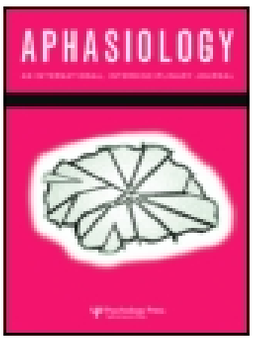

Aphasiology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/paph20

Enhancing the classification of aphasia: a statistical analysis using connected speech

Davida Fromm, Joel Greenhouse, Mitchell Pudil, Yichun Shi & Brian MacWhinney

To cite this article: Davida Fromm, Joel Greenhouse, Mitchell Pudil, Yichun Shi & Brian MacWhinney (2021): Enhancing the classification of aphasia: a statistical analysis using connected speech, Aphasiology, DOI: 10.1080/02687038.2021.1975636

To link to this article: https://doi.org/10.1080/02687038.2021.1975636

APHASIOLOGY
https://doi.org/10.1080/02687038.2021.1975636

Enhancing the classification of aphasia: a statistical analysis using connected speech

Davida Fromm[a], Joel Greenhouse[b], Mitchell Pudil[b], Yichun Shi[b] and Brian MacWhinney[a]

[a]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA; [b]Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Background: Large-shared databases and automated language analyses allow for the application of new data analysis techniques that can shed new light on the connected speech of people with aphasia (PWA).

Aims: To identify coherent clusters of PWA based on language output using unsupervised statistical algorithms and to identify features that are most strongly associated with those clusters.

Methods & Procedures: Clustering and classification methods were applied to language production data from 168 PWA. Language samples were from a standard discourse protocol tapping four genres: free speech personal narratives, picture descriptions, Cinderella storytelling, and procedural discourse.

Outcomes & Results: Seven distinct clusters of PWA were identified by the K-means algorithm. Using the random forest algorithm, a classification tree was proposed and validated, showing 91% agreement with the cluster assignments. This representative tree used only two variables to divide the data into distinct groups: total words from free speech tasks and total closed-class words from the Cinderella storytelling task.

Conclusion: Connected speech data can be used to distinguish PWA into coherent groups, providing insight into traditional aphasia classifications, factors that may guide discourse research and clinical work.

ARTICLE HISTORY
Received 9 November 2020
Accepted 30 August 2021

KEYWORDS
Aphasia; discourse; clustering methods; classification

Introduction

Two seemingly opposite things are simultaneously true of every individual with aphasia. Each individual's presentation is unique, and yet, the general pattern of expressive and receptive language skills will most likely fall into one of the several typical patterns. As many have observed, the common groupings allow for generalisations that help guide assessment and treatment planning, communication among professionals, understanding of neural organization of language, and targeted scientific investigations (Bartlett & Pashek, 1994; Beeson & Bayles, 1997; Crary et al., 1992; Hillis, 2007; Marshall, 2010). Within

CONTACT Davida Fromm fromm@andrew.cmu.edu Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

Supplemental data for this article can be accessed here.

those groups, though, there can be big differences in severity and other variables that make them far from homogeneous. The question is whether large shared databases and automated language analyses allow for the application of new data analysis techniques that can teach us something new about language skills in aphasia, help us be more efficient and effective in our clinical research and management, and understand more about the brain and language.

The aphasia literature includes many good articles on aphasia classification and aphasia syndromes. The topic has been analysed, reviewed, summarized, and challenged from many perspectives including neuropsychological theories, psycholinguistic models, expressive and receptive language symptoms, correlations of language symptoms with lesion sites, neuroimaging techniques, neural networks, statistical modeling, and data mining techniques (Akhutina, 2016; Ardila, 2010; Axer, Jantzen, Berks, Südfeld et al., 2000; Axer, Jantzen, Berks, von Keyserlingk et al., 2000; Axer, Jantzen, von Keyserlingk et al., 2000; Bates et al., 2005; Caplan, 2003; Dick et al., 2001; Hillis, 2007; Jantzen et al., 2002; Kasselimis et al., 2017; Marshall, 2010; McNeil & Kimelman, 2001). Scientific advances continue to bring new data that allow for reinterpretations of conventional knowledge.

Aphasia syndrome classification: classic model

The classic neurobiological model, based on discoveries and insights over almost 100 years by luminaries like Broca, Wernicke, Lichtheim, and Geschwind, has informed much of aphasia syndrome classification for clinical and research purposes (Ben Shalom & Poeppel, 2008; Sundet & Engvik, 1985; Tremblay & Dick, 2016). Plenty has also been written about problems with the typical classification (i.e., anomic, Broca's, conduction, global, transcortical, Wernicke's, etc.), for example: the classification features (e.g., impairments in naming and grammar) overlap aphasia category boundaries, the groups are not homogeneous, the syndromes may relate more to the vascular organization of the brain than to any particular site of lesion, and not all people with aphasia (PWA) fit into the traditional types (Axer et al., 2000; Caplan, 2003; Schwartz, 1984). Caplan (2003) also explained that the classical syndromes relate to performance on overall language performance (e.g., comprehension, naming, repeating) as opposed to performance of specific language tasks. Other issues include individual variability in brain anatomy, language networks, and the fact that terms such as "comprehension" and "fluency" are too generic (Kasselimis et al., 2017).

As this project focused on discourse variables, we concentrate here on aphasia-type classification research based on features of connected speech. One example of a discourse feature that overlaps category boundaries is the production of grammatical errors in connected speech. Errors of grammatical omission and simplification (agrammatic errors) are typically associated with nonfluent aphasia (mostly Broca's aphasia), whereas other grammatical errors (paragrammatic errors of substitution, omission, addition) occur more often in fluent aphasia (mostly Wernicke's aphasia). Paragrammatic errors are more difficult to identify and may coexist with agrammatic errors in some PWA (Matchin et al., 2020). Further, several methods exist to quantify agrammatism in aphasia, such as Northwestern Narrative Language Analysis (NNLA; Thompson et al., 1995) and Quantitative Production Analysis (QPA; Saffran et al., 1989), but no such methods are available for paragrammatism. Gordon (1998) showed that almost half of a group of 24

practicing speech-language pathologists (SLPs) identified grammatical factors as the most salient characteristic contributing to impressions of fluency in aphasia. She found that only 3 of 10 PWAs were unanimously classified as fluent or non-fluent by 24 practicing clinicians rating expressive language samples. Speech fluency, however, is also compromised by naming impairments that are common in both fluent and non-fluent aphasia types. Word-finding problems can appear in connected speech as hesitations, fillers, revisions, and sentence fragments that may be perceived as fluency deficits and agrammatic (or non-fluent) aphasia (Clough & Gordon, 2020). Thus, grammatical deficits are not necessarily a clear and distinct classifying feature for a particular type of aphasia or even a determination of fluent versus non-fluent aphasia. Interestingly, dissociations between fluency and grammatical production have been reported in narrative speech of individuals with primary progressive aphasia (Thompson et al., 2012). A more nuanced approach to the measurement of grammar usage as well as fluency in aphasia could further inform basic characterizations of connected speech in the traditional aphasia syndromes.

The Aphasia Quotient (AQ) subtest scores of the Western Aphasia Battery-R (WAB; Kertesz, 2007) are often used to determine type of aphasia, based on spontaneous speech fluency, auditory comprehension, repetition, and naming performance. Most relevant to this project is that the spontaneous speech fluency score (≥5 vs. ≤4) separates the types into fluent and non-fluent categories. The score is a rating based on subjective judgments mostly about quantity and grammaticality of output along with other features, such as word-finding difficulty, paraphasias, and hesitations. Since its initial version in 1982, several articles have addressed WAB classification issues (Crary et al., 1992; Ferro & Kertesz, 1987; John et al., 2017; Swindell et al., 1984; Trupe, 1984; Wertz et al., 1984). At the level of aphasia types, Crary et al. (1992) used a cluster analysis and found that only 30% of their 47 participants' original WAB aphasia types corresponded with the classification that resulted from their Q-analysis. Each factor contained participants with multiple WAB aphasia types. As an example, one of the three factors included participants whose original types (based on AQ subtest scores) were global, anomic, Broca's, conduction, and Wernicke's. Wertz et al. (1984) found that the WAB and another commonly used aphasia test, the Boston Diagnostic Aphasia Exam (Goodglass & Kaplan, 1983), agreed much better in measuring severity than in classifying aphasia types. It seems useful, then, to consider ways to enhance the current classification system with regard to characteristics of spontaneous speech.

Aphasia classification: multivariate and machine learning approaches

Recently, researchers have taken a more empirical approach and applied new techniques to aphasia classification. (Readers are urged to consult Wilson & Hula, 2019 for a survey of recent multivariate approaches to studying the multifactorial aspects of aphasia.) For example, logistic regression, a statistical approach used to model a binary variable, was used to predict fluent or non-fluent group membership based on judgments of picture descriptions from a mixed group of adults with no brain damage, unilateral left hemisphere stroke with predominant frontal lobe damage, semantic dementia, Alzheimer's disease, mixed dementia, traumatic brain injury, and posterior cerebral artery stroke with damage to the posterior and inferior region of the temporal lobe (Park et al., 2011). Results

showed that of five predictors, three were able to discriminate fluency status with over 95% accuracy: productivity, speech rate, and audible struggle. In fact, productivity had the greatest influence on judgments, as listeners (three doctoral level SLPs) were 18 times more likely to judge speakers "fluent" if they verbalized for more than 50% of the time, that is, produced more output (Park et al., 2011).

Another approach to classification uses neural networks, which look for similarities among example inputs and then classify similar cases into groups. This is a machine learning technique, in which a computer is trained to learn by examples that have been classified in advance. Using this approach with spontaneous speech subtest scores from the Aachen Aphasia Test (AAT; Huber et al., 1984), Axer, Jantzen, Berks, von Keyserlingk (2000) found that the neural network was better at classifying Broca's and global aphasia (both over 95% correct) than anomic and Wernicke's aphasia (75% and 83%, respectively). A second test with four different subtest scores (melody of speech, grammar, repetition, and reading aloud) yielded improved classifications, all above 90% correct for Broca's, global, and Wernicke's groups, and improved but lower classification accuracy for the anomic (83%) group. The authors concluded that the anomic group was not as homogenous as the others. In another report, Axer, Jantzen, von Keyserlingk (2000) used AAT data from 254 patients to compare the neural network model with a fuzzy model. Briefly, these models use fuzzy logic, which is more like human reasoning and decision making and allows for partial membership in a particular group. The fuzzy model still involves training based on classifications made to a training set before classifying cases from the test set. This model succeeded in building neural networks that successfully classified 87% of patients based on spontaneous speech data alone and 92% of patients based on additional comprehensive data (e.g., including comprehension, repetition, reading, writing). The standard for comparison was an expert's diagnosis of aphasia type using the basic aphasia syndromes: Broca's, Wernicke's, global, anomic, conduction. These authors encouraged interdisciplinary and collaborative work to add cases and test models to better understand the language impairment and facilitate the diagnostic process.

Using the AAT database of 146 patients (with diagnoses of Broca, Wernicke, global, or anomic aphasia), Akbarzadeh-T and Moshtagh-Khorasani (2007) also compared neural networks and a hierarchical fuzzy model to aphasia diagnosis. Like Axer and colleagues, they first used data from the spontaneous speech subtest and then used a more comprehensive set of scores, achieving classification accuracies of 90.82% and 91.89%, respectively. The fuzzy approach achieved classification accuracies of 91.30% with the spontaneous speech data only and 93.61% with the comprehensive set of test scores. Statistical testing showed that for spontaneous speech tests only, the fuzzy model performed better than the neural network. For the comprehensive data set, the only advantage of the fuzzy model was that it required fewer measurements (grammar, compound word repetition, confrontation naming subtests) and calculated more quickly than the neural networks.

Purpose

In this paper, we investigated enhanced systems of aphasia classification by applying unsupervised clustering methods, using only data from the connected speech of PWA. This study differs from those just reviewed in that they used methods to classify

participants according to previously determined aphasia categories (fluent/non-fluent, aphasia types), whereas the goal here is to allow the data to suggest new categories for classification of PWA. Instead of using a "supervised" approach where the classification of new participants is based on an existing classification system, we used an "unsupervised" approach where a new classification of PWA was generated based on similarities of discourse characteristics and not on previously determined aphasia categories. Once the new categories were identified, we then investigated the characteristics of each new category, that is, which variables best describe each new category, using supervised statistical learning methods (e.g., classification trees). The use of objective, quantitative criteria along with modern statistical methods may help identify key characteristics of spoken discourse that can better inform the process of aphasia classification for both clinical and research purposes. With over 500 different measures used in the literature to analyze connected speech, clinicians and researchers have been working to identify a core outcome set for aphasia treatment research (Bryant et al., 2016). Results of this study can make an important contribution to recent calls for a more systematic approach to discourse measurement and analysis in aphasia studies (Dietz & Boyle, 2018; Kintz & Wright, 2018; Stark et al., 2021).

Methods

Participants and materials

Data were collected from the AphasiaBank database (https://aphasia.talkbank.org/), a shared database of multimedia interactions for the study of communication in aphasia. At the time of this study, the database contained transcriptions of standard discourse protocols from 306 PWA and extensive demographic data on all participants. In addition, three standardized measures were administered: (1) the AQ subtests from the WAB-R; (2) the short form of the Boston Naming Test-Second Edition (BNT; Kaplan et al., 2001); and (3) the Verb Naming Test from the Northwestern Assessment of Verbs and Sentences-Revised (VNT; Cho-Reyes & Thompson, 2012). Word-level and sentence-level repetition skills were measured using a non-standardized AphasiaBank repetition test (https://aphasia.talkbank.org/protocol/repetition.pdf) with three parts: (1) closed and open word lists of increasing length; (2) sentences of increasing length; and (3) sentences with no errors, semantic errors (e.g., "The bird was caught by the worm"), and interference effects (e.g., "Count to ten as fast as you can").

Participants from 23 sites around the United States and Canada were tested with a standard discourse protocol comprising samples of free speech (stroke story, recovery, important event), picture descriptions (Broken Window, Refused Umbrella, Cat Rescue), storytelling (Cinderella), and procedural discourse (making a peanut butter and jelly sandwich). A script was used for administration of the discourse tasks so that investigator prompts were consistent throughout. All materials (stimulus pictures, script instructions, list of tests administered, demographics, test results, etc.) are at the AphasiaBank website given earlier.[1] All discourse tasks and testing, with the exception of the WAB-R and the comprehension tests, were recorded on video. (Participant characteristics are summarized below in the Statistical Analysis section.)

Language sample transcription and analysis

Discourse samples went through a detailed process of transcription, coding, and checking. Transcription was done using CHAT format, which operates closely with the CLAN programs that allow for the analysis of a wide range of linguistic and discourse structures (MacWhinney, 2000). These transcripts and their media files are also at the AphasiaBank website (with password protection). We coded word repetitions, revisions, fillers, sound fragments, gestures, and unintelligible output. Word-level errors were coded in four primary categories: phonological, semantic, neologistic, and morphological. Utterances that were non-task related (e.g., comments on the task, questions about the task) were excluded from analysis. Two full-time, trained transcribers with at least a Bachelor's degree in Linguistics or SLP reviewed each transcription, and the two reached forced-choice agreement on any discrepancies.

A variety of CLAN analyses were used to extract discourse variables for analysis. After transcripts were prepared and checked by at least two experienced transcribers, the MOR program was used to automatically create a morphological tier (%mor) and a grammatical relations tier (%gra) in the transcript. The %mor tier provides part-of-speech and morphological tags for each of the words in the utterance (excluding repetitions and revised content); the %gra tier provides a grammatical dependency parsing based on binary grammatical relations. Information from these tiers is used in the CLAN analyses that measure morphosyntactic and lexical aspects of the discourse (e.g., noun to verb ratio, number of embeddings) that are explained below. The morphological tagging accuracy of CLAN has consistently been between 95% and 97% (Huang, 2016; MacWhinney et al., 2011).

The variables used in the analysis included discourse data from each of the four discourse genres as well as demographic variables and scores from the BNT, VNT, and the AphasiaBank Repetition Test. The discourse variables included basic measures that are used to characterize connected speech in aphasia and can be automatically computed from CHAT transcripts. They represent the major categories of discourse analysis foci used in 165 studies of discourse in aphasia reviewed by Bryant et al. (2016) and included measures of fluency, rate of speech, amount of output, morphology, lexical word class usage, word error frequency and type, and lexical diversity. The only variables that required manual coding were the word-level error codes. Table 1 contains a complete list of the variables. A benefit of the machine learning technique approach is that it takes all the variables and identifies which are the most important for both separating and defining groups. We can cast a wide net and let the program reveal the ones that work best to both distinguish and then unify the participants.

Four CLAN programs were used to generate variables for the analysis (https://talkbank.org/manuals/CLAN.pdf). Each command was run on the full set of transcripts for each of the discourse genres.

(1) The GEM command was used to extract segments representing the different discourse genres from the master transcripts, creating a set of CHAT files that included free speech tasks, picture description tasks, the storytelling task, and the procedural discourse task for each participant.

Table 1. List of variables.

Category	Variable
Demographics	sex
	race
	handedness
	education
	aphasia etiology
	aphasia duration
Boston Naming Test	total score
Verb Naming Test	obligatory 1-place – total
	obligatory 2-place – total
	optional 2-place – total
	obligatory 3-place – total
	optional 3-place – total
	total score
AphasiaBank Repetition Test	I.A. closed word list – longest word string
	I.B. open word list – span score, any order
	I.B. open word list – span score, serial order
	II.A. sentences – longest successful
	II.A. sentences – # words correct
	II.A. sentences – # readministered
	II.B. sentences – # words correct
	II.B. sentences – # readministered
	II.B. sentences, no error – # words correct
	II.B. sentences, no error – # readministered
	II.B. sentences, semantic error – # words correct
	II.B. sentences, semantic error – # readministered
	II.B. sentences, interference effect – # words correct
	II.B. sentences, interference effect – # readministered
	II.B. sentences, interference effect – # commands followed
	II.B. sentences, interference effect – # questions answered

(*Continued*)

Table 1. (Continued).

Category	Variable
Discourse measures	duration
-Free speech	total utterances
-Picture descriptions	mean length of utterance (MLU) in words
-Cinderella story	total words
-Procedural	words per minute
	moving average type-token ratio (MATTR)
	idea density
	noun_verb ratio
	open_closed class ratio
	# and % open class words
	# and % closed class words
	# of revisions
	# of repetitions
	# and % of utterances with revisions
	# and % of utterances with repetitions
	# and % of utterances with revisions and/or repetitions
	# and % word errors (including repetitions and revisions)
	# morphological errors (including repetitions and revisions)
	# morphological errors missing an obligatory marker
	# and % neologistic errors (including repetitions and revisions)
	# and % phonemic errors (including repetitions and revisions)
	# and % semantic errors (including repetitions and revisions)
	# dysfluencies in words (including repetitions and revisions)
	% morphological errors (of all errors)
	# unintelligible segments
	# fillers
	fillers per minute
	fillers per words (including repetitions and revisions)
	# sound fragments
	% nouns
	% verbs
	% determiners
	% prepositions
	% adjectives
	% adverbs
	% conjunctions
	% pronouns
	# and % utterances coded as jargon

(2) The EVAL command (Forbes et al., 2012) was used to generate 21 outcome measures for the analyses, including a variety of part-of-speech and grammatical variables.
(3) The FREQ command was used to (a) compute the moving average type-token Ratio (MATTR; Covington, 2007), a strong and unbiased measure of lexical diversity in aphasia (Fergadiotis et al., 2013); (b) count the number of unintelligible segments in each transcript; (c) count the number of filled pauses and sound fragments; and (d) compute the total of each type of word-level error (semantic, morphological, phonological, neologistic, and within-word dysfluencies) including those made in repetitions and revisions.
(4) The KWAL and FREQ commands were used to compute the number of utterances with repetitions and revisions. The KWAL command first pulls out all the utterances with a repetition or revision, creates a new file of those utterances, and then the FREQ command counts the number of utterances in that new file.

Statistical analysis

As noted earlier, our goal was to investigate enhanced systems of aphasia classification using unsupervised clustering methods, specifically K-means clustering. When we clustered the observations of a data set, we sought to partition them into distinct groups so that the observations within each group were quite similar to each other, i.e., the within-cluster variation was as small as possible, while observations in different groups were quite different from each other. This approach is called unsupervised because we try to discover structure, in this case, distinct clusters of PWA, based on discourse characteristics of connected speech. A practical issue in the application of clustering analysis is deciding how many clusters, K, to look for in the data. We used the elbow plot to identify a value of $K = k$, where the change in total within-cluster variation is small relative to the value for $k - 1$ or $k + 1$ clusters. If a single value of K does not stand out from the plot, i.e., the "elbow" is not sharp, we try several values to see if the interpretation of the identified groups changes very much and look for the choice of K with the most useful, interpretable solution.

After assigning labels to identify the clusters of PWA identified by K-means clustering, we next used a set of supervised statistical methods based on classification trees to investigate the key characteristics of spoken discourse that best characterize the new groupings. The use of a single classification tree is highly sensitive to the set of observations used to create that tree. Although K-means clustering does a good job of distinguishing clinical profiles, it does not specify which discourse measures work best for accurate classification. To address this, we applied random forest methods to create a prediction model that can be used in actual clinical practice based on a connected speech sample. Random forest methods involve a computationally intensive approach for generating multiple classification trees using subsets of the predictor variables (e.g., features of connected speech) from subsamples of the data set to assign each PWA to one of the K-means clusters. Random forest analysis itself would not be used in clinical practice. Rather, it is a representative classification tree selected from the random forest that could guide clinical research and practice.

Specifically, random forest constructs a collection of classification trees, say B trees, based on randomly selected sets of observations and predictor variables selected from the data set. Each tree assigns a PWA to a label, i.e., a cluster group. Across the B trees we determine the most prevalent label assigned to a PWA and choose that label as the predicted K-means cluster group for that PWA (Breiman, 2001; James et al., 2013). A feature of the random forest approach, called "feature bagging", is that instead of using all the discourse variables in developing each classification tree, a randomly selected subset of variables is used for each tree, which has the effect of making each individual tree more unique and reduces correlation between trees, improving the random forest's overall performance (Breiman, 2001). An added advantage of using feature bagging is that it can be used to help identify the set of discourse variables that is the most informative in contributing to the classification of PWA into clusters.

To assess how well the predictions from the random forest model did, we compared the classifications of a subset of the PWA who were held out of the training sample that created the random forest prediction model and cross-classified their predicted labels to their assigned labels based on K-means clustering. Here, we were looking for the degree-of-agreement between the two. Finally, to investigate further what we could learn about the new classification system based on the discourse characteristics of connected speech, we cross-classified the PWA using the traditional aphasia classifications versus the new enhanced system to better understand where there was agreement and, more importantly, where the two schema differed. Statistical analyses were done using R, version 3.4.1.

The Data. The AphasiaBank database included a total of 306 PWA. This study consisted of the 168 participants with complete observations (for all 221 variables) at the time of this study. Table 2 provides demographic characteristics of the PWA sample.

Table 2. Participants' (N = 168) Characteristics.

Variable	Mean (SD) or Frequency
Age	62 (12) years
Sex	96 males, 72 females
Handedness	149 right, 14 left, 5 ambidextrous
Education	15.3 (2.9) years
Race	141 White, 18 African American, 5 Hispanic/Latino, 2 Asian, 1 Native Hawaiian/Pacific Islander, 1 Mixed
Time post-onset	5.4 (5.0) years
Aphasia etiology	164 stroke, 4 other
WAB-R type	61 Anomic, 35 Broca, 32 Conduction, 11 Wernicke, 8 Transcortical motor, 21 not aphasic*

Note: WAB-R – Western Aphasia Battery-Revised (Kertesz, 2007)

** "not aphasic" refers to participants whose WAB-R Aphasia Quotient was above the test battery's "normal or nonaphasic" cutoff of 93.8 but who still considered themselves (and were considered by their clinicians) to be aphasic.*

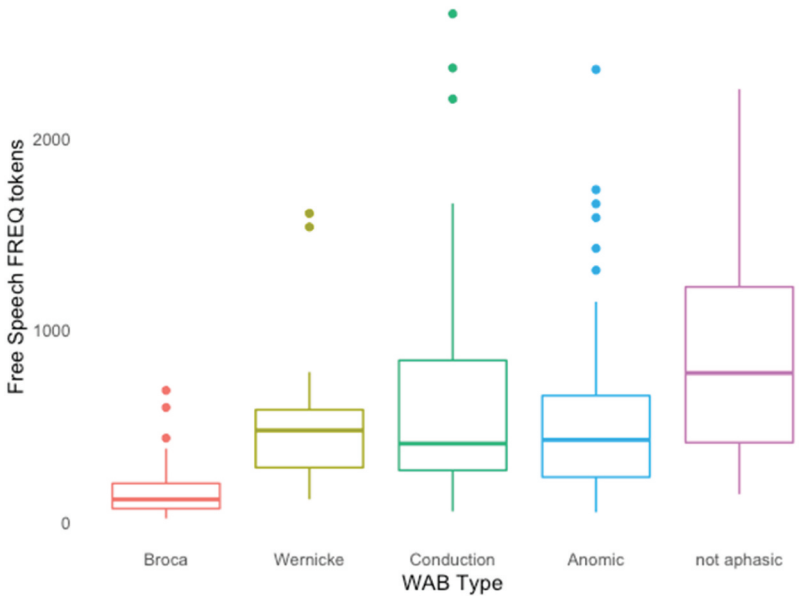

Figure 1. Total number of words on free speech discourse tasks by aphasia type. Note: In this box plot, the bottoms and tops of the boxes represent the first and third quartiles, respectively, of the distribution (number of words on free speech discourse tasks). The horizontal line inside each box represents the median. The vertical lines extending above and below the boxes represent the minimum and maximum data points, excluding outliers which are represented by dots and signify data points outside 99.3% of the distribution.

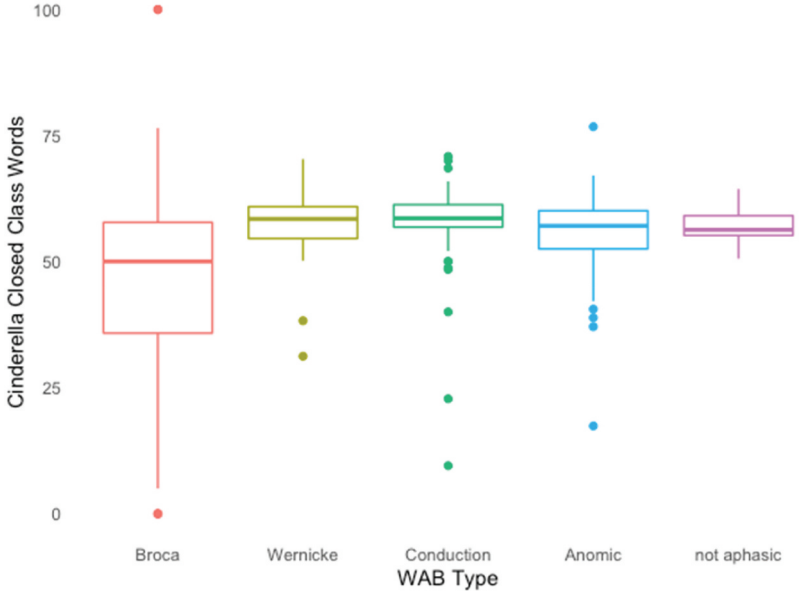

Figure 2. % of closed class words on Cinderella task by aphasia type.

Results

Exploratory data analysis

First, we used exploratory data analysis techniques to examine the performance on the discourse variables for each traditional aphasia type. Box plots in Figures 1 and 2 provide examples of these results. Figure 1 shows the total number of words produced on the free speech tasks, and Figure 2 shows the percent of closed-class words produced on the Cinderella storytelling task. Both figures clearly show that traditional aphasia types have limited value in discriminating performance on each of these discourse variables alone. That is, knowing that an individual produced approximately 550 words on the free speech task, might indicate that the individual probably did not have Broca's aphasia, but it would not help discriminate among the four fluent types of aphasia. Likewise, knowing that, on average, more than half of the words used in the Cinderella story were closed-class words might tell you the same thing. These results indicate, unsurprisingly, that there is a great overlap across aphasia types when looking at single dimensions of discourse.

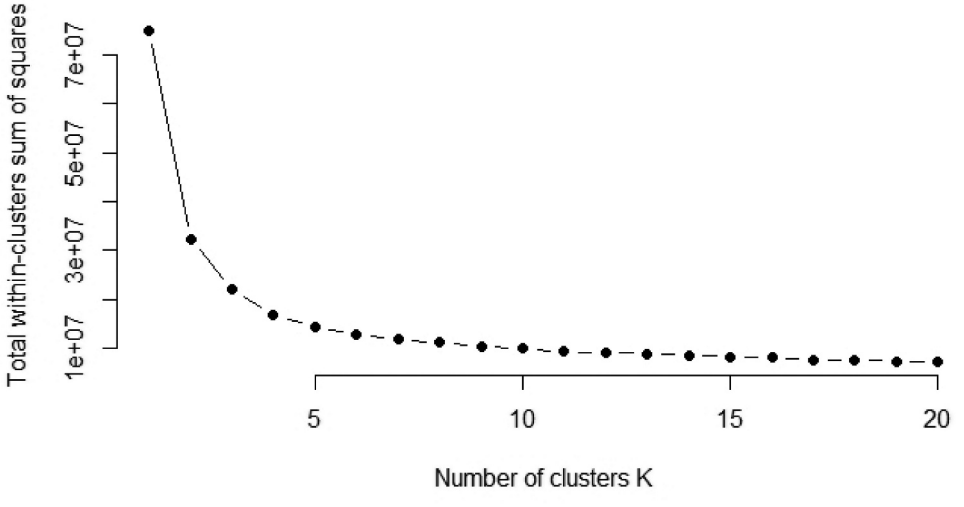

Figure 3. Total within-cluster sum of squares for different K values in K-means.

Table 3. Number of individuals per cluster by K-means.

K-means Cluster	Count	Fraction
a	20	0.12
b	20	0.12
c	10	0.06
d	28	0.17
e	35	0.21
f	27	0.16
g	29	0.17

The next steps involved using K-means clustering, an unsupervised machine learning approach, to best identify coherent groups of PWA using our total collection of discourse variables. Two supplemental data tables can be accessed at supplement tables with means and standard deviations for the formal and informal testing (S-Table 1) and the discourse data across all four genres (S-Table 2).

Identifying new aphasia clusters based on K-means

To discover new clusters based on language behaviors and patterns, we applied the K-means algorithm to the language discourse data set. We first applied the elbow method to identify the optimal *K* value (number of clusters) for the K-means algorithm. Figure 3, the elbow plot, shows the total within-cluster sum of squares for different values of *K*. Based on the plot, we picked $K = 7$ for the number of clusters, as this value provides a low within-cluster sum of squares, meaning that the variability of participants within each cluster would be quite low and differ minimally from $K = 6$ or $K = 8$. We explore the sensitivity of our results in choosing $K = 7$ versus 6 or 8 later.

We applied the K-means algorithm with $K = 7$, and labeled these 7 clusters alphabetically from *a* to *g* in arbitrary order. Table 3 displays the number of participants who were categorized into each of the new aphasia clusters. Clusters *e* and *c* are the largest and smallest, respectively; the other clusters consist of roughly similar numbers of PWA.

The first author (DF) reviewed video files of the PWA participants to describe the general perceptual characteristics of the connected speech in the seven new clusters (*a*–*g*). Table 4 displays the key characteristics for these seven clusters. The *a*–*g* clusters revealed coherent and distinctive connected speech groupings based primarily on characteristics of amount of output (total words, total utterances) and fluency. In this case, *fluency* meant that the connected speech had some normal syntactic sentence structure and melodic line. The other distinguishing characteristics involved the relative frequency of behaviors, such as word errors (e.g., paraphasias, neologisms), repetitions, revisions, fillers, and sound fragments (incomplete word attempts), all of which were present to some extent in all samples and all of which can be considered aspects of fluency (Gordon, 1998). Clusters *a* and *f* were typical of non-fluent aphasia, with *a* being more severe and containing mostly single words with no complete sentences. The samples in cluster *a* had some word errors and jargon as well as some fillers and fragments; the samples in cluster *f* also had word

Table 4. Characteristics of aphasia K-means clusters.

K-means Cluster	Fluency	Amount of output (length and complexity)
a	Non-fluent	Very limited Mostly single words
b	Fluent	Mostly normal
c	Fluent	Lengthy output
d	Fluent	Reasonable
e	Fluent	Reasonable
f	Non-fluent	Limited
g	Non-fluent/fluent	Somewhat limited Slow with pauses

errors, fillers, fragments. The characteristics of cluster *g* had elements of both non-fluent and fluent aphasia with some more syntactic sentence structure but slow production and notable pausing as well as word errors, repetitions, revisions, fillers, and fragments. Within the fluent clusters, *d* and *e* had less output than *b* and *c*. Utterances in cluster *d* samples were less grammatically complex than those in cluster *e*, and they contained more word errors, repetitions, revisions, fillers, and fragments. Cluster *c* samples had more word errors, repetitions, revisions, fillers, and fragments than cluster *b*, which were the samples that showed the least overall impairment. Based on this expert assessment, we concluded that these clusters of PWA developed by the K-means algorithm were able to differentiate and thereby characterize basic underlying language patterns in individuals with aphasia. These will be discussed more in the Discussion section.

The enhanced classification system

With the new clusters of PWA identified based on K-means (with $K = 7$, as described in Table 3), we tagged the participants with their associated K-means cluster label (*a*–*g*) and used random forest methods, a supervised machine learning approach, to develop a prediction model based on classification trees to assign new PWA to groups based on their language output.

The importance of each variable in the new classification prediction model derived from the random forest algorithm is displayed in Figure 4. The prefixes on each variable in Figure 4 correspond to the specific discourse genre: F = free speech narratives, P = picture description, C = Cinderella storytelling. However, more important is the category of the

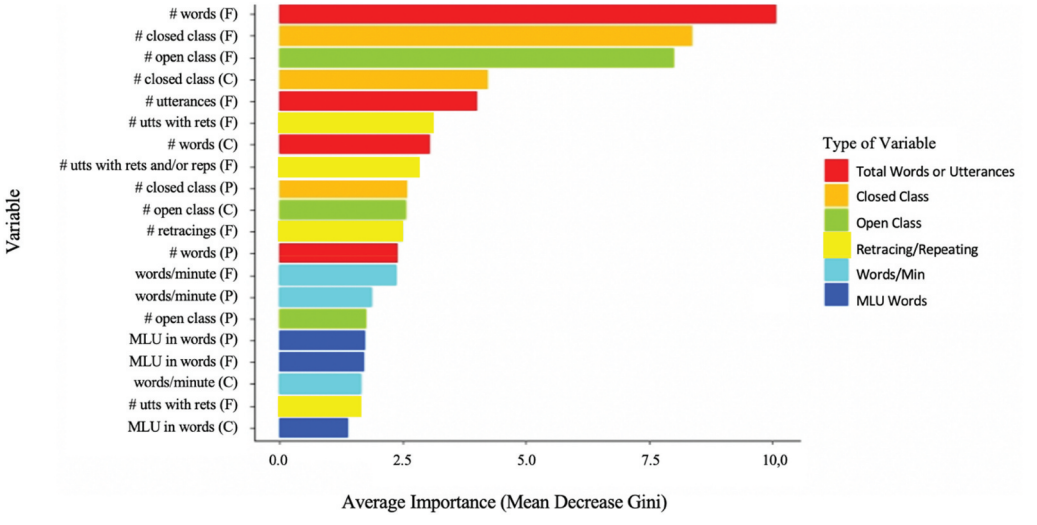

Figure 4. Average variable importance for enhanced aphasia classification by random forest ($k = 6$, 7, and 8). Note: (F) = free speech narratives, (C) = Cinderella storytelling, (P) = picture descriptions, MLU = mean length of utterance, rets = retracings, reps = repetitions

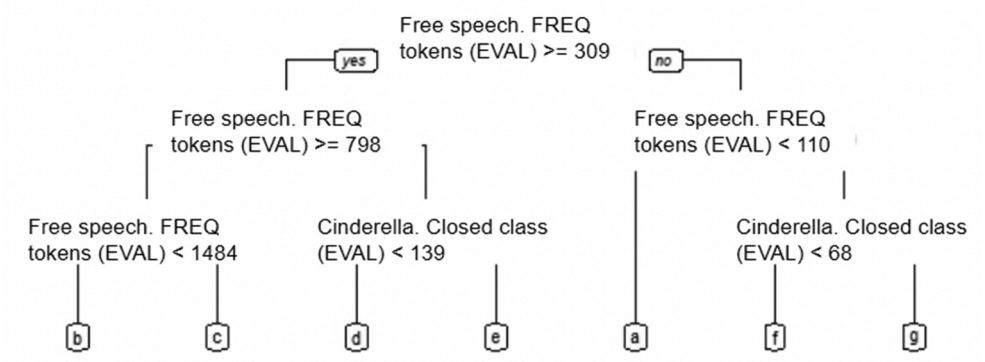

Figure 5. Best decision tree from 10-fold cross validation.

variable (refer to Table 1 for the full list of variables). The figure illustrates that the amount of output (measured by total number of tokens, total number of utterances, etc.) and the lexical content of the output (number of open class and closed-class words) were among the most important categories of discourse variables used in the model to assign subjects to a K-means cluster category.

A representative classification tree from cross-validation that assigns participants into one of the *a*–*g* classes is displayed in Figure 5. The classification tree provides a simple and mathematically sound model for assigning PWA to groups based on two variables: *total number of words in the free speech tasks* and *total number of closed-class words in the Cinderella storytelling*. Starting at the top of the figure, the first split is based on the free speech task: less than 309, yes or no. Then, if a PWA's total word count on the free speech tasks is less than 309 and even less than 110 in the free speech task, that PWA would be assigned to the *a* cluster. If a PWA's total word count on the free speech tasks is greater than 309 but less than 798, and the number of closed-class words is greater than 139, that PWA would be in the *e* cluster.

Given that only two variables, total number of tokens from the free speech task and the total number of closed-class words from the Cinderella task, were used to create the splits within the tree classifier, we re-examined the results from the feature selection from the random forest analysis. As Figure 4 shows, the two variables from the classification tree (Figure 5) are the first and fourth most important variables selected by random forest. To assess the robustness of our results, we also redid the analyses with $K = 6$ and $K = 8$ clusters and found results similar to the model presented here using $K = 7$ clusters, meaning in each case two types of variables on average were identified using random forest, one being total words or utterances and the other, the number of closed-class words. Thus, we find that the features that predict the K-means clusters are the same whether $K = 6$, 7, or 8.

We also validated the accuracy of the tree by comparing the classification tree grouping results from random forest analysis to the K-means clustering results for the testing data set, the training data set, and the full data set. Our tree classifier achieved high accuracy (agreement = 0.86) on the testing data set, the training data set (agreement = 0.92), and the entire data set (agreement = 0.91). This provides further confidence

Table 5. Confusion matrix of K-means predictions (columns) and tree classification system (rows).

K-means							
Tree	a	b	c	d	e	f	g
a	20	0	0	0	0	2	0
b	0	20	0	0	0	0	0
c	0	0	10	0	0	0	0
d	0	0	0	24	1	0	1
e	0	0	0	4	34	0	1
f	0	0	0	0	0	24	3
g	0	0	0	0	0	1	24

that the tree serves as a simple yet robust algorithm to identify clusters of PWA based on measures defined by the K-means algorithm. The confusion matrix in Table 5 provides a cross-classification of the *a–g* groups based on the prediction model using the tree classifier based on the two features (rows) compared to the K-means cluster assignments (columns) for all PWA in the full data set. For example, there were 34 PWA assigned to cluster *e* by the unsupervised K-means algorithm who were also placed into group *e* using the prediction model based on the tree classifier's measures (total number of words on free speech tasks and total number of closed-class words on Cinderella task); one individual from the K-means *e* cluster was placed into group *d* by the tree classifier. The values along the main diagonal of Table 5 show the strong agreement between the two classification methods. There are relatively few cases of the main diagonal that would indicate a lack of agreement between the two classification methods. With only a few exceptions ($n = 13$), the participants were placed in the same groups with both classifiers. Of the 13 exceptions, 11 K-means predictions were in neighboring tree groupings (e.g., *e* instead of *d, g* instead of *f, f* instead of *a* and *g*). For example, two individuals in the K-means *d* cluster were classified in the *e* tree group because the number of closed-class words in their Cinderella stories was 134 and 137, just below the 139 word cutoff, which would have placed them in the *d* group, suggesting that these few examples of lack of agreement were minor.

Comparison of original and enhanced classification systems

We were also interested in comparing and contrasting the new classification system with the traditional WAB-R classification system. Table 6 illustrates the concordance between WAB-R aphasia classification and the *a-g* classes from the predictions from the best

Table 6. Confusion matrix of traditional WAB-R aphasia types and tree classification system.

	Tree Classifications						
WAB-R Types	a	b	c	d	e	f	g
Anomic	4	8	3	11	20	6	9
Broca's	17	0	0	3	2	11	2
Conduction	0	5	3	8	7	2	7
Not aphasic	0	7	3	3	4	0	4
TCM	1	0	0	0	0	6	1
Wernicke's	0	0	1	1	6	2	2

Key: TCM = Transcortical motor

classification tree. Interestingly, many of the WAB-defined aphasia types span multiple *a-g* groups under the enhanced classification system. For example, the 35 PWA identified by WAB-R classification as Broca's are located in five of the tree classification clusters, though mostly concentrated in clusters *a* (n = 17) and *f* (n = 11). These results will be discussed in more depth in the next section.

Discussion

In this study, we explored an enhanced aphasia classification system based on spoken language samples and some repetition and naming test scores from individuals with aphasia. Seven distinct groups were identified by a K-means algorithm with the optimal K-value suggested by the elbow method. Using the new groupings suggested by K-means clustering as the response variable, we discovered a classification scheme using the random forest algorithm to classify individuals with aphasia. As the tree classifier was capable of achieving high accuracy with the testing set in predicting participants' new aphasia groupings, we are confident that the tree is robust and are encouraged that this model will help clinicians and researchers identify salient characteristics that better characterize the language output of PWA.

Aphasia clusters

The clustering had a surprisingly simple structure, as only two measures created splits within the tree classifier: total number of words in the free speech (personal narrative) tasks and number of closed-class words from the Cinderella storytelling task. These results support both intuitive and empirically demonstrated factors that are important in evaluating language ability. Total number of words represents the amount of speech output the person generates, a measure of productivity. Amount of closed-class words (e.g., pronouns, determiners, conjunctions, prepositions), concerns the type of words in the output and can be viewed as a crude measure of grammaticality. In a study of 27 common auditory-perceptual features of connected speech in aphasia, factor analysis showed that logopenia (paucity of speech) and agrammatism were two of the four underlying factors accounting for 79% of the variance in speech samples (Casilio et al., 2019). Interestingly, proportion of closed-class words was one of two measures used to assess grammatical deficits in a recent article mapping articulatory and grammatical aspect of fluency deficits in aphasia and was significantly correlated with aphasia severity but not correlated with consistent lesion locations (Mirman et al., 2019). Both measures, total words and proportion of closed-class words are elements of the QPA (Saffran et al., 1989) and NNLA (Thompson et al., 1995), two systematic approaches to measuring syntactic ability in connected speech. The fact that number of words and number of closed-class words were key elements in the clusters of discourse from a large group of PWA demonstrates that the ability to produce more speech output and form utterances with more grammatical words serves to differentiate and separate PWAs into different groups. It would be interesting to investigate these new groups with analyses based on a multi-level discourse-processing model, as Sherratt (2007) did with discourse from non-brain-damaged males. She analysed seven areas based on the model (e.g., relevance, discourse grammar, fluency) using 23 specific outcome measures. One of the findings that bears on the results

of this study, was that longer samples (more output) were associated with an increased proportion of cohesive ties (especially conjunctions and lexical ties) and syntactic complexity.

The variable importance plot (Figure 4) showed that in addition to the productivity measures (total words, total utterances) and word class frequencies (closed class, open class), the next most important measures involved retracings (revisions), measured as both the raw total number of retracings and the frequency of utterances that included any number of retracings. Retracings are a very telling feature of connected speech in aphasia. As a rule, individuals with Wernicke's type fluent aphasia are truly fluent, in the sense that they produce speech in an uninterrupted flow with infrequent repetitions and revisions, sound fragments, fillers, word-finding pauses, and hesitations (Pallickal, 2020). The connected speech of individuals with conduction type fluent aphasia, on the other hand, typically contains multiple attempts at self-correction. In fact, conduit d'approche, successive attempts at a target word, is a classic feature of conduction aphasia (Bartha & Benke, 2003). Other PWA also exhibit revisions that can be revealing in a number of ways. They may indicate an ability to self-monitor in real time, demonstrating the speaker's knowledge of the target word, awareness of the error, and desire to repair it. In less fluent contexts, they may indicate a coexisting apraxia of speech, a motor speech disorder that often co-occurs with aphasia and manifests with inconsistent speech sound errors, articulatory groping, restarts, and attempts at revisions (Haley et al., 2021; Strand et al., 2014; Van der Merwe, 2007). As seen in the excerpts from CHAT files in Figure 6, retracings (marked with [//]) may occur following paraphasias and reformulations due to a host of factors, but they commonly result from problems finding and producing words. Other CHAT markings in these utterances include: repetitions marked with [/]; target words in square brackets with a colon next to error productions; phonetic transcriptions of error productions tagged with @u; sound fragments preceded by &+ symbols; and (.) for a short pause. Along with the quantity of output and number of closed-class words, retracings are easy to identify and helpful in discriminating among speakers with aphasia. Therapeutic goals may include ways to modify the retracings to work toward maximizing communication success but also minimizing the frequency of these behaviors.

Aphasia cluster characterizations and comparison with traditional aphasia types

Along with amount of output and fluency, the subjective features that were most salient in describing the perceptual characteristics of the clusters were repetitions, revisions, paraphasias, fillers, and sound fragments. These features correspond to results of the random forest analysis discussed earlier. Interestingly, they also correspond to factors that accounted for 79% of the variability in the Casilio et al. (2019) study of auditory-perceptual speech ratings of connected speech in aphasia: logopenia (paucity of output), agrammatism, paraphasia, and motor speech (e.g., pauses, reduced rate, halting, and effortful speech). The results of the K-means clustering analysis in the current study revealed two distinctly non-fluent groups, with halting production, mostly single words or short phrases. Participants in group *a* had language output typical of individuals with severe Broca's aphasia, and those in group *f* resembled speakers with more traditional, less severe, chronic Broca's aphasia. Cluster *g* straddled the fluency fence with somewhat limited output but more propositional utterances with more grammatical

*PAR: well kick [//] cook [: kick] [//] no [//] cook [: kick] [//] <no that's not> [//] kʊp@u [: kick] [//] &-uh kɛt@u [: kick] [/] kɛt@u [: kick] ball.

*PAR: so &+i &+i (.) the [/] the [/] &+s <the boy> [//] the [/] the [/] &+s <the man> [//] <the [/] the man> [//] the woman has [//] &+le lets go of the umbrella [/] &-uh &+le umbrella.

*PAR: and he's tryin(g) to get <a chair [: ladder]> [//] the desk [: ladder] [//] table [: ladder].

*PAR: the &-uh læmbəˑ@u [: window] [//] ræmbəˑ@u [: window] [//] rændo@u [: window] [//] wɛndəˑ@u [: window].

*PAR: and then at [//] the [/] &+m <the husband [: mother] &+nay I'm sorry> [//] <the wife [: mother]> [//] (.) or <the &-uh (.) mother> [/] the mother &-um too [//] was sad &+an about the [/] (.) the boy.

*PAR: &-um above <the king> [/] <the king in> [//] <or the to> [//] or [/] <or maybe the queen> [//] the [/] the [/] &+k the &-uh prince &-um (.) were trying to make something that [/] that nice.

*PAR: but he was so &+m mark [: smart] [//] &-uh smɑrk@u [: smart].

*PAR: I'm doing [//] &-uh talking &+red &-uh ready [: better] [//] better than &+u &+o juʒuə@u [: usual].

*PAR: and it startles the man <next to him> [//] next to it .

*PAR: and <they wanted> [//] <they [/] they didn't> [//] I wasn't able to talk .

*PAR: and it starts raining off [: on] [//] on him when he leaves the house .

*PAR: &-uh <I [/] I &-uh was> [//] &-uh &-uh I remember about a year ago .

*PAR: but sometimes is [//] &-uh early mornin(g) is a bad day .

Figure 6. Examples of retracing in CHAT transcripts.

elements but also grammatical errors, pronoun substitutions, paraphasias, noticeable pauses, repetitions, and revisions. The other four groups were distinctly fluent, meaning the connected speech had some normal syntactic sentence structure and melodic line. The *b* cluster was the one that most resembled normal discourse output, and those samples had fewer total words and fewer of the other behaviors (e.g., word errors, repetitions, revisions, filler, and fragments) than cluster *c*. Cluster *c* had the lengthiest

output along with some word errors, repetitions, revisions, fillers, and fragments. As we know from discourse efficiency measures, this type of lengthy output can signify more word finding problems, circumlocution, semantic jargon, paragrammatism, and irrelevant or empty speech.

Comparing the clusters to the WAB-R aphasia types is interesting but should be done with caution. The intention here is not to recommend new aphasia subtypes or classification schemes, but to present the idea of coherent clusters of PWA that are based on connected speech data using unsupervised statistical algorithms. In addition to fluency in connected speech, the WAB-R determinations are based on subtests of auditory comprehension, repetition, and naming. Thus, we expect some dispersion of WAB-R aphasia types across the seven clusters that resulted from the K-means procedure, as seen in Table 6. Individuals with Wernicke's aphasia and transcortical motor aphasia were predominantly in single clusters (*e* and *f*, respectively), while participants with Broca's aphasia were mostly distributed across two clusters, where cluster *a* resembled severe Broca's aphasia and *f* resembled a more traditional, less severe, chronic presentation. Many articles over the years have discussed the multifaceted nature of Broca's aphasia, debating its unified syndrome status as well as its associated lesion location(s) (Caramazza et al., 2001; Drai & Grodzinsky, 2006; Fridriksson et al., 2015).

The dispersion was most marked for the anomic aphasia group. A study by John et al. (2017) compared WAB-R types with clinical impressions and found that only two of the 14 participants with anomic aphasia according to the WAB-R were judged to have anomic aphasia; 10 were judged to have Broca's and two to have Wernicke's aphasia. Though other groups also showed discrepancies in that study, the anomic group showed the most, including the crossover from what is considered to be a fluent aphasia into a non-fluent aphasia type. Casilio et al. (2019) commented that the auditory-perceptual factor loadings for individuals with the same type of aphasia (e.g., Broca's or Wernicke's) showed considerable diversity. Conversely, participants with similar auditory-perceptual profiles had different WAB-R aphasia types. Axer, Jantzen, Berks (2000) commented that their neural network classifier using spontaneous speech (six spontaneous speech subtests of the AAT) was inadequate at classifying anomic aphasia, which they hypothesized was not as homogenous a group as the other aphasia syndromes. Notably, the PWA with conduction aphasia and those who scored above the WAB-R aphasia cutoff were also dispersed across the clusters, though the PWA classified as "not aphasic" by the WAB did not fall into either of the non-fluent clusters.

A primary explanation for classification discrepancies with the WAB can be traced to the gating function of the fluency rating scale in the spontaneous speech section (Trupe, 1984). This portion of the test contains six conversational questions and one picture description. Fluency rating scores of 5 and above divide the fluent (anomic, Wernicke's, conduction, or transcortical sensory) and non-fluent (4 and below) aphasia types (global, Broca's, transcortical motor, or isolation). These ratings are based on a limited amount of connected speech assessment, much of which does not even require propositional phrases. In fact, all six of the conversational questions (e.g., *How are you today, Have you been here before, What is your first and last name*) could be appropriately answered without propositional phrases or sentences. Thus, the single picture description could be the basis for the fluency rating which distinguishes the fluent from non-fluent aphasia types. Someone with a fluent aphasia who has word finding problems and limited output

in describing the picture could be classified as having Broca's or transcortical motor aphasia, depending on the other subtest scores. Likewise, someone with a non-fluent Broca's aphasia who managed to produce some grammatical words and at least two propositional sentences could be classified as having conduction, anomic, or Wernicke's aphasia, depending on the other subtest scores.

The WAB was designed to be a comprehensive assessment to determine presence, severity, and type of aphasia based on linguistic skills in spontaneous speech, auditory comprehension, naming, and repetition. As such, the WAB-based classifications may lack the ability to identify the underlying differences in certain language output skills and establish distinct aphasia groups based on those behaviors. Our results provide another opportunity to raise some caution about strict adherence to these aphasia types for clinical and research purposes, specifically when the focus is on connected speech.

Discourse genres

Of the four discourse genres included in this analysis, the Cinderella story and free speech genres were the most useful in identifying and distinguishing clusters. They represent narrative and everyday discourse, the two main genres that clinicians and researchers typically elicit (Pritchard et al., 2018). The procedural discourse genre was represented by only one simple task, the sandwich task, and therefore had much less quantity and complexity of output than any of the others. The picture description tasks included three separate picture stimuli and prompts. These tasks can also be considered narratives or everyday discourse, but they are more expository in nature, similar to describing situations or events. Though two are sequenced and one is not, these expository tasks (the single "Cat Rescue" task and the sequenced "Broken Window" task) were shown to cluster together in analyses of discourse microstructure (Stark et al., 2021). All these stimuli are black-and-white drawings, perhaps not as engaging or relevant to the participants. The free speech tasks, on the other hand, included multiple prompts about the participants' speech, their stroke, their recovery, and an important event in their lives. It is conducted as a structured conversation, has more personal relevance, and is more like a normal, familiar conversational interaction than any of the other tasks. The Cinderella storytelling task is the second most frequently reported language sampling technique (the first being the Cookie Theft picture) used to elicit narratives in aphasia (Bryant et al., 2016). It has been shown to be useful in generating rich language samples that involve some orientation, precipitating action, and resolution (Armstrong, 2000). Many articles attest to its use in identifying language characteristics of fluent and non-fluent participants, highlighting recovery patterns, and showing changes following treatment (Bird & Franklin, 1996; DeDe & Salis, 2020; S. G. H. Dalton & Richardson, 2019; Jacobs, 2001; Stark, 2010; Thompson et al., 2003; Webster et al., 2007). In a large study of main concept production in five discourse tasks, the Cinderella task had the greatest number of large effect sizes indicating performance differences among subtypes of PWA (Dalton & Richardson, 2019). Results from our study – having included a large number of discourse measures from a range of discourse tasks as well as naming and repetition test scores – demonstrate that the quantity of grammatical words used in the Cinderella task and the total number of words produced in the personal narratives are uniquely valuable in distinguishing and identifying aphasic discourse.

Conclusions

The classification system developed here is capable of identifying underlying differences in individuals within the same WAB-defined aphasia types, and grouping them into new clusters. This discourse-based clustering system mapped the participants into different clusters based on a remarkably small number of discourse measures. Interestingly, test scores were not useful in distinguishing groups. Given the two related facts that 100% of SLPs elicit spoken discourse when analyzing PWA (Bryant et al., 2017) and that the aphasia literature reports over 500 different measures to analyze connected speech in aphasia (Bryant et al., 2016), these results answer the call for a more systematic approach to discourse measurement and analysis in aphasia studies (Dietz & Boyle, 2018; Kintz & Wright, 2018; Stark et al., 2021). Understandably, the clusters presented here relate to specific tasks administered as part of a specific discourse protocol. Thus, the decision tree divisions, for example, are not universally prescriptive. Nor are the perceptual descriptions of the clusters intended to be a unique or novel classification scheme. The analyses done here also involved transcription, which is time-consuming and not always conducive in clinical settings. Using the CLAN editor and automated CLAN analyses makes the entire process much more efficient, but still a potential barrier for busy clinicians. The intention was to illustrate how the use of unsupervised statistical techniques can shed light on salient targets to guide efforts to establish the most effective measures for assessment, treatment, and research of connected speech in aphasia.

Connected speech is the most functional and ecologically valid level of expressive language to assess (Craig et al., 1993; Gordon, 2020; Prins & Bastiaanse, 2004) but also challenging to quantify and analyze. We believe that the enhanced clustering scheme is meaningful both in research and clinical contexts. Machine learning approaches have begun to be used in other areas of discourse impairments such as Alzheimer's disease (Fraser et al., 2016), mild cognitive impairment (Lundholm Fors et al., 2018), and semantic dementia (Garrard et al., 2014). Our study revealed a robust system to classify multivariate groupings through the use of unsupervised machine learning with a large discourse data set from a large group of individuals with aphasia. The new system, based on the language patterns of individuals with aphasia, is capable of differentiating PWA based on clinically relevant and ecologically valid behaviors. Clinicians could combine the traditional aphasia-type classification system with the new tree classifier to gain insights about the associated characteristics with each aphasia type, and plan for more targeted treatment and outcome goals. Amounts of output, closed-class function words, revisions, repetitions, fillers, sound fragments, and word errors are straightforward and salient characteristics that can be identified and measured for assessment and treatment purposes. It can be argued that these are all essential aspects of what is perceived as *fluency* in connected speech. Similarly, researchers could further investigate the connections between the comprehensive traditional aphasia typing and the clusters revealed by this analysis of discourse. In summary, we are confident that our research provides interesting and practical information regarding language patterns of individuals with aphasia that can be applied to both clinical treatment and research contexts.

Future work

There are a number of ways for future work to improve on and extend this study. The final data set, using participants who had complete data, consisted of 168 participants and 221 variables. Though we have applied a train-test split and used cross validation to prevent overfitting in this paper, we hope that future research could (i) replicate the K-means and random forest procedure to verify whether similar results could be obtained independently, and (ii) apply our decision tree algorithm to classify new PWA, i.e., to assess the out-of-sample performance of the algorithm. Secondly, we used K-means clustering, which is a well-known clustering algorithm to perform the clustering task. Some disadvantages for K-means algorithm are that the algorithm requires pre-determining the optimal K-value, and the clustering results do depend on the initial random choice of cluster centers. Hence, future research can be dedicated to applying more consistent clustering methods such as density-based clustering or mean-shift clustering to investigate the underlying groupings of PWA based on language pattern. Interpretation of the model (Table 4) was based on subjective, qualitative auditory-perceptual judgments of one individual not blinded to the group assignments. Here, the goal was to learn about the application of these procedures to discourse data and present the results for continued scientific inquiry. Future research in this area should design more sophisticated validation procedures and include measures of rater reliability for perceptual judgments. In summary, while at this point, the tree structure in Figure 5 is likely the most easily interpretable and usable model available, it is possible that better classification models can come when we have a greater sample size to work with and include other variables in the analysis. Finally, relevant to the tree classifier, the results are raw frequencies of discourse variables specific to the tasks used in the AphasiaBank standard discourse protocol. Future studies should include other discourse variables and other tasks to test both the reliability and generalizability of these results. It would also be interesting to use this or some other unsupervised machine modeling approach to relate discourse variables to lesion-symptom mapping to further understand these important brain behavior relationships.

Note

1. Participant-related data are password protected and restricted to members of the AphasiaBank consortium group. Licensed SLPs, educators, and researchers who would like access can send an email request to Brian MacWhinney (macw@cmu.edu) with contact information, affiliation, and a brief general statement about how they envision using the resources.

Acknowledgments

Open Access funding provided by the Qatar National Library.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Institute on Deafness and Other Communication Disorders [Grant R01-DC008524] (2007-2022, awarded to MacWhinney).

ORCID

Davida Fromm http://orcid.org/0000-0002-4704-7709
Joel Greenhouse http://orcid.org/0000-0001-5087-9648
Brian MacWhinney http://orcid.org/0000-0002-4988-1342

References

Akbarzadeh-T, M. R., & Moshtagh-Khorasani, M. (2007). A hierarchical fuzzy rule-based approach to aphasia diagnosis. *Journal of Biomedical Informatics*, *40*(5), 465–475. https://doi.org/10.1016/j.jbi.2006.12.005

Akhutina, T. (2016). Luria's classification of aphasias and its theoretical basis. *Aphasiology*, *30*(8), 878–897. https://doi.org/10.1080/02687038.2015.1070950

Ardila, A. (2010). A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology*, *24*(3), 363–394. https://doi.org/10.1080/02687030802553704

Armstrong, E. (2000). Aphasic discourse analysis: The story so far. *Aphasiology*, *14*(9), 875–892. https://doi.org/10.1080/02687030050127685

Axer, H., Jantzen, J., Berks, G., Südfeld, D., & von Keyserlingk, D. G. V. (2000, September). The aphasia database on the web: Description of a model for problems of classification in medicine. In *Proc. ESIT* (Vol. 21), 104–110. ERUDIT.

Axer, H., Jantzen, J., Berks, G., & von Keyserlingk, D. G. V. (2000, September). Aphasia classification using neural networks. In *European Symposium on Intelligent Techniques Aachen* (Vol. 111, p. 5). ERUDIT.

Axer, H., Jantzen, J., & von Keyserlingk, D. G. (2000). An aphasia database on the internet: A model for computer-assisted analysis in aphasiology. *Brain and Language*, *75*(3), 390–398. https://doi.org/10.1006/brln.2000.2362

Bartha, L., & Benke, T. (2003). Acute conduction aphasia: An analysis of 20 cases. *Brain and Language*, *85*(1), 93–108. https://doi.org/10.1016/S0093-934X(02)00502-3

Bartlett, C. L., & Pashek, G. V. (1994). Taxonomic theory and practical implications in aphasia classification. *Aphasiology*, *8*(2), 103–126. https://doi.org/10.1080/02687039408248645

Bates, E., Saygın, A. P., Moineau, S., Marangolo, P., & Pizzamiglio, L. (2005). Analyzing aphasia data in a multidimensional symptom space. *Brain and Language*, *92*(2), 106–116. https://doi.org/10.1016/j.bandl.2004.06.108

Beeson, P. M., & Bayles, K. A. (1997). Aphasia. In Paul David Nussbaum (Ed), *Handbook of neuropsychology and aging* (pp. 298–314). Springer.

Ben Shalom, D., & Poeppel, D. (2008). Functional anatomic models of language: Assembling the pieces. *The Neuroscientist*, *14*(1), 119–127. https://doi.org/10.1177/1073858407305726

Bird, H., & Franklin, S. (1996). Cinderella revisited: A comparison of fluent and non-fluent aphasic speech. *Journal of Neurolinguistics*, *9*(3), 187–206. https://doi.org/10.1016/0911-6044(96)00006-1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

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

Caplan, D. (2003). Aphasic syndromes. In K. M. Heilman and E. Valenstein (Eds.), *Clinical Neuropsychology*, 14–34. Oxford University Press. https://doi.org/10.1046/j.1468-1331.2003.00655.x

Caramazza, A., Capitani, E., Rey, A., & Berndt, R. S. (2001). Agrammatic Broca's aphasia is not associated with a single pattern of comprehension performance. *Brain and Language*, *76*(2), 158–184. https://doi.org/10.1006/brln.1999.2275

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology*, *28*(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Cho-Reyes, S., & Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology*, *26*(10), 1250–1277. https://doi.org/10.1080/02687038.2012.693584

Clough, S., & Gordon, J. K. (2020). Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology*, *34*(5), 515–539. https://doi.org/10.1080/02687038.2020.1727709

Covington, M. A. (2007). *MATTR user manual (CASPR research report 2007–05)*. http://ai1.ai.uga.edu/caspr/MATTR-Manual.pdf

Craig, H. K., Hinckley, J. J., Winkelseth, M., Carry, L., Walley, J., Bardach, L., Higman, B., Hilfinger, P., Schall, C., & Sheimo, D. (1993). Quantifying connected speech samples of adults with chronic aphasia. *Aphasiology*, *7*(2), 155–163. https://doi.org/10.1080/02687039308249503

Crary, M. A., Wertz, R. T., & Deal, J. L. (1992). Classifying aphasias: Cluster analysis of Western aphasia battery and Boston diagnostic aphasia examination results. *Aphasiology*, *6*(1), 29–36. https://doi.org/10.1080/02687039208248575

Dalton, S. G. H., & Richardson, J. D. (2019). A large-scale comparison of main concept production between persons with aphasia and persons without brain injury. *American Journal of Speech-Language Pathology*, *28*(1S), 293–320. https://doi.org/10.1044/2018_AJSLP-17-0166

DeDe, G., & Salis, C. (2020). Temporal and episodic analyses of the story of Cinderella in latent aphasia. *American Journal of Speech-Language Pathology*, *29*(1S), 449–462. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0210

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review*, *108*(4), 759. https://doi.org/10.1037/0033-295X.108.4.759

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology*, *32*(4), 487–492. https://doi.org/10.1080/02687038.2017.1398814

Drai, D., & Grodzinsky, Y. (2006). A new empirical angle on the variability debate: Quantitative neurosyntactic analyses of a large data set from Broca's Aphasia. *Brain and Language*, *96*(2), 117–128. https://doi.org/10.1016/j.bandl.2004.10.016

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)

Ferro, J. M., & Kertesz, A. (1987). Comparative classification of aphasic disorders. *Journal of clinical and experimental Neuropsychology*, *9*(4), 365–375.

Forbes, M. M., Fromm, D., & MacWhinney, B. (2012, August). AphasiaBank: A resource for clinicians. In *Seminars in Speech and Language* (Vol. 33, No. 03, pp. 217–222). Thieme Medical Publishers.

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

Fridriksson, J., Fillmore, P., Guo, D., & Rorden, C. (2015). Chronic Broca's aphasia is caused by damage to Broca's and Wernicke's areas. Cerebral Cortex, 25(12), 4689–4696.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., & Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, *55*, 122–129. https://doi.org/10.1016/j.cortex.2013.05.008

Goodglass, H., & Kaplan, E. (1983). *Boston diagnostic aphasia examination* (*2nd* ed.). Lea & Febiger.

Gordon, J. K. (1998). The fluency dimension in aphasia. *Aphasiology*, *12*(7–8), 673–688. https://doi.org/10.1080/02687039808249565

Gordon, J. K. (2020). Factor analysis of spontaneous speech in aphasia. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4127–4147. https://doi.org/10.1044/2020_JSLHR-20-00340

Haley, K. L., Cunningham, K. T., Jacks, A., Richardson, J. D., Harmon, T., & Turkeltaub, P. E. (2021). Repeated word production is inconsistent in both aphasia and apraxia of speech. *Aphasiology*, *35* (4), 518–538. https://doi.org/10.1080/02687038.2020.1727837

Hillis, A. E. (2007). Aphasia progress in the last quarter of a century. *Neurology*, *69*(2), 200–213. https://doi.org/10.1212/01.wnl.0000265600.69385.6f

Huang, R. (2016). *An evaluation of POS taggers for the CHILDES corpus* [Unpublished dissertation]. City University of New York. Retrieved March 15, 2021 from https://academicworks.cuny.edu/gc_etds/1577/

Huber, W. Poeck, K. and Willmes, K. 1984, The Aachen aphasia test. *Advances in Neurology*, *42*, 291–303.

Jacobs, B. J. (2001). Social validity of changes in informativeness and efficiency of aphasic discourse following linguistic specific treatment (LST). *Brain and Language*, *78*(1), 115–127. https://doi.org/10.1006/brln.2001.2452

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Jantzen, J., Axer, H., & von Keyserlingk, D. G. (2002). Diagnosis of aphasia using neural and fuzzy techniques. In Advances in computational intelligence and learning (pp. 461–474). Springer, Dordrecht

John, A. A., Javali, M., Mahale, R., Mehta, A., Acharya, P. T., & Srinivasa, R. (2017). Clinical impression and Western aphasia battery classification of aphasia in acute ischemic stroke: Is there a discrepancy? *Journal of Neurosciences in Rural Practice*, *8*(1), 074–078. https://doi.org/10.4103/0976-3147.193531

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test–Second Edition*. Pro-Ed.

Kasselimis, D. S., Simos, P. G., Peppas, C., Evdokimidis, I., & Potagas, C. (2017). The unbridged gap between clinical diagnosis and contemporary research on aphasia: A short discussion on the validity and clinical utility of taxonomic categories. *Brain and Language*, *164*, 63–67. https://doi.org/10.1016/j.bandl.2016.10.005

Kertesz, A. (2007). *Western aphasia battery–revised*. Pearson.

Kintz, S., & Wright, H. H. (2018). Discourse measurement in aphasia research. *Aphasiology*, *32*(4), 472–474.

Lundholm Fors, K., Fraser, K., & Kokkinakis, D. (2018). Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. In Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth (pp. 705–709). IOS Press

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (*3rd* ed.). Erlbaum.

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

Marshall, J. (2010). Classification of aphasia: Are there benefits for practice? *Aphasiology*, *24*(3), 408–412. https://doi.org/10.1080/02687030802553688

Matchin, W., Basilakos, A., Stark, B. C., den Ouden, D. B., Fridriksson, J., & Hickok, G. (2020). Agrammatism and paragrammatism: A cortical double dissociation revealed by lesion-symptom mapping. *Neurobiology of Language*, *1*(2), 208–225. https://doi.org/10.1162/nol_a_00010

McNeil, M. R., & Kimelman, M. D. (2001). Darley and the nature of aphasia: The defining and classifying controversies. *Aphasiology*, *15*(3), 221–229. https://doi.org/10.1080/02687040042000223

Mirman, D., Kraft, A. E., Harvey, D. Y., Brecher, A. R., & Schwartz, M. F. (2019). Mapping articulatory and grammatical subcomponents of fluency deficits in post-stroke aphasia. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(5), 1286–1298. https://doi.org/10.3758/s13415-019-00729-9

Pallickal, M. (2020). Discourse in Wernicke's aphasia. *Aphasiology*, *34*(9), 1138–1163. https://doi.org/10.1080/02687038.2020.1739616

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

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

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.

Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative-analysis of agrammatic production – Procedure and data. *Brain and Language*, *37*(3), 440–479. https://doi.org/10.1016/0093-934X(89)90030-8

Schwartz, M. F. (1984). What the classical aphasia categories can't do for us and why. *Brain and Language*, *21*(1), 3–8. https://doi.org/10.1016/0093-934X(84)90031-2

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology*, *21*(3–4), 375–393. https://doi.org/10.1080/02687030600911435

Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A. E., Roberts, A., den Ouden, D. B., Brock, K., McKinney-Bock, K., Paek, E. J., Harmon, T. G., Yoon, S. O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S., & Sharma, S. (2021). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology*, *30*(1S), 491–502. https://doi.org/10.1044/2020_AJSLP-19-00093

Stark, B. C., & Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience*, *36*(5), 562–585.

Stark, J. A. (2010). Content analysis of the fairy tale Cinderella–A longitudinal single-case study of narrative production: "From rags to riches". *Aphasiology*, *24*(6–8), 709–724. https://doi.org/10.1080/02687030903524729

Strand, E. A., Duffy, J. R., Clark, H. M., & Josephs, K. (2014). The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech. *Journal of Communication Disorders*, *51*, 43–50. https://doi.org/10.1016/j.jcomdis.2014.06.008

Sundet, K., & Engvik, H. (1985). The validity of aphasic subtypes. Scandinavian Journal of Psychology, 26(1), 219–226

Swindell, C. S., Holland, A. L., & Fromm, D. (1984). Classification of aphasia: WAB type versus clinical impression. In Clinical Aphasiology: Proceedings of the Conference 1984 (pp. 48–54). BRK Publishers

Thompson, C. K., Cho, S., Hsu, C. J., Wieneke, C., Rademaker, A., Weitner, B. B., Mesulam, M. M., & Weintraub, S. (2012). Dissociations between fluency and agrammatism in primary progressive aphasia. *Aphasiology*, *26*(1), 20–43. https://doi.org/10.1080/02687038.2011.584691

Thompson, C. K., Shapiro, L. P., Kiran, S., & Sobecks, J. (2003). The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia. *Journal of Speech, Language, and Hearing Research*, *46*(3), 591–607. https://doi.org/10.1044/1092-4388(2003/047)

Thompson, C. K., Shapiro, L. P., Li, L., & Schendel, L. (1995). Analysis of verbs and verb-argument structure: A method for quantification of aphasic language production. *Clinical Aphasiology*, *23*, 121–140.

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. Brain and Language, 162, 60–71

Trupe, E. H. (1984). Reliability of rating spontaneous speech in the Western aphasia battery: Implications for classification. In *Clinical Aphasiology: Proceedings of the Conference 1984* (pp. 55–69). Seabrook Island, SC: BRK Publishers.

Van der Merwe, A. (2007). Self-correction in apraxia of speech: The effect of treatment. *Aphasiology*, *21*(6–8), 658–669. https://doi.org/10.1080/02687030701192174

Webster, J., Franklin, S., & Howard, D. (2007). An analysis of thematic and phrasal structure in people with aphasia: What more can we learn from the story of Cinderella? *Journal of Neurolinguistics*, *20* (5), 363–394. https://doi.org/10.1016/j.jneuroling.2007.02.002

Wertz, R. T., Deal, J. L., & Robinson, A. J. (1984). Classifying the aphasias: A comparison of the Boston diagnostic aphasia examination and the Western aphasia battery. In *Clinical Aphasiology: Proceedings of the Conference 1984* (pp. 40–47). Seabrook Island, SC: BRK Publishers.

Wilson, S. M., & Hula, W. D. (2019). Multivariate approaches to understanding aphasia and its neural substrates. *Current Neurology and Neuroscience Reports*, *19*(8), 53. https://doi.org/10.1007/s11910-019-0971-6