of the estimated marginal means were performed for models with a significant interaction term.

Results

RQ1: Do the WIM and the MATTR-5 Differ Among APHASIA, TYPICAL, and NABW Groups?

The mean MATTR-5 for APHASIA was 0.90 (*SD* = 0.06), as demonstrated in Figure 1. A mean of 0.95 (*SD* = 0.03) was found for NABW, and a mean of 0.97 (*SD* = 0.01) was found for TYPICAL. The ANOVA for the MATTR-5 indicated a statistically significant effect of diagnostic group, $F(2, 475) = 164.77$, $p < .001$, $\omega^2 = .41$. The effect size may be considered medium to large (Kirk, 1996). Post hoc *t* tests corrected for multiple comparison indicated a significant group difference between each pairwise comparison ($p < .01$). As shown in Figure 2, the mean WIM was 3.68 (*SD* = 0.66) for APHASIA, 4.18 (*SD* = 0.27) for NABW, and 4.5 (*SD* = 0.34) for TYPICAL. The ANOVA for the WIM also indicated a significant effect, $F(2, 475) = 143.72$, $p < .001$, $\omega^2 = .37$. The effect size is considered medium to large. Post hoc *t* tests corrected for multiple comparison indicated a significant group difference between each pairwise comparison ($p < .01$).

Figure 1. Boxplot of the Moving-Average Type–Token Ratio (MATTR-5) for people with aphasia, people without aphasia per a language battery (NABW), and neurotypical controls (TYPICAL). Whiskers represent the range of extreme values within 1.5 times the interquartile range. NABW = people who are not aphasic per the Western Aphasia Battery.

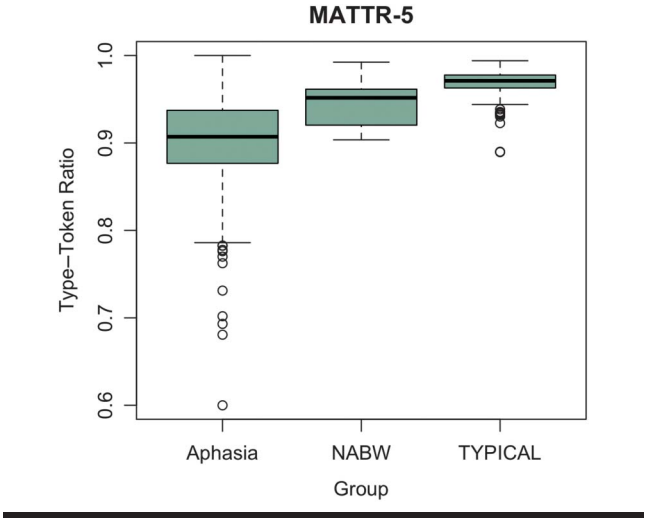
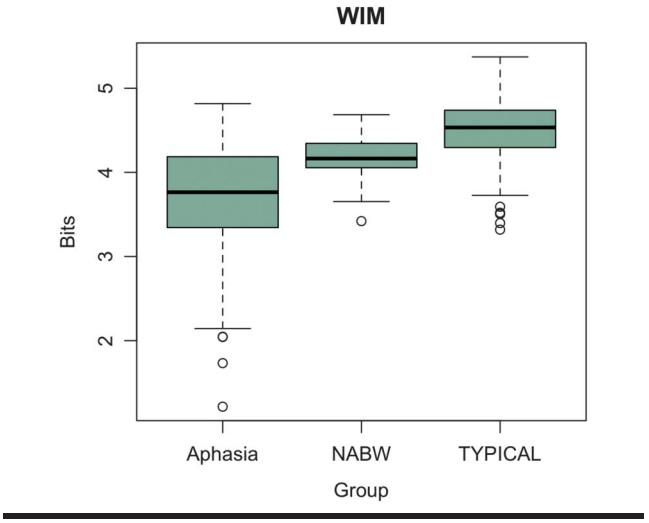


Figure 2. Boxplot of the Word Information Measure (WIM) for people with aphasia, people without aphasia per a language battery (NABW), and neurotypical controls (TYPICAL). Whiskers represent the range of extreme values within 1.5 times the interquartile range. NABW = people who are not aphasic per the Western Aphasia Battery.



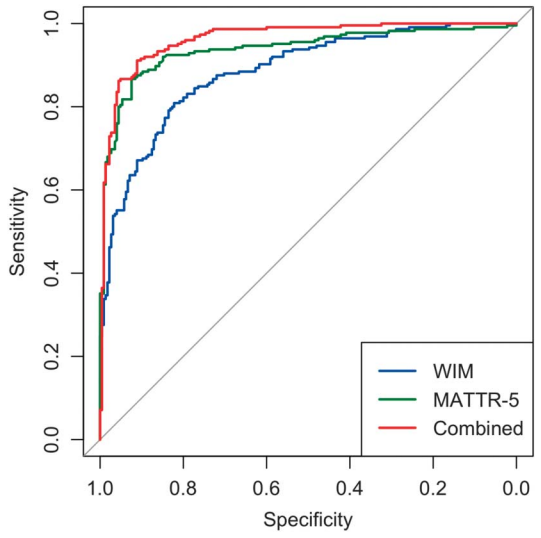
RQ2: Do the WIM and the MATTR-5 Predict APHASIA and TYPICAL Group Membership Correctly?

The ROC curves are shown in Figure 3. The areas under the curve were 0.88 for the WIM, 0.94 for the MATTR-5, and 0.96 for the combined model. Delong's test indicated that the AUC for the WIM and the AUC for the MATTR-5 were statistically different, $U = 3.08$, $p = .002$. The optimal cutoff for predicting aphasia versus control per Youden's (1950) method was as follows: WIM = 4.23 bits and MATTR-5 = 0.95 TTR. For the MATTR-5, the specificity was 92.4% and the sensitivity was 86.7%. For the WIM, the specificity was 82.2% and the sensitivity was 80.1%. For the combined model, both the sensitivity and specificity were 91.1%.

RQ3: Do the WIM and the MATTR-5 Demonstrate Expected Variation by Aphasia Severity and Aphasia Subtype?

Severity. The means of the MATTR-5 for each severity classification were as follows: mild = 0.91 ($SD = 0.05$), moderate = 0.89 ($SD = 0.06$), severe = 0.89 ($SD = 0.06$), and profound = 0.84 ($SD = 0.15$). The ANOVA of the MATTR-5 across severity categories indicated a significant difference in the model, $F(3, 221) = 3.43$, $p = .018$, $\omega^2 = .03$. The effect size may be considered small. However, no significant group differences survived correction for multiple post hoc comparisons ($p = .059$ for severe and mild aphasia; $p > .20$ for each other comparison). The means of the WIM for each severity classification were as follows: mild = 3.90 ($SD = 0.52$), moderate = 3.56 ($SD = 0.64$), severe = 3.49 ($SD = 0.76$), and profound = 2.40 ($SD = 1.04$). ANOVA of the WIM across these classifications indicated a significant difference, $F(3, 221) = 12.36$, $p < .001$, $\omega^2 = .13$.

Figure 3. Receiver operating characteristic curve for the Word Information Measure (WIM), Moving-Average Type–Token Ratio (MATTR), and the combined model including both measures.



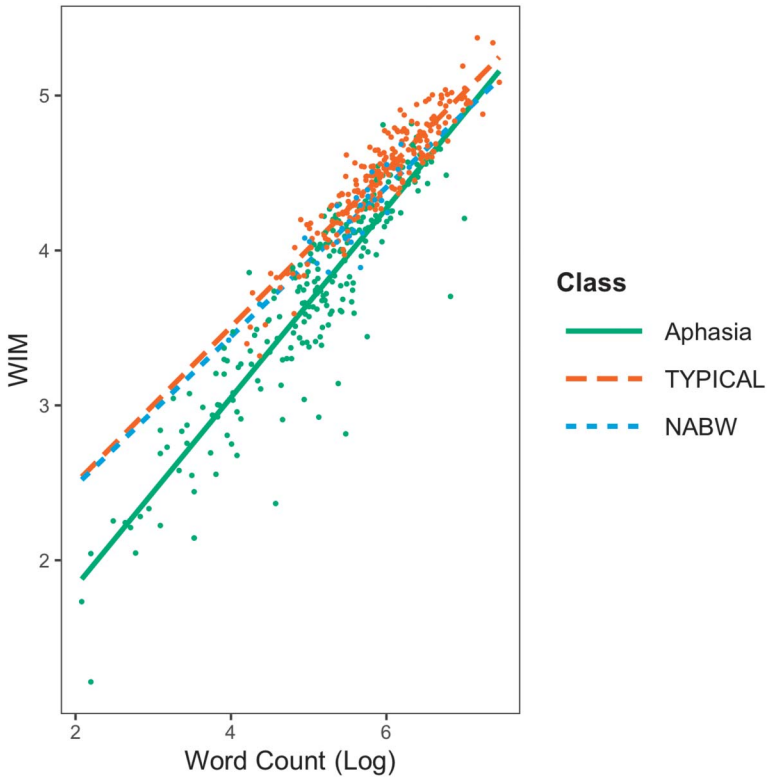
The effect size may be considered small. Post hoc testing showed significant differences between all pairs of severity levels ($p < .01$), except between moderate aphasia and severe aphasia ($p = .62$).

Subtype. The means of the MATTR-5 for each aphasia subtype were as follows: Broca = 0.88 ($SD = 0.07$), Wernicke = 0.90 ($SD = 0.05$), conduction = 0.90 ($SD = 0.06$), and anomic = 0.91 ($SD = 0.05$). The ANOVA of the MATTR-5 across subtypes indicated no statistically significant differences in the model, $F(3, 221) = 1.99$, $p = .127$. The means of the WIM for each aphasia subtype were as follows: Broca = 3.2 ($SD = 0.68$), Wernicke = 3.93 ($SD = 0.50$), conduction = 3.89 ($SD = 0.56$), and anomic = 3.85 ($SD = 0.55$). The ANOVA of the WIM across Western Aphasia Battery (Kertesz, 2006) subtypes (anomic, Broca, conduction, and Wernicke) indicated a significant difference in the model, $F(3, 221) = 19.43$, $p < .001$, $\omega^2 = .20$. The effect size may be considered medium sized. Post hoc t tests corrected for multiple comparisons indicated a difference between Broca aphasia and each other subtype ($p < .001$) but no statistically significant differences among the fluent subtypes ($p = 1$).

RQ4: How Does Word Count Predict the MATTR-5 and the WIM?

As expected with the logarithmic scale of Shannon entropy and linear scale of word count, the scatter plot of the initial regression model for the WIM on word count indicated a quadratic relationship between the two variables. Inspection of the fitted versus residuals plot revealed heteroskedasticity of residuals, thus violating an assumption of linear regression. A log transform of the predictor variable improved the model fit and resulted in homoskedasticity. When diagnosis was added as an interaction term, the adjusted R^2 was .89. Significant findings were found for the main effect of word count, $F(1, 472) = 3834.21$, $p < .001$. The interaction between diagnosis group and word count was also significant, $F(1, 472) = 8.58$, $p = .0002$. Post hoc testing of the estimated marginal means for the WIM model indicated a significant group difference between APHASIA and TYPICAL, $t(-12.6)$, $p < .001$, and APHASIA and NABW, $t(-4.66)$, $p < .001$, but not TYPICAL and NABW, $t(2.21)$, $p = .07$. As visualized in Figure 4, the slope for APHASIA was greater than those for the other two groups, with increasing verbal productivity predicting a greater increase in the WIM. For the linear regression of the MATTR-5 on word count, no data transformation was indicated by inspection of the Q–Q plot or residuals versus fitted plot. However, for purposes of comparison between the measures, we also performed a logarithmic transformation of the word count. An adjusted R^2 of .41 was for the regression of the MATTR-5 on word count with diagnosis as an interaction term. The main effect of word count was statistically significant, $F(1, 472) = 6.74$, $p = .01$. However, the interaction between diagnostic group and word count was not significant, $F(2, 472) = 1.39$, $p = .25$. This model is visualized in Figure 5.

Figure 4. Regression of the Word Information Measure (WIM) on number of words produced by each speaker with diagnosis of aphasia, neurotypical control (TYPICAL), and mild aphasia not measured by a language battery (NABW) included as an interaction term. NABW = people who are not aphasic per the Western Aphasia Battery.



Discussion

Lexical Diversity Is Sensitive to the Presence, Severity, and Subtype of Aphasia

Both the WIM and the MATTR-5 differentiated among APHASIA, TYPICAL, and NABW. Effect sizes were medium to large, and group differences were statistically significant after post hoc testing. These results are consistent with the findings of Fromm et al. (2017) that the MATTR differed among people with anomic aphasia, TYPICAL, and not aphasic by the Western Aphasia Battery–Revised (Kertesz, 2006). Our findings, using a larger sample with more diverse profiles of aphasia, confirm that the MATTR-5 can detect linguistic impairment in left-hemisphere stroke survivors with an Aphasia Quotient past the cutoff on the Western Aphasia Battery–Revised. The results also support our hypothesis that the WIM differentiates among these groups.

Our results indicate that lexical diversity is highly predictive of individual cases of aphasia. Both measures individually had moderate high predictive accuracy despite the fact that our analysis involved a single discourse task. When each of the measures was compared individually, the MATTR-5 had the better accuracy for identifying the presence of aphasia. These findings are consistent with previous work that found the MATTR was helpful in predicting cases of dementia and mild cognitive impairment (Fraser et al., 2016; Masrani et al., 2017; Mueller et al., 2016).

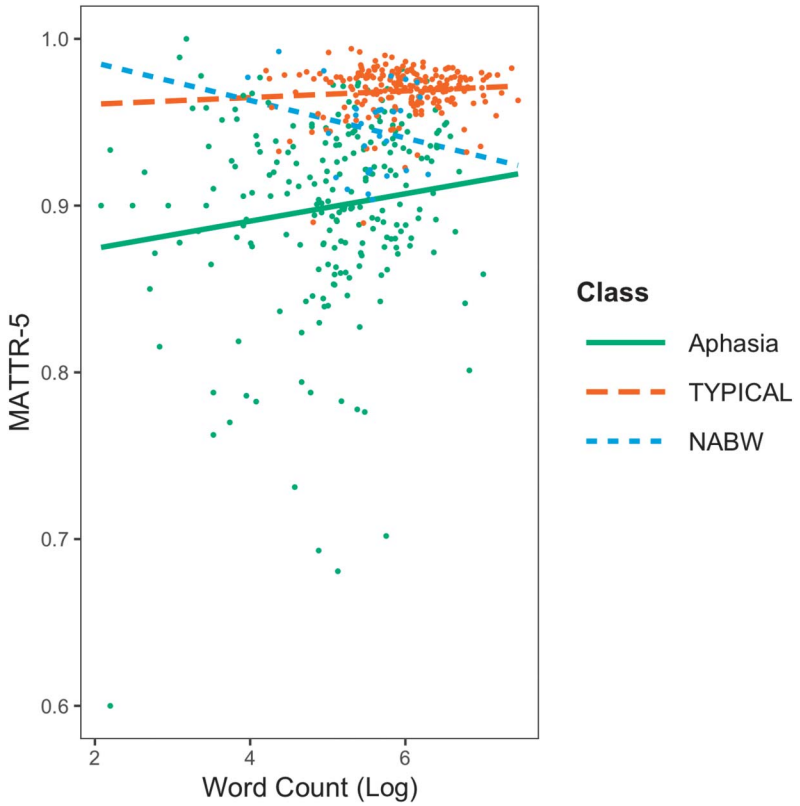
Our results indicate that the length-variant measure of the WIM demonstrated better sensitivity than the MATTR-5 for detecting severity variation among people with aphasia. While much work has been performed in quantitative linguistics to develop measures of lexical diversity that are robust to the length of a sample, our results indicate that the WIM may be particularly useful because of its length-variant characteristics. Unlike other applications such as mental health disorders or some pediatric language disorders, verbal productivity varies greatly among people with aphasia and impacts discourse performance dramatically. As a result, length-invariant measures such as the MATTR may sometimes produce unexpected results. Consider the following example, which is the entire transcript of the Cinderella passage for a speaker with an Aphasia Quotient of 28 and a subtype of Broca's aphasia:

> a birch much I try pretty I don't know

When the analysis window of 5 for the MATTR-5 is applied,

> [a birch much I try] = (5 types/5 tokens) = 1
> [birch much I try pretty] = (5 types/5 tokens) = 1

Figure 5. Regression of the Moving-Average Type–Token Ratio (MATTR-5) on number of words produced by each speaker with a diagnosis of aphasia, neurotypical control (TYPICAL), and mild aphasia not measured by a language battery (NABW) included as an interaction term. NABW = people who are not aphasic per the Western Aphasia Battery.



[much I try pretty I] = (5 types/ 5 tokens) = 0.8
[I try pretty I don't] = (4 types/5 tokens) = 0.8
[try pretty I don't know] = (5 types/5 tokens) = 1.0

An MATTR-5 of 0.92 is obtained as the mean of these values. By consulting Figure 1, we learn that this score is higher than the median for the APHASIA group, ranking this speaker above the majority of people with aphasia. Because the MATTR is length invariant, it is sometimes insensitive to the speaker's sparse verbal output with very few types or tokens produced. Contrastingly, this same transcript obtains a WIM score of 2.04 bits, making this speaker with an Aphasia Quotient of 28 an outlier below the first quartile. Because it accounts for verbal productivity and the relative preponderance of types to tokens, we suggest that the WIM may better reflect the overall severity of discourse production impairment in this population with varying fluency.

To further explore this relationship of verbal productivity and the WIM and MATTR-5, we performed an analysis with each method regressed on word count with diagnostic group as an interaction term. For the WIM, the expected relationship was found, with the WIM increasing incrementally as word count increased. This relationship is likely a function of both the inherent length variance of the measure and the fact that individuals with aphasia who produce more words more often produce more types to tokens. For individuals with aphasia, increases in word count predicted greater increases in the WIM than for neurotypical speakers or individuals with mild aphasia not detected by a language battery. This finding in aphasia differs from the investigation of Parish-Morris et al. (2018) where increases in word count of people with ASD were predicted to be more modest than increases in entropy (equivalent to the WIM) compared to TYPICAL. We also found a positive relationship between word count and the MATTR-5, though this index of lexical diversity is length invariant. This relationship did not differ between diagnostic groups. While an association in increasing word count and the MATTR-5 may be related to severity for people with aphasia, the reasons for this relationship among neurotypical speakers are not clear. Possibly, it may reflect differences in participant effort or task administration that result in some speakers producing both a higher word count and a greater MATTR-5. Future investigations explaining performance variation among neurotypical speakers are needed to improve diagnostic precision.

Local Lexical Diversity

Lexical diversity has traditionally been understood to refer broadly to "how many different words are used"

(Johansson, 2009, p. 61) or "the range and variety of vocabulary used" (McCarthy & Jarvis, 2007, p. 549). Our results support that each of the candidate measures captures a different level of analysis for lexical diversity. The MATTR-5 detects repetition of words in very close proximity, with a local level of analysis similar to cohesion analysis (Armstrong, 1991). Individuals with aphasia can produce repeated words in discourse for a variety of reasons, including failures of lexical retrieval, speech production difficulties, and grammatical deficits resulting in repetitive morphosyntactic structures. Our results suggest that this local lexical diversity appears to be highly suitable to distinguish neurotypical speakers from individuals with left-hemisphere focal lesions, which is consistent with many studies using the MATTR to predict mild cognitive impairment and dementia (Fraser et al., 2014; Masrani et al., 2017; Mueller et al., 2016). Given its comparatively greater sensitivity to a diagnosis of aphasia, the MATTR-5 may also be beneficial in quantifying the discourse deficits of people with very mild aphasia, for whom there is an urgent need to objectively characterize deficits in order to justify rehabilitation services (Cavanaugh & Haley, 2020).

Global Lexical Diversity

If a clinician needs to measure the range of a speaker's vocabulary, the MATTR-5 is likely not suitable to capture this more global form of lexical diversity. As in the example above, an individual who produces very few words and very few word types can still have a moderately high MATTR-5 score due to a preponderance of word types in their limited output or the presence of agrammatism, which reduces the number of normally occurring repetitive functor words. The WIM appears to be a more suitable candidate to capture global lexical diversity, or the entire range of a speaker's output in a discourse sample. It is a length-variant measure, meaning that it also measures verbal productivity in addition to the relative preponderance of types to tokens. Unlike the TTR, the length variance of the WIM is fixed in the expected direction with improvements in verbal productivity predicting improvements in the WIM. However, unlike simple word count, the relationship of the WIM and verbal productivity is not linear. Adding highly redundant words to a sample will not meaningfully increase the WIM score. Instead, a speaker needs to produce a diversity of word types to increase the WIM score.

This length variance may provide the WIM with some advantages as an outcome measurement. Our findings indicate that it was more sensitive to aphasia severity and aphasia subtype, unlike the MATTR, which did not differ between people with mild aphasia and people with severe aphasia. The WIM is also more reproducible with no potential for incommensurate results due to unreported or varying sampling windows (e.g., Bunker et al., 2018; Masrani et al., 2017), making it more suitable for inclusion in a core outcome set (Armstrong, 2018; Dietz & Boyle, 2018; Wallace et al., 2016). In addition, the WIM can be more broadly applied in aphasia since it allows meaningful measurement of lexical diversity in people with very severe aphasia who produce fewer than five words.

Limitations

A limitation of the study is that our discourse analysis was based on a single task, and lexical diversity is known to be mediated by the type of discourse task (Stark, 2019). Future studies should examine the number and type of tasks optimal for sensitivity and specificity and develop standardization procedures to minimize examiner effects (Wright & Capilouto, 2009). Additionally, the Cinderella task, though widely used, is relatively simplistic, involving a person retelling children's story after reviewing pictures representing it. More linguistically challenging and ecologically valid tasks may improve the specificity and sensitivity of the measures. Finally, though this study has a large sample size, the group of NABW was relatively small. Future studies should use larger samples for this important group and separately study asymptomatic left-hemisphere stroke survivors and symptomatic left-hemisphere stroke survivors. Finally, the Western Aphasia Battery–Aphasia Quotient is a relatively coarse measurement, which itself is strongly predicted by language fluency (Crary & Gonzalez Rothi, 1989), so it is not surprising that the length-variant measure of the WIM was more predictive of the Aphasia Quotient. Future studies should investigate these indices of lexical diversity in relation to diverse measures of impairment, including more comprehensive language batteries or other discourse measures.

Future Directions

We investigated two methods of analyzing lexical diversity in discourse, which satisfy the concerns reported by clinicians for implementation of discourse assessment (Bryant et al., 2017). They can be very quickly applied and require no training, with both measures currently available via online calculators with general user interfaces (Michalke, 2019; PlanetCalc, 2019). With these tools, a clinician could simply copy and paste a transcribed discourse simply into the web client and immediately receive a score of lexical diversity. These methods appear to be promising avenues to guide better diagnosis and treatment of communication disorders that affect this crucial domain of communication. Both measures require psychometric validation for people with neurogenic communication disorders. Temporal stability and test–retest reliability need to be demonstrated for these measures to be considered in outcome assessment. In addition, standardized methods of eliciting discourse tasks should be considered given the potential for examiner instructions to influence lexical diversity (Wright & Capilouto, 2009). To contribute to standardization of analysis windows for the MATTR, we have provided performance norms for the neurotypical speakers at suggested windows of five, 10, 15, 25, and 50 words. We cannot currently report norms for the people with aphasia since many individuals in this sample did not produce 10 or 15 words

on the Cinderella passage. Future studies should report performance norms at these suggested windows for people with aphasia by administering more discourse elicitation tasks.

Finally, though the analysis methods are automated in this current study, the generation of discourse samples still requires manual transcription. To improve implementation, speech recognition technology may be a solution to achieve point-to-point automation (Fraser et al., 2013; Jacks et al., 2019; Themistocleous et al., 2019).

Acknowledgments

Thank you to Davida Fromm, Michael Covington, Peter Halpin, Tyler Rinker, and Meik Michalke for consultation.

References

Armstrong, E. (1991). The potential of cohesion analysis in the analysis and treatment of aphasic discourse. *Clinical Linguistics & Phonetics, 5*(1), 39–51. https://doi.org/10.3109/02699209108985501

Armstrong, E. (2018). The challenges of consensus and validity in establishing core outcome sets. *Aphasiology, 32*(4), 465–468. https://doi.org/10.1080/02687038.2017.1398804

Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., & Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology, 66*(9), 1405–1413. https://doi.org/10.1212/01.wnl.0000210435.72614.38

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics, 30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Bunker, L. D., Wright, S., & Wambaugh, J. L. (2018). Language changes following combined aphasia and apraxia of speech treatment. *American Journal of Speech-Language Pathology, 27*(1S), 323–335. https://doi.org/10.1044/2018_AJSLP-16-0193

Cavanaugh, R., & Haley, K. L. (2020). Subjective communication difficulties of very mild aphasia: Survey of aphasia assessment measures implemented in clinical and research settings. *American Journal of Speech-Language Pathology, 29*(1S), 437–448. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0222

Covington, M. A., He, C., Brown, C., Naçi, L., McClain, J. T., Fjordbak, B. S., Semple, J., & Brown, J. (2005). Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research, 77*(1), 85–98. https://doi.org/10.1016/j.schres.2005.01.016

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Crary, M. A., & Gonzalez Rothi, L. J. (1989). Predicting the Western Aphasia Battery Aphasia Quotient. *Journal of Speech and Hearing Disorders, 54*(2), 163–166. https://doi.org/10.1044/jshd.5402.163

Davidson, B., Howe, T., Worrall, L., Hickson, L., & Togher, L. (2008). Social participation for older people with aphasia: The impact of communication disability on friendships. *Topics in Stroke Rehabilitation, 15*(4), 325–340. https://doi.org/10.1310/tsr1504-325

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: Have we reached the tipping point? *Aphasiology, 32*(4), 459–464. https://doi.org/10.1080/02687038.2017.1398803

Elbourn, E., Kenny, B., Power, E., Honan, C., McDonald, S., Tate, R., Holland, A., MacWhinney, B., & Togher, L. (2019). Discourse recovery after severe traumatic brain injury: Exploring the first year. *Brain Injury, 33*(2), 143–159. https://doi.org/10.1080/02699052.2018.1539246

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58*(3), 840–852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22*(2), S397–S408. https://doi.org/10.1044/1058-0360(2013/12-0083)

Fraser, K. C., Hirst, G., Graham, N. L., Meltzer, J. A., Black, S. E., & Rochon, E. (2014). Comparison of different feature sets for identification of variants in progressive aphasia. In P. Resnik, R. Resnik & M. Mitchell (Eds.), *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 17–26). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3203

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Fraser, K., Rudzicz, F., Graham, N., & Rochon, E. (2013). Automatic speech recognition in the diagnosis of primary progressive aphasia. In J. Alexandersson, P. Ljunglöf, K. F. McCoy F. Portet, B. Roark, F. Rudzicz, & M. Vacher (Eds.), *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies* (pp. 47–54). Association for Computational Linguistics.

Fromm, D., Forbes, M., Holland, A., Dalton, S. G., Richardson, J., & MacWhinney, B. (2017). Discourse characteristics in aphasia beyond the western aphasia battery cutoff. *American Journal of Speech-Language Pathology, 26*(3), 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071

Heaps, H. S. (1978). *Information retrieval, computational and theoretical aspects*. Academic Press.

Jacks, A., Haley, K. L., Bishop, G., & Harmon, T. G. (2019). Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions. *Folia Phoniatrica et Logopaedica, 71*(5–6), 282–292. https://doi.org/10.1159/000499156

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers in linguistics, 53,* 61–79.

Kagan, A., Simmons-Mackie, N., Rowland, A., Huijbregts, M., Shumway, E., McEwen, S., Threats, T., & Sharp, S. (2008). Counting what counts: A framework for capturing real-life outcomes of aphasia intervention. *Aphasiology, 22*(3), 258–280. https://doi.org/10.1080/02687030701282595

Kertesz, A. (2006). *Western Aphasia Battery–Revised*. Pro-Ed.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications, 34*(1), 366–374. https://doi.org/10.1016/j.eswa.2006.09.004

Laliberté, M. P., Alary Gauvreau, C., & Le Dorze, G. (2016). A pilot study on how speech-language pathologists include social participation in aphasia rehabilitation. *Aphasiology, 30*(10), 1117–1133. https://doi.org/10.1080/02687038.2015.1100708

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics, 26*(4), 657. https://doi.org/10.1162/coli.2000.26.4.657

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

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

Maddy, K. M., Howell, D. M., & Capilouto, G. J. (2015). Current practices regarding discourse analysis and treatment following non-aphasic brain injury: A qualitative study. *Journal of Interactional Research in Communication Disorders, 6*(2), 211–236. https://doi.org/10.1558/jircd.v7i1.25519

Malyutina, S., Richardson, J. D., & den Ouden, D. B. (2016). Verb argument structure in narrative speech: Mining aphasiabank. *Seminars in Speech and Language, 37*(1), 34–47. https://doi.org/10.1055/s-0036-1572383

Masrani, V., Murray, G., Field, T., & Carenini, G. (2017). Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In K. B. Cohen, D. Demner-Fushman, S. Ananiadou, & J. Tsujii (Eds.), *BioNLP 2017* (pp. 232–237). Association for Computational Linguistics.

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

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Michalke, M. E. (2019). *koRpus text analysis*. Retrieved from https://ripley.psycho.hhu.de/koRpus/

Michalke, M. E., Brown, E., Mirisola, A., & Brulet, A. (2018). *koRpus: An R package for text analysis* (0.11.5) [Computer software]. https://rdrr.io/cran/koRpus/

Mueller, K. D., Koscik, R. L., Turkstra, L. S., Riedeman, S. K., LaRue, A., Clark, L. R., Hermann, B., Sager, M. A., & Johnson, S. C. (2016). Connected language in late middle-aged adults at risk for Alzheimer's disease. *Journal of Disease, 54*(4), 1539–1550. https://doi.org/10.3233/JAD-160252

Parish-Morris, J., Sariyanidi, E., Zampella, C., Bartley, G. K., Ferguson, E., Pallathra, A. A., Bateman, L., Plate, S., Cola, M., Pandey, J., Brodkin, E. S., Schultz, R. T., & Tunç, B. (2018). Oral-Motor and lexical diversity during naturalistic conversations in adults with autism spectrum disorder. In K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik, & P. Resnik (Eds.), *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 147–157). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0616

Park, H., Rogalski, Y., Rodriguez, A. D., Zlatar, Z., Benjamin, M., Harnish, S., Bennett, J., Rosenbek, J. C., Crosson, B., & Reilly, J. (2011). Perceptual cues used by listeners to discriminate fluent from nonfluent narrative discourse. *Aphasiology, 25*(9), 998–1015. https://doi.org/10.1080/02687038.2011.570770

PlanetCalc. (2019). *Calculator for Shannon entropy*. Retrieved from https://planetcalc.com/2476/

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders, 53*(6), 1078–1093. https://doi.org/10.1111/1460-6984.12420

Rinker, T. W. (2013). *qdap: Quantitative discourse analysis package*. University at Buffalo/SUNY.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(77). https://doi.org/10.1186/1471-2105-12-77

Seshan, V. E., Gönen, M., & Begg, C. B. (2013). Comparing ROC curves derived from regression models. *Statistics in Medicine, 32*(9), 1483–1493. https://doi.org/10.1002/sim.5648

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simmons-Mackie, N., Threats, T. T., & Kagan, A. (2005). Outcome assessment in aphasia: A survey. *Journal of Communication Disorders, 38*(1), 1–27. https://doi.org/10.1016/j.jcomdis.2004.03.007

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology, 28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

Templin, M. C. (1957). *Certain language skills in children; their development and interrelationships*. University of Minnesota Press. https://doi.org/10.5749/j.ctttv2st

Themistocleous, C., Webster, K., Ficek, B., & Tsapkini, K. (2019). *Quantification of PPA effects on part-of-speech using computational grammars*. Platform presentation, The 49th Clinical Aphasiology Conference, May 30, 2019, Whitefish, MT.

Verna, A., Davidson, B., & Rose, T. (2009). Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology, 11*(3), 191–205. https://doi.org/10.1080/17549500902726059

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2016). Core outcomes in aphasia treatment research: An e-Delphi consensus study of international aphasia researchers. *American Journal of Speech-Language Pathology, 25*(4S), S729–S742. https://doi.org/10.1044/2016_AJSLP-15-0150

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., & Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain: A Journal of Neurology, 133*(Pt. 7), 2069–2088. https://doi.org/10.1093/brain/awq129

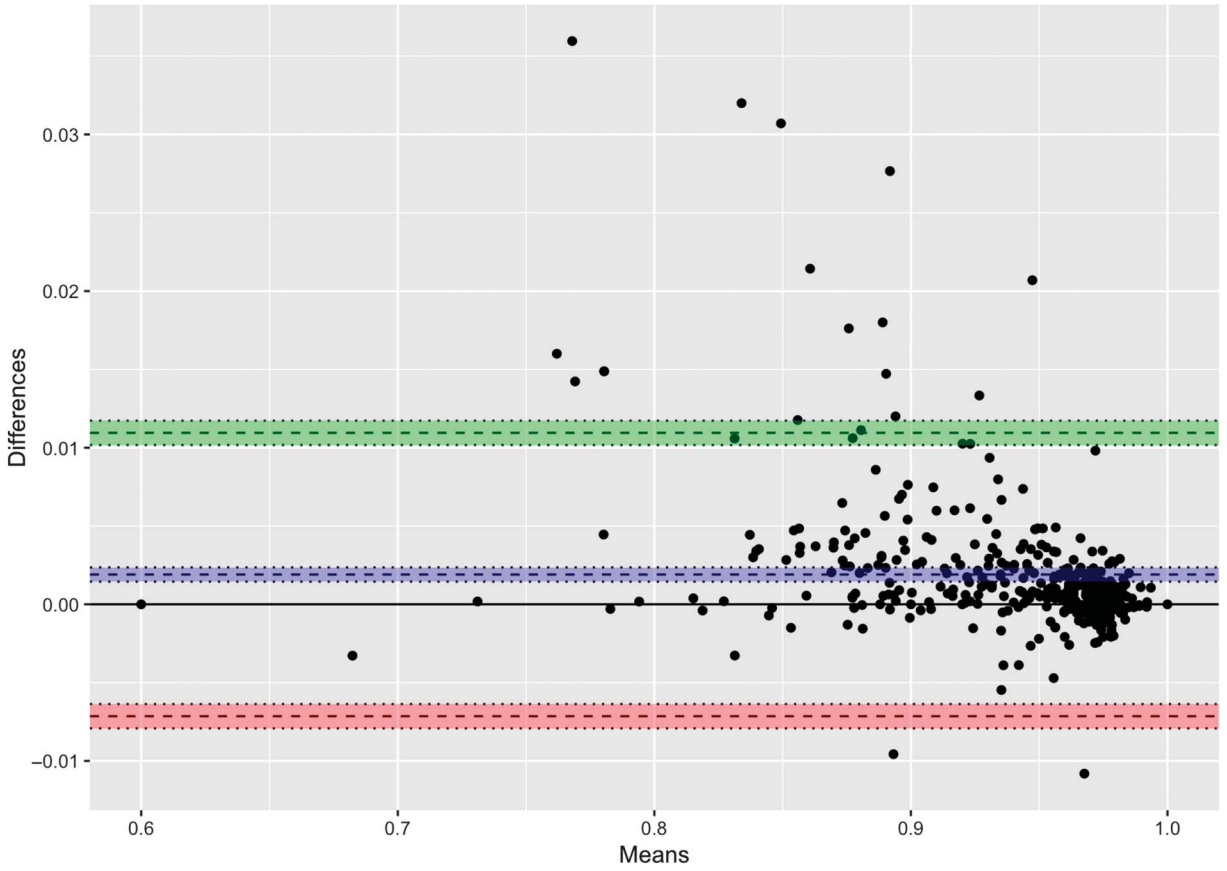
Worrall, L., Sherratt, S., Rogers, P., Howe, T., Hersh, D., Ferguson, A., & Davidson, B. (2011). What people with aphasia want: Their goals according to the ICF. *Aphasiology, 25*(3), 309–322. https://doi.org/10.1080/02687038.2010.508530

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology, 23*(10), 1295–1308. https://doi.org/10.1080/02687030902826844

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Appendix A

Bland–Altman analysis between the Moving-Average Type–Token Ratio as computed by the R package koRpus and the baseline software.



Appendix B

Neurotypical Performance on the Cinderella Task

Measure	*M*	*SD*
MATTR-5	0.969	0.014
MATTR-10	0.916	0.022
MATTR-25	0.805	0.035
MATTR-50	0.701	0.040
WIM	4.500	0.341

Note. The number after MATTR represents the selection of analysis window. MATTR = Moving-Average Type–Token Ratio; WIM = Word Information Measure.