JSLHR

Research Article

Factor Analysis of Spontaneous Speech in Aphasia

Jean K. Gordon[a]

Purpose: Spontaneous speech tasks are critically important for characterizing spoken language production deficits in aphasia and for assessing the impact of therapy. The utility of such tasks arises from the complex interaction of linguistic demands (word retrieval, sentence formulation, articulation). However, this complexity also makes spontaneous speech hugely variable and difficult to assess. The current study aimed to simplify the problem by identifying latent factors underlying performance in spontaneous speech in aphasia. The ecological validity of the factors was examined by examining how well the factor structures corresponded to traditionally defined aphasia subtypes.
Method: A factor analysis was conducted on 17 microlinguistic measures of narratives from 274 individuals with aphasia in AphasiaBank. The resulting factor scores were compared across aphasia subtypes. Supervised (linear discriminant analysis) and unsupervised (latent profile analysis) classification techniques were then conducted on the factor scores and the solutions compared to traditional aphasia subtypes.
Results: Six factors were identified. Two reflected aspects of fluency, one at the phrase level (Phrase Building) and one at the narrative level (Narrative Productivity). Two other factors reflected the accuracy of productions, one at the word level (Semantic Anomaly) and one at the utterance level (Grammatical Error). The other two factors reflected the complexity of sentence structures (Grammatical Complexity) and the use of repair behaviors (Repair), respectively. Linear discriminant analyses showed that only about two thirds of speakers were classified correctly and that misclassifications were similar to disagreements between clinical diagnoses. The most accurately diagnosed syndromes were the largest groups—Broca's and anomic aphasia. The latent profile analysis also generated profiles similar to Broca's and anomic aphasia but separated some subtypes according to severity.
Conclusions: The factor solution and the classification analyses reflected broad patterns of spontaneous speech performance in a large and representative sample of individuals with aphasia. However, such data-driven approaches present a simplified picture of aphasia patterns, much as traditional syndrome categories do. To ensure ecological validity, a hybrid approach is recommended, balancing population-level analyses with examination of performance at the level of theoretically specified subgroups or individuals.
Supplemental Material: https://doi.org/10.23641/asha.13232354

"Spontaneous speech" refers to spoken language production of a message intended and generated by the speaker. In contrast to repetition and oral reading, spontaneous speech tasks require the speaker to encode the conceptual semantics of the message into linguistic form. Unlike picture-naming tasks, the message to be encoded is propositional. Although the stimulus and structure of the task may vary (e.g., conversation, picture description, procedural narrative, story retelling), spontaneous speech tasks typically involve extended discourse, that is, the production of multiple ideas that are logically connected in some way.

In clinical aphasiology, the analysis of spontaneous speech serves two important and related purposes. First, because spontaneous speech involves generating, formulating, and articulating linguistic messages, it provides the opportunity to investigate multiple levels of linguistic encoding (e.g., lexical retrieval, syntactic formulation, phonological encoding) simultaneously. When combined with more structured testing, spontaneous speech tasks provide valuable diagnostic information about the nature of the communication disorder and its underlying impairments. This is essential to develop appropriate goals and therapy tasks. Second, because spontaneous speech tasks have high ecological validity, they can illustrate how aphasia manifests in everyday language use. Thus, such tasks are the gold standard to assess the impact of aphasia on communicative

[a]Department of Communication Sciences and Disorders, University of Iowa, Iowa City

Correspondence to Jean K. Gordon: jean-k-gordon@uiowa.edu

Editor-in-Chief: Stephen M. Camarata
Editor: Julius Fridriksson

Received June 14, 2020
Revision received August 15, 2020
Accepted September 1, 2020
https://doi.org/10.1044/2020_JSLHR-20-00340

Disclosure: The author has declared that no competing interests existed at the time of publication.

function and the effectiveness of therapy on a person with aphasia's (PWA) life participation.

The complexity of spontaneous speech, which involves multiple linguistic processes and their interactions, generates a vast range of potential patterns of breakdown. Certainly, aphasia presentations are hugely variable; however, systematicities are observed, particularly in patterns of spontaneous speech. Attempting to capture these systematicities has been a long-standing challenge in aphasiology. At the coarsest level, fluent and nonfluent groups can be distinguished, at least on average performance. This distinction is at the heart of traditional models of aphasia diagnosis, such as those instantiated in the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001b), the Western Aphasia Battery (WAB; Kertesz, 1982), and the Aphasia Diagnostic Profiles (Helm-Estabrooks, 1992). However, even this broad distinction is not very reliable (Clough & Gordon, 2020; Gordon, 1998; Trupe, 1984). The traditional approach has often been criticized for its lack of reliability, both in terms of behavioral profiles (e.g., Copland et al., 2018; Howard et al., 2010) and the corresponding neuroanatomy (Halai et al., 2017; Tremblay & Dick, 2016).

A more objective approach than fitting patterns of performance into predefined categories may be to use multivariate methods to derive patterns from the data (Wilson & Hula, 2019). Factor and cluster analytic approaches have a long history in aphasiology, largely coincident with the development and description of comprehensive aphasia batteries (e.g., Goodglass & Kaplan, 1983; Hanson et al., 1982; Jones & Wepman, 1961; Kertesz & Phipps, 1977; Powell et al., 1979; Schuell et al., 1962). Factor analyses on aphasia batteries tend to generate modality-specific factors, with spoken language abilities usually fractionating into two to four factors. For example, Goodglass and Kaplan (1983) derived seven factors from the performance of 242 PWAs on the Boston Diagnostic Aphasia Examination, which were identified as auditory comprehension, reading, writing, repetition–recitation, fluency, nonverbal oral agility, and paraphasia. A more recent analysis of 355 PWAs (Fong et al., 2019) identified five factors corresponding to auditory comprehension/praxis, naming/reading, articulation/repetition, grammatical comprehension, and phonological processing. As noted by Fong et al. (2019), differences arose in part because of variability in the subtests included—all the writing subtests had to be excluded from their own study because of missing data. Examples from other tests echo this point. Hanson et al. (1982) tested 118 PWAs on the 18 subtests of the Porch Index of Communicative Ability (Porch, 1967) and found five factors that corresponded closely to the modality of language use: speaking, writing, comprehension, gesturing, and copying. By contrast, Kertesz and Phipps (1977) conducted a principal components analysis on the performance of 142 PWAs based on 10 subtests of the WAB representing only auditory/oral modalities and generated four components: severity (accounting for the vast majority of the variance), fluency/comprehension, repetition/information content, and naming.

More recently, multivariate analyses of aphasia have turned their focus to the correspondence of language functions and their neural substrates. In a series of studies, Lambon Ralph and colleagues (Alyaha et al., 2018, 2020; Butler et al., 2014; Halai et al., 2017) conducted principal components analyses on a large set of language and cognitive measures to identify latent cognitive–linguistic factors characteristic of aphasia that could then be aligned with lesion sites. The set of measures—selected to comprehensively capture input and output phonological and semantic processing—and the participant samples overlapped considerably across studies. The core factors derived were therefore similar as well, reflecting phonological processing, semantic processing, comprehension, and cognition. When measures of connected speech (speech rate, mean length of utterances [MLU], total words) were added to the single-word measures to better capture functional language use (Alyaha et al., 2018; Halai et al., 2017), these generated an additional factor–fluency (see also Lacey et al., 2017). In the most recent study from the Lambon Ralph group, Alyaha et al. (2020) included four measures of spontaneous speech: speech rate, type–token ratio (TTR), and the number and proportions of correct information units (CIUs, Nicholas & Brookshire, 1993)[1] in three types of discourse (narrative, procedural, and descriptive discourse). This resulted in the spontaneous speech measures fractionating into three factors: one representing quantity (mostly word counts), one representing informativeness (CIU proportions), and one representing speech rate (labeled "motor speech"). Thus, even when analyses are restricted to a single modality, results of multivariate analyses such as principle components analysis (PCA) are highly dependent on the set of variables initially selected. When the variables are selected to represent specific dimensions of interest, as they were here, those dimensions are likely to pop out as factors.

A significant limitation of many past factor analytic studies is their small sample sizes. A general rule of thumb is that the subject-to-variable ratio should be at least 5 and ideally closer to 20, with a minimum of 100 subjects (Comrey & Lee, 1992; Gorsuch, 1983). Although large samples can compensate to some degree for low subject-to-variable ratios and vice versa, each of the studies described in the previous paragraph included fewer than 50 participants in the factor analysis and had subject-to-variable ratios of lower than 2, below even the most liberal of criteria (e.g., see Costello & Osborne, 2005; Kline & Barrett, 1983).[2] The derived factors of phonology and semantics/comprehension (Butler et al., 2014; Halai et al., 2017) do seem fairly robust, as they were strongly supported by several component variables; this can compensate to some extent for small samples (Beavers et al., 2013). However, the utility of the spontaneous speech factors

[1]These were erroneously described as "content word" counts in the article, as they included several types of function words (as do CIUs).
[2]Note that Alyaha et al. (2018, 2020) misinterpreted the Kaiser–Meyer–Olin index as a measure of sample size adequacy, rather than sampling adequacy, that is, how well the measures sample their shared variance.

(Alyaha et al., 2018, 2020; Halai et al., 2017; Lacey et al., 2017) is limited by the restricted set of variables analyzed.

Only a handful of studies have focused specifically on the structure of spontaneous speech. Several early studies relied on qualitative ratings of seven to 10 dimensions (e.g., Benson, 1967; Goodglass & Kaplan, 1972; Kerschensteiner et al., 1972) to characterize aphasic production. Like the studies above, however, these examinations largely neglected aspects of grammatical formulation and could be faulted for their subjectivity (Wagenaar et al., 1975). In a more recent example, Casilio et al. (2019) expanded the number of ratings to 27 dimensions, including aspects of lexical retrieval, grammar, morphology, and motor speech for 24 PWAs in a personal narrative task. Twenty-three of these were examined in a factor analysis, deriving four factors: paraphasia, logopenia (paucity of speech), agrammatism, and motor speech. Although the primary purpose of the study was to demonstrate the feasibility of the rating tool, the validity of the factor solution can be questioned on the basis of an extremely small subject-to-variable ratio (essentially 1:1). In addition, a common problem with rating scales is that one rating can influence another, the so-called "halo effect" (Thorndike, 1920). Although the authors were able to achieve good levels of reliability on most of the measures and demonstrated concurrent validity across samples by comparison with objective measures, it is likely that shared variance among the ratings was inflated, as the ratings were considerably more strongly intercorrelated than comparable objective measures (e.g., Clough & Gordon, 2020).

Other studies have taken a quantitative approach to measuring spontaneous speech. Wagenaar et al. (1975) examined responses to standard open-ended personal questions from 74 Dutch-speaking PWAs. Thirty measures were taken, including speech rate, utterance length and complexity, several types of lexical errors, word class counts, and measures of grammatical and morphological accuracy. Six factors were generated: fluency, telegraphic speech, grammatical errors, articulation, verbal paraphasia, and empty speech. Vermeulen et al. (1989) conducted a factor analysis of 17 spontaneous speech variables measured on 121 individuals with aphasia, also responding to open-ended questions. They derived five factors: syntactic ability, phonological paraphasia, neologistic paraphasia, articulatory impairment, and vocabulary. A significant shortcoming of this study was the lack of grammatical measures; the interpretation of Factor 1 as representing syntax was indirectly based on loadings of speech rate, MLU, and function word counts. A limitation of both of these studies is the use of open-ended questions, which can introduce potential confounds, such as differences in content and response length, that may be unrelated to the aphasia.

Some studies have specifically focused on components of fluency, an important overall measure of spontaneous speech adequacy. The Quantitative Production Analysis (Berndt et al., 2000; Saffran et al., 1989) provides an in-depth analysis of grammatical production abilities, specifically for characterizing agrammatism. Rochon et al. (2000) examined the performance of 37 nonfluent PWAs on nine grammatical measures obtained from a story-telling task. A factor analysis yielded two factors accounting for 70% of the total variance: Factor 1 mostly reflected sentence-level measures (use of auxiliary verbs, proportion of words in sentences, sentence elaboration), while Factor 2 reflected narrative-level measures (speech rate and lexical ratios—verbs relative to nouns, pronouns relative to nouns, proportion function words). Whether a similar structure would be obtained on a sample that included fluent types of aphasia is an open question. As a precursor to examining therapy outcomes, Feenaughty et al. (2017) subjected five measures of speech fluency (numbers and rates of syllables and pauses, number of different words) collected from 88 PWAs in a picture description task to a principal components analysis, resulting in two factors. However, the five measures were highly interrelated and included a restricted subset of the sources of variances contributing to fluency. To be fair, the goal of the study was not to identify the underlying structure of fluency.

Identifying substrates of fluency was, however, the explicit goal of three recent studies. Nozari and Faroqi-Shah (2017) used a structural equation model to identify latent variables contributing to fluency of production in 112 PWAs. Fluency was measured using speech rate and the WAB Fluency scale, and predictors included naming tests to tap into lexical retrieval skills, measures of syntax and retracing from story-telling samples, and measures of auditory comprehension and working memory. Although the authors reported indirect effects for the latent constructs of comprehension, naming, and working memory (perhaps reflecting overall severity), only the syntax construct had a direct effect on fluency. A large-scale analysis of story-telling narratives in 254 PWAs (Clough & Gordon, 2020; Gordon & Clough, 2020) illustrated that underlying predictors of fluency depend in part on how fluency is measured. In a logistic regression, binary clinical judgments of fluency (i.e., fluent vs. nonfluent) were predicted most strongly by the presence or absence of apraxia of speech (Clough & Gordon, 2020). In linear regressions, the strongest predictors were overall aphasia severity for WAB Fluency scale scores and grammatical complexity for MLU and speech rate, all common proxies for fluency (Gordon & Clough, 2020). These studies differed from the factor analyses, in that they imposed an outcome measure—in this case, fluency—which must be approximated by an imperfect measure. Factor analysis, by contrast, is designed to reveal the underlying structure in a set of data without predicting a specified outcome.

The Current Study

The first goal of the current study was to identify constructs underlying spontaneous speech production in individuals with aphasia using an exploratory factor analysis. Limitations of previous studies were addressed by examining a comprehensive set of variables measuring the major linguistic domains of lexical retrieval and grammatical formulation, as well as the overall fluency of production, and capturing dimensions of accuracy, complexity, diversity, specificity, efficiency, and productivity. The inclusion of a broad range of variables was an attempt to be as agnostic as possible a priori about the underlying structure, as befits an exploratory analysis. The analysis of many variables requires a large subject sample, which was made possible here by taking advantage of the large and carefully vetted data set in AphasiaBank (MacWhinney, 2000). The second goal of the study was to examine how the identified factors correspond to traditionally defined subtypes of aphasia (e.g., Goodglass et al., 2001a; Helm-Estabrooks, 1992; Huber et al., 1984; Kertesz, 1993) to explore the clinical utility of the factor solution in understanding behavioral profiles in aphasia.

Unlike many previous studies, the current study used a factor analysis approach, rather than a PCA, and an oblique rotation, rather than an orthogonal rotation. These were conceptual and methodological decisions. Although these terms are often used interchangeably, PCA operates to simplify the data structure while accounting for all the variance in the measured variables, whereas factor analysis is designed to identify latent factors by accounting for only common variance among the variables. Factor analysis should therefore be the method of choice when the goal is more theoretical, as PCA conflates common variance and error variance (Costello & Osborne, 2005; Wilson & Hula, 2019). The second issue is that orthogonal rotations assume that the underlying components are uncorrelated, whereas oblique rotations do not. Studies of spontaneous speech in aphasia indicate significant intercorrelations among many measures (e.g., Casilio et al., 2019; Clough & Gordon, 2020), even when they are intended to represent different constructs. Thus, the almost exclusive use of orthogonal rotations in past work noted by Wilson and Hula (2019), while it may be methodologically optimal for studies aiming to identify neural correlates (Alyaha et al., 2020; Butler et al., 2014; Halai et al., 2017), runs the risk of making unwarranted assumptions about the nature of the constructs underlying spontaneous speech measures.

Method

This study was approved by the institutional review board at the University of Iowa and partially supported with funds from the American Speech-Language-Hearing Foundation.

Participants With Aphasia

The set of PWAs included all English-speaking individuals in AphasiaBank who carried out the AphasiaBank protocol and for whom at least three spontaneous verbal utterances were produced in the Cinderella story-telling task. Of the 307 unique PWAs in this set (as of April 2020), eight were excluded because the task was not administered or the PWAs declined to carry out or abandoned the task. Another 24 were excluded because the PWAs produced fewer than three spontaneous verbal utterances. One additional PWA was excluded because English was not her first or dominant language. This resulted in a final sample of 274 PWAs, representing a range of aphasia subtypes, ranging in severity from a WAB Aphasia Quotient (AQ) of 12.8–99.6 (*Mdn* = 75), and ranging in time postonset from 1 month to 30 years (*Mdn* = 4 years).

In the AphasiaBank protocol, aphasia syndromes are determined in two different ways: based on WAB categorization guidelines and on clinical impression. It is well known that aphasia syndrome diagnoses are often unreliable (Clough & Gordon, 2020; Gordon, 1998; Holland et al., 1986; Trupe, 1984). In the current sample, this issue was resolved by consulting both diagnostic labels. The documented aphasia type was used when WAB category agreed with clinical impression (n = 156, 57% of PWAs) or when only one of these was available (n = 27, 10%). An additional 10% (n = 28) were considered not aphasic by the WAB AQ cutoff of AQ > 93.8, (Kertesz, 2006) but anomic by clinical impression, which is a severity distinction rather than a classification discrepancy. For these individuals (dubbed "not aphasic by the WAB" [NABW] by Fromm et al., 2017), we used the NABW category to preserve the distinction in severity from those with more severe anomic aphasia. The remaining 23% (n = 63) with discrepant diagnoses were resolved by my own clinical judgment, consulting the audio samples and test scores available in AphasiaBank, thus providing a consensus of two judgments for almost every speaker.[3] A large proportion (n = 24, 38%) of discrepant diagnoses were considered to have Broca's aphasia by clinical judgment but anomic or conduction aphasia by the WAB profile, and most of these were coded as 5 on the WAB Fluency scale. As discussed elsewhere (see Clough & Gordon, 2020; Trupe, 1984), this is an artifact of the WAB scale; a code of 5 indicates residual nonfluency but is categorized as "fluent," resulting in a diagnosis of one of the fluent syndromes.

Following this process, in addition to the 10% (n = 28) classified as NABW, 32% were classified as having Broca's aphasia (n = 88), 31% with anomic aphasia (n = 85), 15% with conduction aphasia (n = 42), 7% with Wernicke's aphasia (n = 20), and 4% with transcortical motor (TCM) aphasia (n = 11). This sample is fairly representative of the distribution of subtypes of nonacute aphasia, in which Broca's aphasia and anomic aphasia predominate (e.g., Basilakos et al., 2014; Borovsky et al., 2007; Feenaughty et al., 2017; Hoffmann & Chen, 2013).

Task

In the AphasiaBank protocol, discourse samples are elicited in several tasks: personal narratives, descriptions of single pictures and series of pictures, a story-telling narrative (Cinderella), and procedural discourse. For this study, the Cinderella story-telling sample was used because narrative tasks tend to produce longer samples and are highly

[3]All but two (3%) of those resolved by the author were resolved in agreement with one of the existing classifications (57% in agreement with clinical judgment; 40% in favor of WAB subtype). An additional speaker had no subtype diagnoses entered and was diagnosed by the author (but was not counted as a discrepancy).

representative of many real-life discourse tasks, such as recounting a day's events (Olness & Ulatowska, 2011). Unlike personal narratives, however, story-telling narratives —referred to as "semispontaneous" (Prins & Bastiaanse, 2004)—have the advantage of relying on a relatively consistent set of propositions dictated by the initial stimulus, thus making the samples more comparable across individuals. As noted above, only PWAs who produced at least three spontaneous utterances in telling the Cinderella story were included. *Spontaneous utterances* were those produced without cueing by the clinician or with only general prompting (e.g., "What else?"; "Then what happened?"). *Verbal utterances* excluded utterances in which only gestures, unintelligible words, or nonwords were produced. Requiring at least three utterances balanced the desire for measures to be as representative as possible of the individual's spontaneous speech with the desire to include as many speakers as possible, with a broadly representative range of severity.

Variables Analyzed

Variables were selected from available information in AphasiaBank to represent, as thoroughly as possible, the ability of the speakers to retrieve words and combine them into sentences and to capture dimensions of accuracy, diversity, specificity, complexity, efficiency, and productivity. It is important to acknowledge that the analysis focused on microlinguistic measures and did not assess the coherence, cohesion, or story structure of the narratives. As others have noted, the primary linguistic impairments of aphasia occur at this level (Armstrong et al., 2013; Bastiaanse & Prins, 1994; Glosser & Deser, 1991); if macrolinguistic deficits are found, they can often be attributed to co-existing microlinguistic deficits (Andreetta & Marini, 2015; Hazamy & Obermeyer, 2020; Wright & Capilouto, 2012). Following is a list of the variables and how they were measured. All measures were collected using the EVAL command in CLAN (MacWhinney, 2000), except where noted otherwise.

Narrative-Level Measures

Two narrative-level measures were included. *Total utterances* is a measure of how much narrative was produced, without taking into account the appropriateness or efficiency of the content. In addition to indicating the ability to produce extended discourse, the amount of speech produced is an important covariate, as it can affect the production of other measures (Cahana-Amitay & Jenkins, 2018). *Speech rate,* measured in words per minute (WPM), tapped into the efficiency of production and reflects aspects of grammatical formulation, lexical retrieval ability, and motor speech (Gordon & Clough, 2020). By default, EVAL excludes in its word count nonlinguistic fillers, retraced material, fragments, and errors for which a target could not be identified.

Utterance-Level Measures

Seven utterance-level measures captured aspects of syntactic and semantic sufficiency. MLU (measured in words) can be affected by lexical retrieval abilities but mostly reflects grammatical complexity (Gordon & Clough, 2020). *Repairing* reflects the efficiency of combining words into phrases and was measured by calculating the proportion of utterances that contained repetition or retracing (self-corrections or changes). These were combined because it is often impossible to identify the reason for the repair behavior. Two more direct measures of grammatical complexity were also included: The proportion of grammatical relations that mark *embeddings,* as defined in the CLAN manual, was calculated by using FREQ to count these codes from the MOR line and dividing by the number of utterances. The *sentence complexity ratio* from the new C-NNLA function in CLAN (Fromm et al., 2020) calculates the proportion of sentences containing such embeddings. Thus, although the two complexity measures make use of the same counts, they are calculated using different denominators. Grammatical accuracy was measured at the utterance level by counting the proportion of utterances coded as *grammatical errors* [+gram]. Semantic specificity was captured by the proportion of utterances coded as either *circumlocutions* [+cir] *or empty speech* [+es]. As with retraces and repetitions, these two measures were combined because they are often difficult to distinguish, and both reflect word retrieval difficulties. The proportion of utterances consisting of *jargon* [+jar] reflected semantic content, in that jargon is defined by its lack of meaning (in the case of neologistic or nonword jargon) or lack of accurate meaning (in the case of semantic or real-word jargon).

Word-Level Measures

Ten measures at the word level were included to tap into aspects of grammatical formulation, morphology, and accuracy of lexical retrieval. *Content:function word ratio* and *propositional density* are both lexical proportions. A high content:function word ratio is characteristic of agrammatic speech (Gordon, 2006; Saffran et al., 1989), whereas a low content:function word ratio is characteristic of empty speech (Edwards, 2005; Gordon, 2006). Propositional density reflects the ability to combine nouns with words that form ideas or units of meaning (verbs, adjectives, adverbs, prepositions, and conjunctions) and is measured as the proportion of total words consisting of these word classes (following Covington, 2007; Snowden et al., 1996). Morphological complexity was measured by including from the C-NNLA program the proportion of verbs that were inflected for number and/or tense markers (*verb marking*) and the proportion of nouns that were marked for number and possessive markers (*noun marking*). Morphological accuracy was captured by summing all of CLAN's *morphological error* codes [* m:a] as a proportion of total words.

Lexical diversity was captured using the moving average type–token ratio (MATTR; Covington & McFall, 2008) in CLAN. A moving average window compensates for variability in sample length by calculating TTR over successive windows, or blocks of text, of a given size and averaging them. In order to capture the scope of lexical diversity appropriate to each speaker, MATTR was calculated using up to three different window sizes depending on sample length (5, 10, and 20 words) and averaging the ratios. Using a small window size ensured that TTR could be calculated for all speakers (the minimum number of words produced was 6). However, a window size of 5 largely reflects word repetition within an utterance (the average MLU for the speakers was 6.2 words). Thus, for speakers who produced longer samples, measuring TTR over larger windows provided an index of lexical repetition across successive utterances in the narrative. Averaging TTR across windows provided one index that could be compared across all speakers. Lexical accuracy was captured using four error codes, calculated as a proportion of total words: *phonologically related errors* [* p]; *semantically related errors* [* s:r]; *unrelated word errors*, that is, real-word errors that were unrelated to the target [* s:ur], perseverated [* s:per], or for which the target was unknown [* s:uk]; and *neologistic errors* [* n].

Data Analysis

After compiling the data, values for each measure were *z*-score transformed across the sample of PWAs to put them on the same scale. Prior to carrying out the factor analysis, the "factorability" of the data was assessed using functions and guidelines from Field et al. (2012) and the *psych* package in R (Revelle, 2019, 2020). First, the intercorrelations among variables were examined for redundant variables (those with any correlations >.90) and for "unique" variables (those with no meaningful intercorrelations—here, we used .20 as a minimum). The intercorrelation table is shown in Supplemental Figure S1. Two of the variables (*semantic errors* and *noun marking*) had no intercorrelations above the .20 criterion, indicating that they were not appropriate for the factor analysis. (In preliminary factor analyses, these items were also found to have very high uniqueness values, indicating a lack of shared variance.) Thus, these two variables were deleted from the set. Next, we calculated the determinant of the intercorrelation table, which checks for multicollinearity. After removing the two variables, the determinant was .003, exceeding the minimum criterion of .0001 (Field et al., 2012). The Kaiser–Meyer–Olin test showed an overall measure of sampling adequacy of .73, considered a "middling" result that nevertheless exceeds the minimum recommended value of .60 (Kaiser & Rice, 1974). Bartlett's test of sphericity was significant (χ^2 = 1522, *df* = 136, *p* < .001), indicating that the variables were sufficiently intercorrelated to warrant a factor analysis.

Having met these criteria for factorability, an exploratory factor analysis was conducted on the 17 remaining measures using the *fa* program in R (R Development Core Team, 2020, Version 3.6.3), and selecting the maximum likelihood estimation method to generate factors. Both orthogonal (*varimax*) and oblique (*promax*) rotations were compared, as recommended (Tabachnik & Fidell, 2007). Both methods resulted in the same factor interpretations, although there were differences in the loadings and the order of importance of the factors (i.e., how much variance they accounted for). However, because the oblique factors showed significant intercorrelations, the orthogonal rotation was rejected, and the *promax* rotation is reported below.

A parallel analysis (*fa.parallel*) was also conducted to help determine the number of factors to retain, in combination with the interpretability of the factors (Pohlmann, 2004; Revelle, 2020; Wilson & Hula, 2019). Once the number of factors was determined, the factor loadings were interpreted. Pattern loadings represent the standardized partial regression coefficients of each factor predicting each variable that is controlling for the influence of the other factors. Structure loadings represent simple (zero-order) correlations between each factor and variable, *not* controlling for the influence of the other factors. With orthogonal rotations, pattern and structure loadings are the same; however, with oblique rotations, these matrices generate different loadings. Because they illustrate the unique relationships between factors and variables, pattern loadings are generally considered most useful in interpreting the identity or meaning of each factor. However, because the factors are intercorrelated, it is also helpful to refer to the structure loading matrix. Thus, both are examined below. Recommendations vary on the size of factor loading to interpret, with minimum values hovering around 0.3 or 0.4 (e.g., Beavers et al., 2013). In the current study, for consistency of interpretation, values below 0.2 are not discussed, values between 0.2 and 0.4 are interpreted as weak, values between 0.4 and 0.7 are considered to be moderate, and values over 0.7 are considered to be strong.

Factor scores were then generated for each individual speaker, using the regression method. Factor scores can be generated using either the pattern loadings (default for the *factanal* program in *R*) or the structure loadings (default for the *fa* program). Both were examined in the current study. The factor scores were then used in three subsequent analyses to assess their validity. First, to compare factor scores across different aphasia subtypes, one-way analyses of variance (ANOVAs) were conducted for each factor, correcting *p* values for the number of factors assessed. The ANOVAs were first conducted with pattern matrix factor scores and then repeated using structure matrix factor scores. For significant main effects, post hoc (Bonferroni) tests were conducted to identify specific differences between aphasia subtypes. Second, to examine how well the identified factors mapped onto traditional aphasia subtypes, two linear discriminant (LD) analyses were conducted using the *MASS* library in *R*. An LD analysis calculates linear combinations of predictors that best discriminate among a set of provided categories (in this case, aphasia subtypes). Aphasia subtype classification accuracy predicted from factor structure scores (first LD analysis) was compared to classification accuracy from the original variables (second LD analysis). Third, to determine how the speakers would be classified without reference to predefined categories, an unsupervised classification method—latent profile analysis (LPA)—was conducted using the *mclust* function in the *tidyLPA* package (Rosenberg et al., 2018). The generated profiles were compared to traditional aphasia subtype classifications.

Results

Figure 1 shows the *z* scores for each of the original variables, averaged across the six aphasia subtypes in the sample. The figure illustrates the data that went into the factor analysis and is referred to later to help interpret the findings. Corresponding mean values and standard deviations for each subtype are available in Supplemental Table S1. As mentioned above, intercorrelations among the variables are illustrated graphically in Supplemental Figure S1.

Factor Analysis

The parallel analysis indicated that the eigenvalues for six factors exceeded the random eigenvalues generated by simulated or resampled data (see scree plot in Supplemental Figure S2). The loadings for each factor in the six-factor solution are shown in Table 1, along with their respective sums of squared loadings, and the proportion of variance accounted for by each. The communalities (h2) of each variable, that is, the proportion of variance in that variable accounted for by all the factors, are shown in the next to last column. In the final column is the complexity of each variable, that is, the extent to which it cross-loads on more than one factor.

According to the pattern matrix, Factor 1 (*Grammatical Complexity*) accounted for 13% of the variance, excluding variance shared with other factors. This factor shows strong loadings on the embedding index (0.85) and the complexity ratio (0.93), a moderate loading on utterance length (0.55), and weak loadings on speech rate (0.29) and repairs (0.20). This is a relatively straightforward factor that reflects the ability to produce grammatically complex structures and produce speech relatively quickly. The positive loading of repairs, although weak, suggests that doing so occasionally results in errors or false starts that require repairs.

Factor 2 (*Phrase Building*) accounted for 10% of the variance. This factor showed a strong loading of propositional density (0.94) as well as moderate loadings of verb marking (0.55) and lexical diversity (0.52). Weak positive loadings were found for utterance length (0.25) and grammatical errors, and weak negative loadings for neologistic (−0.30) and phonological (−0.31) errors. This combination of measures appears to reflect phrase-level formulation, that is, the ability to combine words into phrases promoted by both lexical retrieval (MATTR) and grammatical formulation abilities (verb marking). Propositional density appears to reflect both of these, since it is a lexical measure but counts words used to formulate phrasal units (e.g., verbs, adjectives, prepositions). Similarly, the role of lexical diversity here might arise simply from the ability to retrieve words of different grammatical classes, rather than a high diversity of content words. As for Factor 1, an unexpected positive loading is noted; the weak positive loading on grammatical errors suggests that phrase-building entails a cost in some individuals. The distinction between Grammatical Complexity (Factor 1) and Phrase Building suggests that Factor 1 is capturing a higher level of grammatical formulation reflecting syntactic skills, whereas Phrase Building reflects skills needed to produce basic word combinations—lexical retrieval and verb inflection.

Factor 3 (*Semantic Anomaly,* accounting for another 10% of the variance) clearly reflects the integrity of lexical semantics, with a large loading on jargon (1.06), as well

Figure 1. Mean *z* scores of original spontaneous speech variables for each aphasia subtype. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; WER = Wernicke's aphasia; TCM = transcortical motor aphasia; NABW = not aphasic by the Western Aphasia Battery; TotalUtts = number of total utterances; WPM = speech rate in words per minute; MLU = mean length of utterances in words; PropDens = propositional density; ConFun = content:function word ratio; Repair = % utterances with repetition or retracing; MorphErr = % utterances with morphological errors; NeoErr = % neologistic errors of total words; PhonErr = % phonologically related words of total words; SemErr = % semantically related words of total words; URWdErr = % unrelated word errors of total words; MATTR = moving average type–token ratio; CircEs = % utterances with circumlocution or empty speech; GrammErr = % utterances with grammatical errors; Jar = % utterances containing jargon; ComplexGram = % grammatically complex utterances; NounMarking = % nouns inflected; VbMarking = % verbs inflected; ComplexRatio = % grammatically complex sentences.

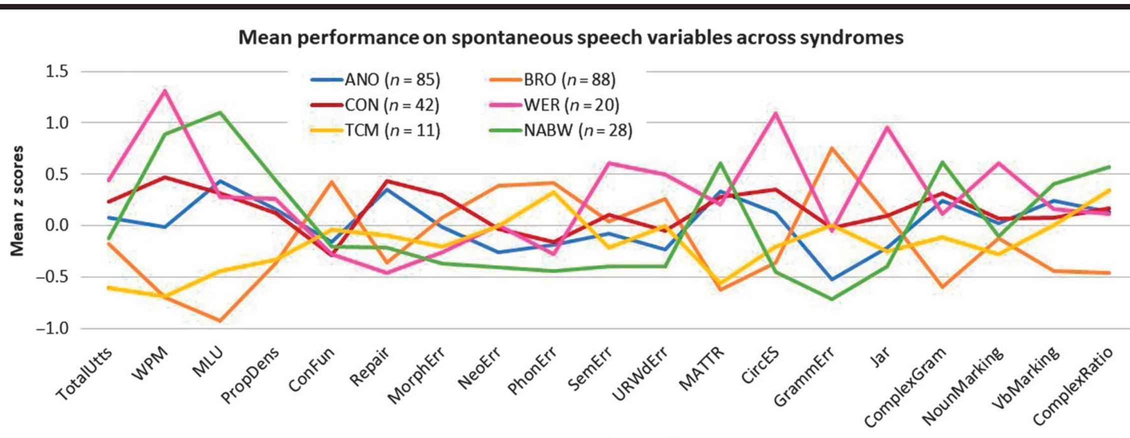


Table 1. Factor pattern matrix loadings.

Variable	Factor 1 Gram Comp	Factor 2 Phrase Bldg	Factor 3 Sem Anom	Factor 4 Gram Err	Factor 5 Narr Prod	Factor 6 Repair	h2	com
Factor pattern matrix (partial regression coefficients)								
Total Utts					**0.47**		0.17	1.3
WPM	0.29				**0.91**	**–0.43**	1.00	1.6
MLU	**0.55**	0.25					0.77	1.7
Prop Dens		**0.94**		0.37			0.58	1.3
Con:Fun		–0.26		0.26	–0.20		0.32	4.3
Repair	0.20					**0.59**	0.37	1.3
Morph Err				0.31		0.22	0.14	2.1
Neo Err		–0.30	**0.46**				0.45	1.9
Phon Err		–0.31					0.23	1.7
UR Wd Err			**0.64**				0.49	1.1
MATTR		**0.52**					0.41	1.3
Circ/ES					0.29		0.14	1.8
Gramm Err		0.30		**1.14**			1.00	1.2
Jargon			**1.06**				1.00	1.1
Embedding	**0.85**						0.77	1.1
Verb Marking		**0.55**					0.26	1.2
Complex Ratio	**0.93**						0.71	1.1
Importance of factors (sum of squared loadings; variance accounted for)								
SS loadings	2.21	1.65	1.75	1.36	1.21	0.61		
Proportion Var	0.13	0.10	0.10	0.08	0.07	0.04		
Cumulative Var	0.13	0.23	0.33	0.41	0.48	0.52		

Note. Only loadings > 0.2 are shown; loadings > 0.4 are shown in bold. The pattern matrix table shows regression coefficients, which, in the case of oblique rotations, are not equivalent to correlations and therefore are not constrained to be less than 1. Gram Comp = Grammatical Complexity; Phrase Bldg = Phrase Building; Sem Anom = Semantic Anomaly; Gram Err = Grammatical Errors; Narr Prod = Narrative Productivity; h2 = communalities; com = variable complexity (degree of cross-loading). Please see Figure 1 or Supplemental Table S1 for variable acronyms. See Supplemental Table S2 for complete pattern matrix.

as moderate loadings on unrelated word errors (0.64) and neologisms (0.46). What these errors have in common is their lack of meaning, in contrast to semantically *related* errors (which, as discussed above, were excluded after preliminary analyses showed that they did not contribute to any factors). The label "semantic anomaly" rather than "semantic error" reflects this finding.

Factor 4 (*Grammatical Error,* 8% of the variance) largely affects the proportion of ungrammatical utterances (1.14), with weaker loadings on morphological errors (0.31), propositional density (0.37), and content:function word ratio (0.26). The combination of these variables—with the exception of propositional density—is characteristic of telegraphic, agrammatic production. The role of propositional density here is somewhat surprising, since agrammatism is characterized by excessive reliance on content words, particularly nouns, which are not included at all in propositional density counts. The association of grammatical errors with propositional density was also a feature of Factor 2 (Phrase Building). The cross-loading of these two variables on both factors indicates that the variables themselves may be multidimensional. For example, different patterns of grammatical errors may be conflated. This issue is revisited below.

Factor 5 (*Narrative Productivity,* accounting for 7% of the variance) loads heavily on speech rate (WPM, 0.91), with a moderate loading on total utterances (0.47), a weak positive loading on circumlocution/empty speech (0.29), and a weak negative loading on content:function word ratio (−0.20). This factor appears to reflect narrative-level fluency or productivity, that is, the ability to produce more words overall, relatively quickly. The coloading of circumlocution/empty speech here suggests that the ability to maintain flow of speech may be achieved at the expense of content specificity in at least some of the speakers who score highly on this factor.

Factor 6 (*Repairs,* 4% of the variance) showed a moderate positive loading on the repair measure (0.59), a moderate negative loading on speech rate (−0.43), and a weak loading of morphological errors (0.22). The interpretation of this factor is quite straightforward, indicating that repair behaviors result in slower speech production and that repairs are associated, in part, with morphological errors. However, the finding that the factor only accounted for 4% of the variance and affected only three variables calls its importance into question. An alternative five-factor solution resulted in a more parsimonious but less interpretable model, with repairs loading on the Phrase Construction factor. Thus, I opted to retain the six-factor solution indicated by the parallel analysis.

Together, these factors accounted for 52% of the variance in spontaneous speech scores. This is the sum of the unique proportions of variance accounted for by each factor, as shown in Table 1. However, Table 2—which shows the structure matrix loadings, the sums of squared loadings and total variance accounted for by each factor, and the intercorrelations among the factors—illustrates that there is

Table 2. Factor structure matrix loadings.

Variable	Gram Comp	Phrase Bldg	Sem Anom	Gram Err	Narr Prod	Repair
Total Utts					0.37	
WPM	**0.59**	**0.47**		–0.39	**0.86**	
MLU	**0.79**	**0.69**	–0.32	**–0.55**	**0.53**	0.20
Prop Dens	0.33	**0.70**	–0.31		0.29	
Con:Fun		**–0.44**		**0.46**	**–0.41**	–0.26
Repair	0.20	0.25			0.20	**0.58**
Morph Err				0.28		0.20
Neo Err	–0.24	**–0.52**	**0.57**	0.27	–0.31	–0.25
Phon Err	–0.25	**–0.45**		0.33	–0.38	
UR Wd Err	–0.20	–0.31	**0.68**	0.06	–0.23	–0.28
MATTR	**0.43**	**0.61**	–0.31	–0.36	0.30	
Circ/ES	0.24				0.34	
Gramm Err	–0.33	**–0.41**		**0.97**	–0.37	
Jargon			**0.96**			
Embedding	**0.86**	**0.49**	–0.25	–0.34	0.35	0.21
Verb Marking		**0.48**		–0.30		
Complex Ratio	**0.82**	0.31		–0.25		
Importance of factors (sum of squared loadings; variance accounted for)						
SS loadings	3.12	3.19	2.20	2.37	2.27	0.84
Proportion Var	0.18	0.19	0.13	0.14	0.13	0.05
Factor score intercorrelations						
Gram Comp		0.49	–0.25	–0.39	0.37	0.03
Phrase Bldg			–0.37	–0.58	0.52	0.27
Sem Anom				0.02	–0.18	–0.20
Gram Err					–0.44	–0.13
Narr Prod						0.31

Note. Only loadings > 0.2 are shown; loadings > 0.4 are shown in bold. Cumulative variance accounted for is not shown in this table, as the proportions of variance accounted for by each factor overlap with each other and therefore cannot be summed. Gram Comp = Grammatical Complexity; Phrase Bldg = Phrase Building; Sem Anom = Semantic Anomaly; Gram Err = Grammatical Errors; Narr Prod = Narrative Productivity. Please see Figure 1 or Supplemental Table S1 for variable acronyms. See Supplemental Table S3 for complete structure matrix.

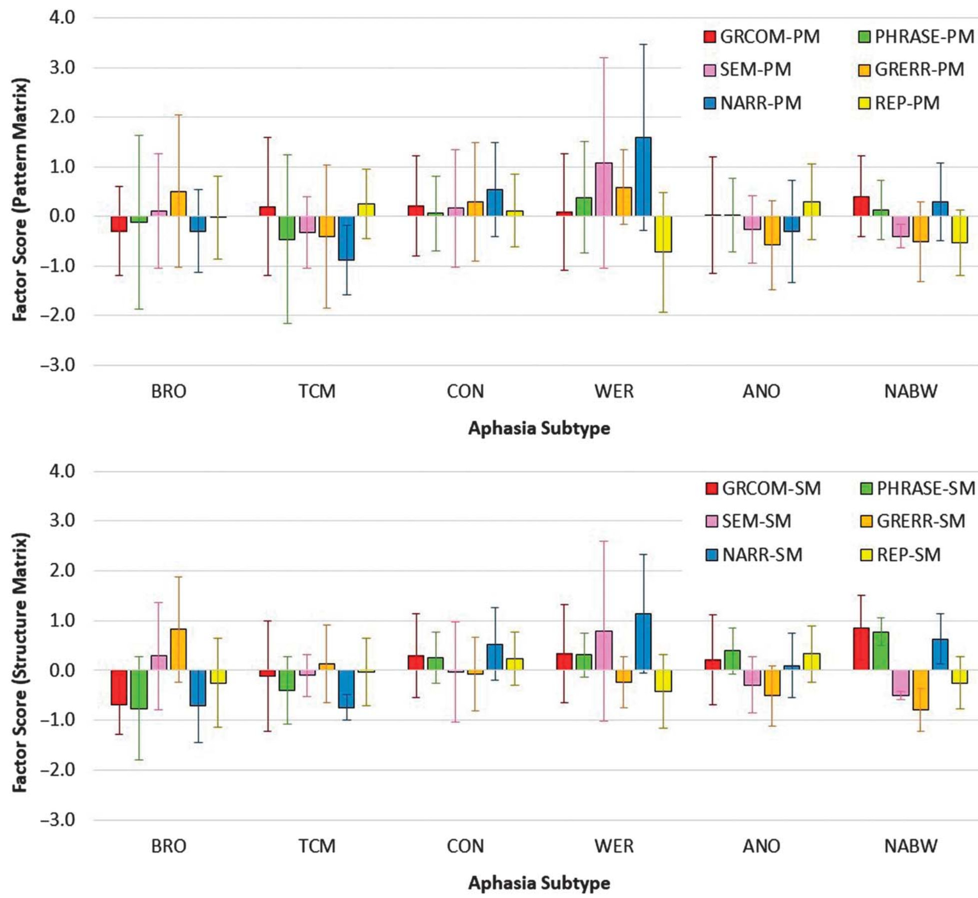
considerable overlap in the variance accounted for by the factors. In particular, Factors 1 (Grammatical Complexity) and 2 (Phrase Building) showed a strong positive correlation of $r = .49$, and both were negatively correlated with Factor 4 (Grammatical Error): $r = -.39$ (Factor 1) and $r = -.58$ (Factor 2). Factor 5 (Narrative Productivity) was also strongly associated with Factor 2 ($r = .52$). The relatedness of the factors can also be seen in the greater number of cross-loadings. For example, the pattern matrix suggests that only Factor 1 contributes to utterance length; however, the structure matrix shows moderate to strong loadings of four of the six factors (Grammatical Complexity, Phrase Building, Grammatical Error, and Narrative Productivity) on MLU and weak loadings on both of the others. The structure matrix is also helpful in understanding the interactions among the variables and the factors. For example, repair behaviors are positively associated with Grammatical Complexity, Phrase Building, and Narrative Productivity, which seems counterintuitive. These are weak loadings, but the consistently positive association provides a rationale for further investigation, which is pursued below in the comparison of aphasia syndromes.

Aphasia Syndrome Comparison

To examine how well these factors mapped onto aphasia syndromes, a set of factor scores were first generated (weighted estimates for each speaker, based on their individual performance on the measured variables and the factor variable loadings). Figure 2 shows the profiles of factor scores that were calculated for each speaker and then averaged across aphasia subtypes. In the top graph, factor scores were calculated from the pattern matrix; in the bottom graph, factor scores were calculated from the structure matrix. The structure matrix factor scores appear to do a better job of discriminating the subtypes, as indicated by smaller error bars and—to some extent—more distinct profiles. This is likely because the structure matrix factor scores take into account the relationships between factors that are present in actual performance indicators (recall that pattern matrix loadings factor out the contributions of other factors in predicting variables).

Comparing structure matrix factor scores for each aphasia subtype still reflects a high degree of variability within subtypes, but with some expected patterns differentiating them. Speakers with Broca's aphasia showed lower-than-average scores on the Grammatical Complexity, Phrase Building, and Narrative Productivity factors, accompanied by higher-than-average Grammatical Error scores, indicating difficulty building grammatical units. Phrase Building and Narrative Productivity were also relatively low in TCM aphasia, but without the Grammatical Errors that characterized Broca's aphasia. By contrast, those with Wernicke's aphasia showed high Narrative Productivity

Figure 2. Mean factor scores for each aphasia subtype. Top: Factor scores calculated from pattern matrix loadings. Bottom: Factor scores calculated from structure matrix loadings. Error bars indicate standard deviations. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; WER = Wernicke's aphasia; TCM = transcortical motor aphasia; NABW = not aphasic by the Western Aphasia Battery; GRCOM = Grammatical Complexity factor; PHRASE = Phrase Building factor; SEM = Semantic Anomaly factor; GRERR = Grammatical Error factor; NARR = Narrative Productivity factor; REP = Repair factor; PM = Pattern Matrix scores; SM = Structure Matrix scores.



and Semantic Anomaly scores, reflecting the pattern of copious but errorful speech typical of this syndrome. Anomic aphasia was characterized by relatively high Phrase Building, on average, and a paucity of Grammatical Error, but without the high Narrative Productivity of Wernicke's aphasia and with relatively high Repair scores. Those classified as NABW showed the highest mean scores on Grammatical Complexity and Phrase Building and the lowest mean scores on Semantic Anomaly and Grammatical Errors, as would be expected in this mildest form of aphasia. This profile was very similar to the profile of anomic aphasia, but with higher scores on the "good" indicators and lower scores on the error indicators.

Because the syndromes are so variable, it is necessary to assess differences statistically. Results of the one-way ANOVAs for each factor (presented in Table 3) revealed that, when using factor scores derived from the pattern matrix, neither of the first two factors showed significant differences among subtypes (Grammatical Complexity: $p = .02$; Phrase Building: $p = .44$), but subtypes were significantly different on all other factors (all ps < .001). By contrast, all factors showed significant differences among subtypes using factor scores derived from the structure matrix (all ps <.001).

The ANOVA results confirm the impression from Figure 2 that structure matrix scores distinguish subtype profiles better than pattern matrix scores because they take into account factor intercorrelations. Another way of conceptualizing the different information provided by the pattern and structure matrices is to compare the variance accounted for by each set of loadings. The largest difference occurs for Phrase Building, which accounts for 10% of the variance according to the pattern matrix, but 19% in the structure matrix. The ANOVAs using these different scores indicated that it is not the unique variance in Phrase Building that distinguished aphasia subtypes but the variance that this factor shares with the other factors. Another difference is the Grammatical Error factor, which accounts for 8% of the variance in the pattern matrix but 14% in the structure matrix. Removing the shared variance changes the relationship of the factor to aphasia subtypes, most notably for

Table 3. Analyses of variance on factor scores across aphasia subtypes.

Effect	Pattern matrix scores		Structure matrix scores	
	F	*p*	*F*	*p*
Grammatical Complexity	2.7	.02	22.7	**< .001**
Phrase Building	1.0	.44	34.8	**< .001**
Semantic Anomaly	6.5	**< .001**	7.9	**< .001**
Grammatical Error	10.0	**< .001**	32.9	**< .001**
Narrative Productivity	17.4	**< .001**	39.3	**< .001**
Repair	8.2	**< .001**	9.4	**< .001**

Note. Significant *p* values are shown in bold type (corrected: .05/6 factors = .008).

Wernicke's aphasia. Considering the unique variance, Wernicke's aphasia shows high Grammatical Error scores, which is not surprising because Wernicke's aphasia is often characterized by paragrammatism. (This also explains the positive loading of propositional density on the Grammatical Error factor.) However, in the structure matrix, this is no longer the case, because high Grammatical Error scores in the group as a whole are strongly associated with low Narrative Productivity.

Because the structure matrix scores showed clearer distinctions among the aphasia subtypes, all subsequent analyses used only the factor scores generated from the structure matrix. Post hoc comparisons were conducted to follow up on the significant one-way ANOVAs from the structure matrix. Significant differences (see Table 4) support the interpretation of Figure 2 above. Broca's aphasia shows lower scores on Grammatical Complexity, Phrase Building and Narrative Productivity and higher scores on Grammatical Error than all other subtypes, except TCM aphasia, a reflection of the fluency–nonfluency distinction. Among the fluent aphasias, Wernicke's aphasia shows a higher incidence of Semantic Anomaly than either anomic or conduction aphasia or those diagnosed as NABW. Anomic aphasia is characterized by higher Repair scores than Broca's, Wernicke's, or NABW groups. It is notable that anomic and Broca's aphasia are distinguished on every factor, likely representing the difference in severity between the two syndromes, with the exception of the higher Repair score. Relatively high rates of repair are also found in those with conduction aphasia, significantly higher than those of Broca's and Wernicke's aphasia, although it is likely that the repair behaviors of anomic and conduction aphasia are qualitatively different. Conduction aphasia also shows significantly higher scores on the Grammatical Error factor than NABW speakers. Examining Figure 1 suggests that this is due to the relatively high rate of morphological errors in this group, a variable to which this factor contributes.

Table 4. Significant post hoc (Bonferroni) tests of analyses of variance comparing aphasia subtypes on factor scores derived from the structure matrix table. Complete results can be found in Supplemental Table S4.

Subtype comparison	Grammatical Complexity		Phrase Building		Semantic Anomaly		Grammatical Error		Narrative Productivity		Repair	
	diff	*p*	diff	*p*	diff	*p*	diff	*p*	diff	*p*	diff	*p*
ANO–BRO	0.91	**< .001**	1.15	**< .001**	**−0.58**	**.001**	−1.34	**< .001**	0.82	**< .001**	0.58	**< .001**
ANO–CON									−0.43	.025		
ANO–NABW	−0.64	.005							−0.53	.013	0.58	.002
ANO–TCM			0.79	.009					0.84	.005		
ANO–WER					−1.08	**< .001**			−1.04	**< .001**	0.76	**< .001**
BRO–CON	−0.99	**< .001**	−1.01	**< .001**			0.90	**< .001**	−1.25	**< .001**	−0.48	.004
BRO–NABW	−1.54	**< .001**	−1.54	**< .001**	0.79	.002	1.63	**< .001**	−1.35	**< .001**		
BRO–TCM												
BRO–WER	−1.03	**< .001**	−1.07	**< .001**			1.06	**< .001**	−1.86	**< .001**		
CON–NABW			−0.53	.038			0.73	.003				
CON–TCM									1.27	**< .001**		
CON–WER					−0.82	.020			−0.61	.028	0.66	.008
NABW–TCM	0.97	.011	1.18	**< .001**			−0.93	.015	1.37	**< .001**		
NABW–WER					−1.29	**< .001**						
TCM–WER									−1.89	**< .001**		

Note. Only significant differences are shown. The most robust differences ($p \leq .001$) are shown in bold. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; NABW = not aphasic by the Western Aphasia Battery; TCM = transcortical motor aphasia; WER = Wernicke's aphasia.

As a post hoc analysis, the correlations between factors, or between variables and factor scores, can be examined separately for each aphasia subtype. Although an in-depth exploration is beyond the scope of this article, I will discuss one example following up on the observation above of a paradoxical positive association of repair behaviors with several factors. Correlations between the Repair and Narrative factors were strongly positive for Broca's aphasia (r = .64, p < .001) and TCM aphasia (r = .68, p = .02) but negative for Wernicke's aphasia (r = −.68, p < .001) and those NABW (r = −.41, p = .03). Similarly, Repair was positively associated with Grammatical Complexity and Phrase Building for Broca's aphasia but negatively associated with these factors for Wernicke's aphasia. Thus, for those with nonfluent aphasia, being able to repair utterances appears to enhance production of connected speech, whereas it disrupts connected speech in more fluent types of aphasia. The opposing relationships for subtypes account for the weakness of the loadings for the sample as a whole, and the direction of the overall effect was determined by the large number of individuals with Broca's aphasia in the sample. This example illustrates that qualitative differences between subtypes may underlie some of the variance in factor scores.

Oddly, the correlation between Repair and Narrative Productivity was not significant for anomic aphasia (r = −.19, p = .08), for whom repairs were most frequent. However, this turns out to be because, for these speakers, the Repair factor was negatively associated with speech rate (r = −.41, p < .001) but positively associated with MLU (r = .24, p = .03), two variables strongly associated with the Narrative Productivity factor. This observation foreshadows a later finding of the latent profile analysis showing considerable variation within syndromes as well.

Linear Discriminant Analyses

The comparison of aphasia subtypes on their factor scores shows some expected broad distinctions, but a great deal of variability. To assess the extent to which the factors reflect syndrome differences, an LD analysis was conducted on the factor scores. LDs (linear combinations of the factors) were calculated, and these were used to calculate the probability that a given speaker belongs to a given aphasia subtype. The confusion matrix in Table 5 shows the number and proportion of speakers from each subtype that were predicted by the discriminant analysis to belong to each category; the top table shows classifications from factor scores, the bottom from the original variables. Correct classifications are shown on the diagonals (in bold), and the overall rate of correct classification is in the bottom right corner.

Examining the top table first, 61% of the speakers were correctly classified by LDs calculated from the factor scores. Not surprisingly, classifications were most accurate for the largest groups—anomic aphasia (78%) and Broca's aphasia (77%). However, the LD model overdiagnosed anomic aphasia by 44%, categorizing some speakers from every other subtype as being in this group. The least accurately identified subtypes were TCM (0%) and conduction (24%) aphasia. Although misclassification can be partly attributable small samples (particularly for TCM aphasia), another contributing factor is the lack of distinguishing characteristics for these subtypes (particularly for conduction aphasia), as illustrated by Figure 2. The table also illustrates which subtypes are most likely to be confusable with which other subtypes. Speakers with anomic aphasia were most likely to be misclassified with Broca's aphasia (9%) and vice versa (19% of those with Broca's aphasia classified as having anomic aphasia). Speakers with conduction aphasia and those designated as NABW were most likely to be misclassified as anomic (43% and 57%, respectively). The majority (73%) of those with TCM aphasia were classified as having Broca's aphasia. The most common misclassifications for those with Wernicke's aphasia were conduction aphasia (20%) and anomic aphasia (15%).

Are these confusions representative of those made by clinicians? To assess this, 88 misclassifications in the LD model (108 minus 20 confusions between anomic aphasia and NABW) were compared to 55 disagreements in the original data (63 disagreements minus 8 involving subtypes not included in the LD analysis; e.g., Broca's aphasia for global aphasia). Figure 3 shows the breakdown of confusions by proportion of total confusions in each set. The LD model behaved much like clinicians in the frequency with which it confused anomic with Broca's aphasia (28% for the LD model, 25% for clinicians), Broca's with TCM aphasia (9% vs. 11%), and conduction with Wernicke's aphasia (10% vs. 9%). The model was more likely to confuse anomic with conduction aphasia (24% vs. 7%), whereas clinicians were more likely to confuse Broca's with conduction aphasia (22% vs. 9%).

In the bottom of Table 5, showing predictions of LDs calculated from the set of 17 original variables, the patterns of confusions were very similar. However, the overall classification accuracy was somewhat better, at 67% compared to 61% accuracy using the factor scores, suggesting that the availability of more fine-grained data allowed the model to make more specific predictions. Classification accuracy was improved for all the subtypes, except anomic aphasia; although still overdiagnosed by 28%, speakers were slightly less likely to be correctly classified as having anomic aphasia than when factor scores were used. Chi-square analyses showed that the distributions of subtypes predicted by both models were significantly different from the actual distribution. The difference was greater for the factor score LD model (p < .001) than the original variable LD model (p = .004), although the two predicted distributions were not significantly different from each other (p = .130). Both provided significantly better predictions than chance alone (p < .001).

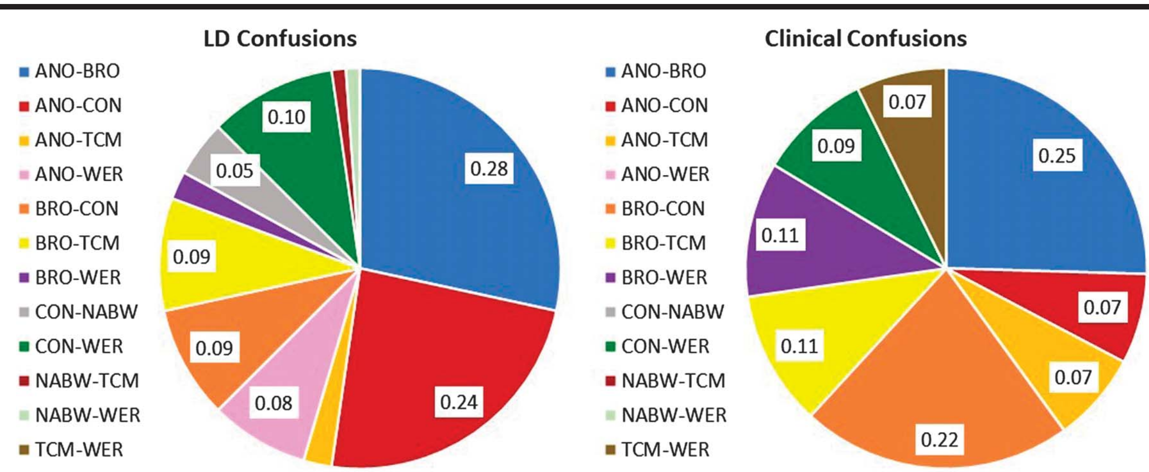
A more stringent assessment of an LD model is whether the LDs generated from one sample can generalize to another. To answer this question, the sample of speakers was split into two sets using a 60–40 split and randomly selecting an equal proportion of speakers from each subtype for each set. LDs were calculated on the factor scores from 60% of the sample (the training set) and used to predict

Table 5. Linear discriminant analysis results.

	ANO	BRO	CON	NABW	TCM	WER	*Predicted sums*
Predicted numbers (%) of each subtype, calculated from factor scores							
ANO	**66**	17	18	16	2	3	*122*
	(0.78)	(0.19)	(0.43)	(0.57)	(0.18)	(0.15)	*(1.44)*
BRO	8	**68**	6	0	8	1	*91*
	(0.09)	**(0.77)**	(0.14)	(0.00)	(0.73)	(0.05)	*(1.03)*
CON	3	2	**10**	1	0	4	*20*
	(0.04)	(0.02)	**(0.24)**	(0.04)	(0.00)	(0.20)	*(0.48)*
NABW	4	0	3	**11**	1	1	*20*
	(0.05)	(0.00)	(0.07)	**(0.39)**	(0.09)	(0.05)	*(0.71)*
TCM	0	0	0	0	**0**	0	*0*
	(0.00)	(0.00)	(0.00)	(0.00)	**(0.00)**	(0.00)	*(0.00)*
WER	4	1	5	0	0	**11**	*21*
	(0.05)	(0.01)	(0.12)	(0.00)	(0.00)	**(0.55)**	*(1.05)*
Actual Sums	*85*	*88*	*42*	*28*	*11*	*20*	**166**
	(1.00)	*(1.00)*	*(1.00)*	*(1.00)*	*(1.00)*	*(1.00)*	**(0.61)**
Predicted numbers (%) of each subtype, calculated from original variables							
ANO	**65**	12	15	11	3	3	*109*
	(0.76)	(0.14)	(0.36)	(0.39)	(0.27)	(0.15)	*(1.28)*
BRO	6	**72**	6	0	6	1	*91*
	(0.07)	**(0.82)**	(0.14)	(0.00)	(0.55)	(0.05)	*(1.03)*
CON	6	3	**16**	1	0	3	*29*
	(0.07)	(0.03)	**(0.38)**	(0.04)	(0.00)	(0.15)	*(0.69)*
NABW	6	0	3	**15**	0	0	*24*
	(0.07)	(0.00)	(0.07)	**(0.54)**	(0.00)	(0.00)	*(0.86)*
TCM	1	0	0	0	**2**	0	*3*
	(0.01)	(0.00)	(0.00)	(0.00)	**(0.18)**	(0.00)	*(0.27)*
WER	1	1	2	1	0	**13**	*18*
	(0.01)	(0.01)	(0.05)	(0.04)	(0.00)	**(0.65)**	*(0.90)*
Actual Sums	*85*	*88*	*42*	*28*	*11*	*20*	**183**
	(1.00)	*(1.00)*	*(1.00)*	*(1.00)*	*(1.00)*	*(1.00)*	**(0.67)**

Note. Accurate classifications are shown on the diagonals (shown in bold). Totals are shown in italics. Proportions are calculated from the actual sums. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; WER = Wernicke's aphasia; TCM = transcortical motor aphasia; NABW = not aphasic by the Western Aphasia Battery.

Figure 3. Comparison of misclassifications by the linear discriminant (LD) model and disagreements between clinical and Western Aphasia Battery diagnoses in the original data. ANO = anomic aphasia; BRO = Broca's aphasia; CON = conduction aphasia; WER = Wernicke's aphasia; TCM = transcortical motor aphasia; NABW = not aphasic by the Western Aphasia Battery.



subtype classification in both the training set (similar to the analysis above) and the remaining 40% (the testing set). Overall classification accuracy was 62% on the training set, similar to the accuracy on the whole set, and showed a similar distribution of errors (lower accuracy for the smaller groups, overdiagnosis of anomic aphasia). For the testing set, the same pattern was shown, but overall accuracy dropped to 57%.

Latent Profile Analysis

The LD models captured broad patterns but were not particularly accurate at predicting aphasia subtypes. At best, only two thirds of individual speakers were accurately classified, suggesting that the subtype labels only partially reflect the sources of variance in spontaneous speech. To determine how the data would be classified in a more agnostic way (i.e., without relying on aphasia subtype labels defined a priori), an LPA was conducted. Solutions were generated for a range of two to 10 profiles, using three models with different assumptions about the data (Model 1: equal variances, covariances fixed to zero; Model 2: varying variances, covariances fixed to zero; and Model 3: equal variances and equal covariances). Of the 27 models generated, the best fit based on a combination of multiple fit indices (Akogul & Erisoglu, 2017) generated seven profiles using Model 2.

Mean factor scores for each latent profile generated by the LPA are shown in Figure 4. Comparing this figure to the factor scores for traditional aphasia subtypes shown in Figure 2 illustrates some striking similarities. Most notably, Profile 3 is a similar but exaggerated version of the profile for Broca's aphasia, with low scores on Grammatical Complexity, Phrase Building, and Narrative Productivity and high scores on Grammatical Error. Profiles 5 and 7 are very similar to the profile for the NABW group. Profile 5 is almost identical, except for the above average Repair score, which is more like anomic aphasia. This is reversed in Profile 7, more like the actual NABW group, and the Grammatical Complexity and Narrative Productivity scores are higher, reflecting the better performing speakers in this group. Profile 4 is similar to the profile for anomic aphasia. Profile 2 is similar to TCM aphasia. There is no single profile similar to Wernicke's aphasia, but the high Semantic Anomaly scores characteristic of this subtype are represented on Profile 1, and the high Narrative Productivity is represented on Profile 6, as if these profiles were modeling more and less severe forms of Wernicke's aphasia, respectively. However, Profile 1 also showed characteristics of Broca's aphasia, with low Phrase Building, Grammatical Complexity, and Narrative Productivity scores and some Grammatical Errors, although not as many as Profile 3. None of the profiles was similar to that of conduction aphasia.

The composition of each latent profile is represented in Table 6 and in the pie charts in Supplemental Figure S3. In Table 6, the profiles are labeled according to visual comparison with the factor score profiles discussed above (see Figure 2) and the makeup of each profile, that is, which aphasia subtypes were represented. The table shows raw numbers and proportions of each subtype that were captured by each latent profile. For example, 35% of speakers with Broca's aphasia fell into Profile 2 ("TCM/High Broca") and 33% fell into Profile 3 ("Low Broca"). More than half (55%) of those with TCM aphasia were in Profile 2, while 18% were put into Profile 3. Over a third (36%) of the speakers with anomic aphasia were characterized by Profile 4 ("Anomic"), and another 25% were characterized by Profile 5 ("High Anomic/NABW"). Those with conduction aphasia were scattered throughout the profiles, but most (24%) were captured by the "Anomic" profile. Of those with Wernicke's aphasia, a third (30%) were in Profile 6 ("High Wernicke"), 25% were in Profile 7 ("High NABW"), and 20% were in Profile 1 ("Low Wernicke/High Broca"). Over half (57%) of those designated NABW were in the Profile 5 ("High Anomic/NABW"), and 18% were in Profile 7 ("High NABW"). Thus, the profiles can

Figure 4. Mean factor scores for each latent profile. Error bars indicate standard deviations. GRCOM = Grammatical Complexity factor; PHRASE = Phrase Building factor; SEM = Semantic Anomaly factor; GRERR = Grammatical Error factor; NARR = Narrative Productivity factor; REP = Repair factor; SM = Structure Matrix scores.

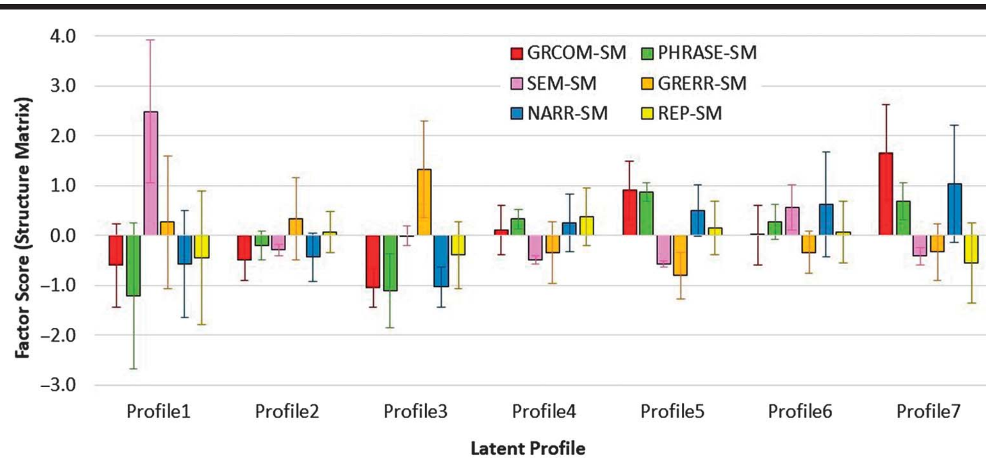


Table 6. Latent profiles generated from factor scores.

Latent profiles	BRO	TCM	ANO	CON	WER	NABW	*Total*
Profile 1	16	0	2	4	4	0	*26*
("Low Wernicke/High Broca")	(0.18)	(0.00)	(0.02)	(0.10)	(0.20)	(0.00)	*(0.09)*
Profile 2	31	6	18	8	1	0	*64*
("TCM/High Broca")	**(0.35)**	**(0.55)**	(0.21)	(0.19)	(0.05)	(0.00)	*(0.23)*
Profile 3	29	2	1	1	0	0	*33*
("Low Broca")	(0.33)	(0.18)	(0.01)	(0.02)	(0.00)	(0.00)	*(0.12)*
Profile 4	5	1	31	10	3	7	*57*
("Anomic")	(0.06)	(0.09)	**(0.36)**	**(0.24)**	(0.15)	(0.25)	*(0.21)*
Profile 5	0	0	21	8	1	16	*46*
("High Anomic/NABW")	(0.00)	(0.00)	(0.25)	(0.19)	(0.05)	**(0.57)**	*(0.17)*
Profile 6	7	1	7	7	6	0	*28*
("High Wernicke")	(0.08)	(0.09)	(0.08)	(0.17)	**(0.30)**	(0.00)	*(0.10)*
Profile 7	0	1	5	4	5	5	*20*
("High NABW")	(0.00)	(0.09)	(0.06)	(0.10)	(0.25)	(0.18)	*(0.07)*
Total	*88*	*11*	*85*	*42*	*20*	*28*	*274*
	(0.32)	*(0.04)*	*(0.31)*	*(0.15)*	*(0.07)*	*(0.10)*	*(1.00)*

Note. Bold type indicates the latent profile grouping of the highest proportion from each subtype (e.g., most speakers with TCM aphasia were grouped into Profile 2). Totals are shown in italics. BRO = Broca's aphasia; TCM = transcortical motor aphasia; ANO = anomic aphasia; CON = conduction aphasia; WER = Wernicke's aphasia; NABW = not aphasic by the Western Aphasia Battery.

be roughly mapped onto traditional aphasia syndromes. However, each profile is made up of no fewer than four aphasia subtypes, and each aphasia subtype is represented on multiple profiles (see Figure 4). Post hoc comparisons of the syndromes that were split into different profiles confirm the interpretation of higher and lower performing subsets. For example, those labeled "Low Broca" in Profile 3 showed significantly worse performance than those labeled "High Broca" in Profile 2 on 13 of the 17 original variables, as well as the WAB Fluency scale score (*t* tests, $p < .05$). Differences were similarly noted between the "Anomic" (Profile 4) and "High Anomic" (Profile 5), "Low Wernicke" (Profile 1) and "High Wernicke"(Profile 6), and "NABW" (Profile 5) and "High NABW" (Profile 7) subsets.

Discussion

What Latent Factors Underlie Spontaneous Speech in Aphasia?

A factor analysis of 17 spontaneous speech variables in 274 individuals with aphasia yielded six factors. Factor 1, Grammatical Complexity, reflected the use of syntactically complex structures and showed maximal distinction between Broca's aphasia on the low end and those designated NABW on the high end. Factor 2 (Phrase Building) represented the ability to retrieve words (particularly proposition-forming words, such as verbs, adjectives, and prepositions) and formulate them into basic phrasal units and can therefore be conceptualized as an utterance-level fluency measure. A separate factor, Narrative Productivity (Factor 5), reflected the ability to produce more speech overall more quickly and therefore represents a narrative-level fluency measure. Factor 3 reflected the production of Semantic Anomalies—jargon, neologistic errors, and unrelated word errors—and was particularly characteristic of Wernicke's aphasia. The fourth factor was Grammatical Error, which contributed to grammatically inaccurate utterances, morphological errors, and an elevated content:function ratio. Speakers with nonfluent aphasia (Broca's, TCM) were distinguished from those with fluent aphasia (anomic, conduction, Wernicke's, NABW) on Grammatical Complexity, Phrase Building, Grammatical Error, and Narrative Productivity. Conduction and anomic aphasia were characterized by high scores on the sixth factor, Repair, which represented the use of repeated and retraced utterances. Thus, the primary dimensions of variability captured were the ability to produce fluent connected speech (Phrase Building, Narrative Productivity, Repair), the accuracy of words and sentences (Semantic Anomaly, Grammatical Error), and the complexity of utterances (Grammatical Complexity).

How Does the Factor Analysis Compare to Previous Studies?

This factor solution showed both similarities and differences compared to previous findings, focusing on those studies that took a comparable approach of including a wide range of spontaneous speech variables. The identification of six factors is in line with findings of Wagenaar et al. (1975) and Vermeulen et al. (1989), which identified, respectively, six factors (n = 107, 30 variables) and five factors (n = 121, 18 variables). Casilio et al. (2019), who had a smaller sample and used ratings, identified four factors (n = 24, 23 variables). The amount of variance accounted for is difficult to compare across studies, because this depends on the method of analysis (PCA vs. true factor analysis) and the set of variables included. Vermeulen et al. used a similar approach to the current study, and their analysis accounted for 50% of the variance, similar to the 52% accounted for here.

As in the current study, both the Wagenaar and Vermeulen studies found that the first factor represented aspects of sentence formulation, including utterance length, speech rate, and either syntactic complexity (Wagenaar) or function word use (Vermeulen). Wagenaar's first factor also included self-corrections, although the current analysis yielded a separate factor for corrections (Repair). In the current study, speech rate also loaded more heavily on Factor 5, partially dissociating phrase-level fluency from narrative-level fluency. This might have been driven by the inclusion of a total utterance count in the current study, which was not included in either of the previous studies. Rochon et al. (2000) described a similar distinction for individuals with nonfluent aphasia. In their results, "the variables in Factor 1 were computed solely on the basis of material that appeared in sentence contexts, whereas those in Factor 2 were computed across the whole narrative sample…reflecting a disposition toward propositional utterances in the case of Factor 1 and ease of word production in the case of Factor 2" (p. 209).

All of the studies identified paraphasia factors overlapping with the Semantic Anomaly factor in the current study. In the Casilio et al. (2019) study, all paraphasias loaded together except for stereotypies and automatisms, which loaded on the agrammatism factor. Wagenaar et al. (1975) found that verbal paraphasias (real-word errors, not otherwise specified) loaded separately from automatisms and neologisms. Vermeulen et al. (1989) differentiated phonological paraphasias, of which they counted several types, from "irrelevant" (i.e., unrelated) word errors and neologisms. Similar to the current study, semantic paraphasias in the Vermeulen study did not make an important contribution to any factor and showed a very low communality—only 14% of the variance in semantic error production was accounted for by the identified factors.

Only Vermeulen et al. (1989) included lexical diversity, which loaded along with empty speech on a factor separate from paraphasias, thus distinguishing relatively normal from relatively abnormal features of language. This "Vocabulary" factor accounted for only 5% of the variance. In the current study, no pure lexical retrieval factor was generated, although measures of lexical retrieval (MATTR and propositional density) were associated with the Phrase Building factor. This may seem surprising given the ubiquity of word retrieval difficulties in aphasia and the obvious importance of word retrieval for spontaneous speech. However, because anomia is a common symptom across aphasia subtypes, it is less useful in distinguishing between them and, therefore, less likely to emerge as an important component of factor analyses. Furthermore, lexical retrieval difficulties are notoriously difficult to measure in spontaneous speech and are usually less marked than in single-word naming tasks (e.g., Fergadiotis et al., 2019; Gordon & Kindred, 2011; Mayer & Murray, 2003). As Vermeulen et al. noted, "in free speech, various tactics (avoiding difficult words, the use of indefinite terms, circumlocutions, and such) are employed that allow the patient to maintain a socially acceptable level of fluency in the face of severe difficulty in finding the right word" (p. 262).

Across studies, differences in factor structure can be related to differences in the variables analyzed, such as the lack of grammatical measures in the Vermeulen study or the lack of articulatory measures or ratings in the current study. In addition, the context in which spontaneous speech samples were elicited can affect the structure of the narrative and the relative importance of different variables (e.g., Alyaha et al., 2020; Armstrong et al., 2013; Fergadiotis & Wright, 2011; Stark, 2019).

How Successful Was the Factor Analysis in Capturing Spontaneous Speech in Aphasia?

The success of this factor solution in accounting for variability in spontaneous speech measures can be assessed by examining the structure of the factor solution and the variance accounted for. Several variables showed low communalities (high uniquenesses)—notably semantic errors and noun marking, which were removed, but also morphological errors, circumlocution/empty speech, and total utterances, which were retained—indicating that they were not well accounted for by the identified factors. Noun markers are disrupted much less often than verb inflections (Goodglass & Berko, 1960); in any case, the noun marking measure only captures the frequency of noun inflections, not their accuracy. Morphological errors can be difficult to identify, particularly in agrammatic speech where they are most likely to occur. There must be sufficient syntactic structure to identify the intended morphology before errors can be coded. This is similar to the idea of pseudo-agrammatism proposed by Goodglass et al. (2001a), who noted that the output of global aphasia is too sparse to be able to reliably detect agrammatism. The lack of salience of total utterances in the current analysis can be accounted for by the use of proportions for the majority of the variables. Although the number of utterances distinguished aphasia subtypes (see Table 1 and Supplemental Figure S1), it showed minimal correlations with all other variables, except speech rate (see Supplemental Figure S2).

The lack of diagnostic value of semantic errors has been noted before and is related to the observation that they might arise for a multitude of reasons, including disruption to semantics or to the availability of a word's phonological form, resulting in production of a co-activated competitor (Borman et al., 2008; Caramazza & Hillis, 1990; Howard & Gatehouse, 2006). Like circumlocution and empty speech, a semantic error might also occur as a compensatory response produced in the attempt to find an elusive word, rather than as a true error of retrieval. Interestingly, the Semantic Anomaly factor did not contribute to semantic errors, suggesting that, at least in connected speech, semantically related errors arise from qualitatively different underlying processes from semantically *unrelated* errors. Dell et al. (1997) noted a similar distinction between related and unrelated errors in naming tasks, attributing the former to noise or decay in the lexical network and the latter to disrupted connections.

Overall, almost half of the variance in the set of spontaneous speech measures was left unaccounted for by the factor solution. Although this was on a par with previous work (e.g., Vermeulen et al., 1989), it raises questions about what else contributes to variability in spontaneous speech production. One aspect of variability not well captured here was motor speech. Although AphasiaBank includes binary classifications of dysarthria and apraxia in its database, these diagnoses are often missing. As both missing data and categorical data are problematic for factor analyses, they were excluded from the current study. In previous studies (Casilio et al., 2019; Clough & Gordon, 2020; Vermeulen et al., 1989; Wagenaar et al., 1975), rating measures have been used to reflect aspects of articulation in a graded manner, but the current study relied solely on the more objectively measurable data that were available in AphasiaBank. Nevertheless, it has been shown that, with appropriate training, ratings can be sufficiently reliable (Casilio et al., 2019; Strand et al., 2014) and may be the best way to capture variability in certain dimensions. (Another idea, raised by an anonymous reviewer, is to count the number of unintelligible utterances as a potential index of motor speech difficulties.) Also unavailable for the full sample in the current study were higher level measures of discourse coherence and cohesion. Although I made the explicit decision to focus on microlinguistic measures, it has been shown that micro- and macrolinguistic levels can interact (e.g., Andreetta & Marini, 2015; Christiansen, 1995; Hazamy & Obermeyer, 2020; Manning & Franklin, 2016; Wright et al., 2010). Other studies have shown that cognitive skills can also contribute to linguistic variability in discourse (e.g., Cahana-Amitay & Jenkins, 2018; Murray, 2012).

Additional sources of variability arise from individual differences existing premorbidly or as a response to stroke. Caramazza (1984) proposed that the performance of people with aphasia reflects four factors: normal variation, the direct consequences of the linguistic deficit, indirect consequences of deficits (e.g., effects of cognitive impairments on linguistic functioning or vice versa), and compensatory operations, or how the PWA responds to their deficits (see also Hazamy & Obermeyer, 2020). These may also interact with each other such that, for example, premorbid style (an aspect of normal variation) affects the types of compensatory strategies or reactions adopted to manage difficulties in linguistic formulation. So far, there is no clear understanding of how best to capture such individual differences. If such sources of "error" cannot be measured, it is more valid to leave them unaccounted for by selecting a factor analytic method that focuses on common variance among linguistic measures than to allow them to contribute to factor loadings, which is what PCA approaches do (Kass & Tinsley, 1979; Wilson & Hula, 2019).

How Well Did the Factor Analysis Correspond to Traditional Aphasia Syndrome Differences?

Using pattern matrix factor scores, which represent the unique variance accounted for by each factor, aphasia subtypes were distinguished (according to one-way ANOVAs) by four of the six factors. Using structure matrix factor scores, all factors significantly differentiated among the aphasia subtypes, suggesting that the shared variance between factors is informative and should be taken into account in modeling different patterns of spontaneous speech in aphasia. Linear discriminant analyses indicated that the factor scores generated correct aphasia subtype classifications only 61% of the time, whereas the original variables correctly classified 67% of the individual speakers. Although some information was lost in the use of factor scores, the accuracy of classification was still quite low even when using the original variables and was not a significant improvement over the factor scores. The level of agreement between the LD analysis and clinical categories was similar, quantitatively and qualitatively, to the kinds of confusions that were found between clinical judgments and WAB scores in the original data.

The latent profile analysis generated profiles that looked remarkably similar to aphasia subtypes in some respects but differed markedly in others. Similarity was strongest for subtypes whose spontaneous speech was most affected (Broca's aphasia) and least affected (NABW). This observation, combined with the division of lower and higher performing subtypes into different profiles, suggests that the LPA paid more attention to severity levels when classifying the data.

Differences between traditional methods of syndrome classification and such data-driven approaches reflect the fuzziness inherent in any attempt to classify naturally occurring data that vary along multiple intercorrelated dimensions. Aphasia syndromes overlap with one another extensively and also show extensive variability within syndromes. Thus, mathematical solutions may make different decisions than clinicians do in deciding which sources of variability are most salient. The extent to which the models matched clinical classifications depended to a large extent on the distribution of subtypes in the sample. As the most frequent categories, Broca's and anomic subtypes were most accurately classified by the LD analysis. In fact, the high degree of overdiagnosis of anomic aphasia suggests that it was treated as a default category by virtue of its frequency in the sample. In the LPA, both Broca's aphasia and anomic aphasia were fairly clearly replicated, Broca's aphasia on Profile 3 and anomic aphasia on Profile 4, and on Profile 5 if NABW is considered a less severe form of anomic aphasia. In addition to frequency of occurrence, distinctiveness of the syndrome also matters. Even though speakers with conduction aphasia made up 15% of the sample (n = 42), this syndrome was not captured by the LPA as well as either Wernicke's aphasia (7%, n = 20) or TCM aphasia (4%, n = 11). Another important factor is severity, as evidenced by the LPA's classification of subtypes of different severity levels into different profile categories. Although clinicians, whose goal is to identify and treat underlying deficits, are generally more focused on qualitative differences, the LPA appears to have given extra weight to severity differences.

Limitations

This attempt to map factor scores onto different aphasia subtypes entailed some important limitations relating to the use of naturalistic data. First, the study included vastly different numbers of participants of each subtype. Because the sample was dominated by Broca's and anomic aphasia, the factors primarily reflect patterns evident in these two subtypes. However, to the extent that this distribution is representative of the larger population of individuals with subacute and chronic aphasia, the factor solution can be considered generalizable. Furthermore, the findings illustrated that even a relatively rare type of aphasia will contribute to the factor solution if the pattern of performance is sufficiently distinctive. For example, although only 7% of the sample had Wernicke's aphasia, one of the factors represented the characteristic production of semantically anomalous speech.

Second, traditional aphasia syndromes are defined on a broader pattern of linguistic behavior than their spontaneous verbal output—repetition and comprehension abilities are particularly relevant. For example, including information about repetition would presumably have enhanced the ability to distinguish conduction aphasia, which did not show a particularly distinctive pattern of factor scores, from other syndromes.

Third, the reliability and validity of any model solution are limited by the quality of the data entering the model. The data in AphasiaBank are valuable in that they are extremely rich and were gathered following a standardized protocol. However, some of the codes are not sufficiently fine-grained, for example, to represent the different types of grammatical errors produced in Broca's and Wernicke's aphasia or the different types of repairs used by those with anomic and conduction aphasia.

Finally, although multivariate normality is not required, factor analysis performs better with data that are continuous, linearly related, and without extreme outliers (Beavers et al., 2013). Similarly, both LD and LPA generate more accurate solutions with "well-behaved" data. However, aphasia data are notoriously messy and frequently skewed. A large sample helps compensate for noisy data, but patterns can be less stable, particularly when examining smaller subgroups.

Conclusions

The factor analysis, conducted on a large and diverse sample of individuals with aphasia, illustrates several important observations about spontaneous speech in aphasia. Results revealed a useful distinction between phrase-level fluency, reflected in basic word combinations, and narrative-level fluency, reflected in the production of more speech more quickly. Other factors reflected characteristic patterns of traditional syndromes—Semantic Anomaly for Wernicke's aphasia; Grammatical Error for Broca's aphasia; Repair for anomic aphasia—but were by no means exclusive to those syndromes. Linear discriminant and latent profile analyses both reflected the most salient qualitative differences between syndromes but were also influenced by quantitative differences in severity. Although the factors and latent profiles identified here contribute to our understanding of how spontaneous speech varies across a widely varying population of individuals with aphasia, they sometimes obscured qualitative differences among aphasia subtypes, such as the different role that repair behaviors can play.

There are many legitimate criticisms of the traditional classifications of aphasia used here, but they serve as a clinically intuitive approximation until the field can establish a more reliable and valid replacement. Importantly, factor analyses also provide a simplified solution to characterizing variation in aphasia. Halai et al. (2017), among many others, criticize aphasia classification schemes as having both "very fuzzy boundaries between 'categories' and considerable variation of profile within each 'category'" (p. 277). However, factor analyses and other data-driven approaches such as LD analysis and LPA clearly do not solve this problem. Nor do they necessarily result in "unified" or "stable" models (Alyaha et al., 2020), as the variability in their work attests. Not only are factor analysis and clustering analyses particularly susceptible to the set of variables included (as discussed above), the participant sample, and the methodology employed, but any model that attempts to provide a single explanation of the problem space is bound to be as much of a simplification as the traditional syndrome approach. The use of orthogonal rotation may solve the ubiquitous problem of fuzzy boundaries between natural kinds, but one wonders how "natural" the newly identified kinds would be.

Like all models, factor analysis involves a trade-off between explanatory power and parsimony, and the optimal balance of this trade-off may depend on the goals of the study. For neuroanatomic localization, parsimony may be preferred, although the explanatory power of neural underpinnings for such broadly defined dimensions of language as "verbal quality" and "verbal quantity" (Alyaha et al., 2020) is likely to be limited. Certainly for understanding behavioral patterns in aphasia subtypes and in individuals, and probably for understanding neuroanatomical models, additional explanatory power is required. To achieve this, a combination of quantitative and qualitative methods is advocated. Quantitative analysis of spontaneous speech must be balanced with qualitative interpretation, including attending to dependencies between variables and levels of analysis (Grande et al., 2008; Prins & Bastiaanse, 2004). The current study attempted to do this by making use of both data-driven and theory-driven approaches. A factor solution for the entire sample was generated, then the factor scores were explored to reveal how well they mapped onto traditional aphasia subtypes differed on the derived factors. Results were compared to two clustering approaches, one supervised (LD analysis) and one unsupervised (LPA). Alternatively, researchers could generate separate factor solutions for different aphasia subtypes or levels of severity of aphasia, given sufficient sample sizes. Regardless of the quantitative approach, a fine-grained analysis is required of the variables that contribute to the factors and how they interact in order to ensure that results are clinically meaningful. Such efforts have been ongoing for decades. As Kertesz and Phipps (1977) noted, "the balance between specificity and objectivity, clinical relevance and mathematical abstraction is yet to be worked out fully" (p. 10). Until then, multivariate analyses of aphasia should weigh findings of broad quantitative trends against careful consideration of qualitative differences among aphasia subtypes and individuals with aphasia.

Author Contributions

Jean K. Gordon: Conceptualization (Lead), Data curation (Lead), Formal analysis (Lead), Funding acquisition (Lead), Investigation (Lead), Methodology (Lead), Writing – original draft (Lead), Writing – review & editing (Lead).

Acknowledgments

This study was partially supported by a New Century Scholar grant from the American Speech-Language-Hearing Foundation. I would also like to express my appreciation to the developers and maintainers of AphasiaBank (particularly Davida Fromm), to the AphasiaBank contributors, and to all the individuals with aphasia who generously allowed AphasiaBank to archive their data.

References

Akogul, S., & Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy, 19*(9), 452. https://doi.org/10.3390/e19090452

Alyaha, R. S. W., Halai, A. D., Conroy, P., & Lambon Ralph, M. A. (2018). Noun and verb processing in aphasia: Behavioural profiles and neural correlates. *NeuroImage: Clinical, 18,* 215–230. https://doi.org/10.1016/j.nicl.2018.01.023

Alyaha, R. S. W., Halai, A. D., Conroy, P., & Lambon Ralph, M. A. (2020). A unified model of post-stroke language deficits including discourse production and their neural correlates. *Brain: A Journal of Neurology, 143*(5), 1541–1554. https://doi.org/10.1093/brain/awaa074

Andreetta, S., & Marini, A. (2015). The effect of lexical deficits on narrative disturbances in fluent aphasia. *Aphasiology, 29*(6), 705–723. https://doi.org/10.1080/02687038.2014.979394

Armstrong, E., Ferguson, A., & Simmons-Mackie, N. (2013). Discourse and functional approaches to aphasia. In I. Papathanasiou, P. Coppens, & C. Potagas (Eds.), *Aphasia and related neurogenic communication disorders* (pp. 217–231). Jones & Bartlett.

Basilakos, A., Fillmore, P., Rorden, C., Guo, D., Bonhila, L., & Fridrikkson, J. (2014). Regional white matter damage predicts speech fluency in chronic post-stroke aphasia. *Frontiers in Human Neuroscience, 17*. https://doi.org/10.3389/fnhum.2014.00845

Bastiaanse, R., & Prins, R. S. (1994). Communicative speech therapy in aphasia: What does it mean, can it be effective and how should it be done? *Aphasiology, 8*(5), 482–488. https://doi.org/10.1080/02687039408248672

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., & Skolits, G. J. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation, 18,* Article 6. https://doi.org/10.7275/qv2q-rk76

Benson, D. F. (1967). Fluency in aphasia: Correlation with radioactive scan localization. *Cortex, 3*(4), 373–394. https://doi.org/10.1016/S0010-9452(67)80025-X

Berndt, R. S., Wayland, S., Rochon, E., Saffran, E., & Schwartz, M. (2000). *Quantitative Production Analysis: A training manual for the analysis of aphasic sentence production*. Psychology Press.

Borman, T., Kulke, F., Wallesch, C.-W., & Blanken, G. (2008). Omissions and semantic errors in aphasia naming: Is there a link? *Brain and Language, 104*(1), 24–32. https://doi.org/10.1016/j.bandl.2007.02.004

Borovsky, A., Saygin, A. P., Bates, E., & Dronkers, N. (2007). Lesion correlates of conversational speech production deficits. *Neuropsychologia, 45*(11), 2525–2533. https://doi.org/10.1016/j.neuropsychologia.2007.03.023

Butler, R. A., Lambon Ralph, M. A., & Woollams, A. M. (2014). Capturing multidimensionality in stroke aphasia: Mapping principal behavioural components to neural structures. *Brain: A Journal of Neurology, 137*(12), 3248–3266. https://doi.org/10.1093/brain/awu286

Cahana-Amitay, D., & Jenkins, T. (2018). Working memory and discourse production in people with aphasia. *Journal of Neurolinguistics, 48,* 90–103. https://doi.org/10.1016/j.jneuroling.2018.04.007

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language, 21,* 9–20. https://doi.org/10.1016/0093-934X(84)90032-4

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex, 26*(1), 95–122. https://doi.org/10.1016/S0010-9452(13)80077-9

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

Christiansen, J. A. (1995). Coherence violations and propositional usage in the narratives of fluent aphasics. *Brain and Language, 51*(2), 291–317. https://doi.org/10.1006/brln.1995.1062

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

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Erlbaum.

Copland, D. A., McNeil, M. R., & Meinzer, M. (2018). Aphasia theory, models, and classification. In L. LaPointe & J. Steirwalt (Eds.), *Aphasia and related neurogenic language disorders* (5th ed., pp. 23–34). Thieme.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10,* Article 7. https://doi.org/10.7275/jyj1-4868

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

Covington, M., & McFall, J. D. (2008). *The moving-average type–token ratio*. Paper presented at the Linguistics Society of America, Chicago, IL, United States.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Edwards, S. (2005). Fluent aphasia. Cambridge University Press. https://doi.org/10.1017/CBO9780511486548

Feenaughty, L., Basilakos, A., Bonilha, L., den Ouden, D.-B., Rorden, C., Stark, B., & Fridrikkson, J. (2017). Non-fluent speech following stroke is caused by impaired efference copy.

Cognitive Neuropsychology, 34(6), 333–346. https://doi.org/10.1080/02643294.2017.1394834
Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modelling confrontation naming and discourse informativeness using structural equation modelling. *Aphasiology, 33*(5), 544–560. https://doi.org/10.1080/02687038.2018.1482404
Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898
Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.
Fong, M. W. M., Van Patten, R., & Fucetola, R. P. (2019). The factor structure of the Boston Diagnostic Aphasia Examination, Third Edition. *Journal of the International Neuropsychological Society, 25*(7), 772–776. https://doi.org/10.1017/S1355617719000237
Fromm, D., Forbes, M., Holland, A. L., 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
Fromm, D., MacWhinney, B., & Thompson, C. K. (2020). Automation of the Northwestern Narrative Language Analysis System. *Journal of Speech, Language, and Hearing Research, 63*(6), 1835–1844. https://doi.org/10.1044/2020_JSLHR-19-00267
Glosser, G., & Deser, T. (1991). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language, 40*(1), 67–88. https://doi.org/10.1016/0093-934X(91)90117-J
Goodglass, H., & Berko, J. (1960). Agrammatism and inflectional morphology in English. *Journal of Speech and Hearing Research, 3*(3), 257–267. https://doi.org/10.1044/jshr.0303.257
Goodglass, H., & Kaplan, E. (1972). *The assessment of aphasia and related disorders* (1st ed.). Lea & Febiger.
Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders* (2nd ed.). Lea & Febiger.
Goodglass, H., Kaplan, E., & Barresi, B. (2001a). *The assessment of aphasia and related disorders* (3rd ed.). Lippincott Williams & Wilkins.
Goodglass, H., Kaplan, E., & Barresi, B. (2001b). *Boston Diagnostic Aphasia Examination–Third Edition*. Lippincott Williams & Wilkins.
Gordon, J. K. (1998). The fluency dimension in aphasia. *Aphasiology, 12*(7/8), 673–688. https://doi.org/10.1080/02687039808249565
Gordon, J. K. (2006). A quantitative production analysis of picture description. *Aphasiology, 20*(2/3/4), 188–204. https://doi.org/10.1080/02687030500472777
Gordon, J. K., & Clough, S. (2020). How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology, 34*(5), 643–663. https://doi.org/10.1080/02687038.2020.1712586
Gordon, J. K., & Kindred, N. K. (2011). Word retrieval in ageing: An exploration of the task constraint hypothesis. *Aphasiology, 25*(6–7), 774–788. https://doi.org/10.1080/02687038.2010.539699
Gorsuch, R. L. (1983). *Factor analysis*. Erlbaum.
Grande, M., Hussmann, K., Bay, E., Christoph, S., Piefke, M., Willmes, K., & Huber, W. (2008). Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia. *International Journal of Language & Communication Disorders, 43*(4), 408–426. https://doi.org/10.1080/13682820701685991
Halai, A. D., Woollams, A. M., & Lambon Ralph, M. A. (2017). Using principal components analysis to capture individual differences with a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex, 86*, 275–289. https://doi.org/10.1016/j.cortex.2016.04.016
Hanson, W. R., Riege, W. H., Metter, E. J., & Inman, V. M. (1982). Factor-derived categories of chronic aphasia. *Brain and Language, 15*(2), 369–380. https://doi.org/10.1016/0093-934X(82)90066-9
Hazamy, A. A., & Obermeyer, J. (2020). Evaluating informative content and global coherence in fluent and non-fluent aphasia. *International Journal of Language & Communication Disorders, 55*(1), 110–120. https://doi.org/10.1111/1460-6984.12507
Helm-Estabrooks, N. (1992). *Aphasia Diagnostic Profiles*. Pro-Ed.
Hoffmann, M., & Chen, R. (2013). The spectrum of aphasia subtypes and etiology in subacute stroke. *Journal of Stroke and Cerebrovascular Diseases, 22*(8), 1385–1392. https://doi.org/10.1016/j.jstrokecerebrovasdis.2013.04.017
Holland, A. L., Fromm, D., & Swindell, C. S. (1986). The labeling problem in aphasia: An illustrative case. *Journal of Speech and Hearing Disorders, 51*(2), 176–180. https://doi.org/10.1044/jshd.5102.176
Howard, D., & Gatehouse, C. (2006). Distinguishing semantic and lexical word retrieval deficits in people with aphasia. *Aphasiology, 20*(9–11), 921–950. https://doi.org/10.1080/02687030600782679
Howard, D., Swinburn, K., & Porter, G. (2010). Putting the CAT out: What the Comprehensive Aphasia Test has to offer. *Aphasiology, 24*(1), 56–74. https://doi.org/10.1080/02687030802453202
Huber, W., Poeck, K., & Willmes, K. (1984). The Aachen Aphasia Test. *Advances in Neurology, 42*, 291–303.
Jones, L. V., & Wepman, J. M. (1961). Dimensions of language performance in aphasia. *Journal of Speech and Hearing Research, 4*(3), 220–232. https://doi.org/10.1044/jshr.0403.220
Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34*(1), 111–117. https://doi.org/10.1177/001316447403400115
Kass, R. A., & Tinsley, H. E. A. (1979). Factor analysis. *Journal of Leisure Research, 11*(2), 120–138. https://doi.org/10.1080/00222216.1979.11969385
Kerschensteiner, M., Poeck, K., & Brunner, E. (1972). The fluency–nonfluency dimension in the classification of aphasic speech. *Cortex, 8*(2), 233–247. https://doi.org/10.1016/S0010-9452(72)80021-2
Kertesz, A. (1982). *Western Aphasia Battery*. Grune & Stratton.
Kertesz, A. (1993). Clinical forms of aphasia. *Acta Neurochirurgica*, 52–58. https://doi.org/10.1007/978-3-7091-9239-9_9
Kertesz, A. (2006). *Western Aphasia Battery–Revised*. Pearson. https://doi.org/10.1037/t15168-000
Kertesz, A., & Phipps, J. B. (1977). Numerical taxonomy of aphasia. *Brain and Language, 4*(1), 1–10. https://doi.org/10.1016/0093-934X(77)90001-3
Kline, P., & Barrett, P. (1983). The factors in personality questionnaires among normal subjects. *Advances in Behaviour Research and Therapy, 5*(3–4), 141–202. https://doi.org/10.1016/0146-6402(83)90017-6
Lacey, E. H., Skipper-Kallal, L. M., Xing, S., Fama, M. E., & Turkeltaub, P. E. (2017). Mapping common aphasia assessments to underlying cognitive processes and their neural substrates. *Neurorehabilitation and Neural Repair, 3*(15), 442–450. https://doi.org/10.1177/1545968316688797
MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3rd ed.). Erlbaum.
Manning, M., & Franklin, S. (2016). Cognitive grammar and aphasic discourse. *Clinical Linguistics & Phonetics, 30*(6), 417–432. https://doi.org/10.3109/02699206.2015.1128981

Murray, L. L. (2012). Attention and other cognitive deficits in aphasia: Presence and relation to language and communication measures. *American Journal of Speech-Language Pathology, 21*(2), S51–S64. https://doi.org/10.1044/1058-0360(2012/11-0067)

Mayer, J. F., & Murray, L. L. (2003). Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology, 17*(5), 481–498. https://doi.org/10.1080/02687030344000148

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Nozari, N., & Faroqi-Shah, Y. (2017). Investigating the origin of nonfluency in aphasia: A path modeling approach to neuropsychology. *Cortex, 95,* 119–135. https://doi.org/10.1016/j.cortex.2017.08.003

Olness, G. S., & Ulatowska, H. K. (2011). Personal narratives in aphasia: Coherence in the context of use. *Aphasiology, 25*(11), 1393–1413. https://doi.org/10.1080/02687038.2011.599365

Pohlmann, J. T. (2004). Use and interpretation of factor analysis in *The Journal of Educational Research*: 1992–2002. *The Journal of Educational Research, 98*(1), 14–22. https://doi.org/10.3200/JOER.98.1.14-23

Porch, B. E. (1967). *The Porch Index of Communicative Ability: Administration, scoring, and interpretation*. Consulting Psychologists.

Powell, B., Clark, E., & Bailey, S. (1979). A cluster-analysis of Schuell test profiles. *The British Journal of Disorders of Communication, 14*(2), 111–122. https://doi.org/10.3109/13682827909011351

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

R Development Core Team. (2020). *R: A language and environment for statistical computing* (Version 3.6.3). R Foundation for Statistical Computing. http://www.R-project.org

Revelle, W. (2019). *psych: Procedures for personality and psychological research* (Version 1.9.12.31). Northwestern University.

Revelle, W. (2020). *How to: Use the psych package for factor analysis and data reduction*. https://cran.r-project.org/web/packages/psychTools/vignettes/factor.pdf

Rochon, E., Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72*(3), 193–218. https://doi.org/10.1006/brln.1999.2285

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2018). tidyLPA: An R package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software, 3*(30), 978. https://doi.org/10.21105/joss.00978

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

Schuell, H., Jenkins, J. J., & Carroll, J. (1962). A factor analysis of the Minnesota test for differential diagnosis of aphasia. *Journal of Speech and Hearing Research, 5*(4), 349–369. https://doi.org/10.1044/jshr.0504.349

Snowden, D., Kemper, S., Mortimer, J. T., Greiner, L. H., Wekstein, D. R., & Marksbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *Journal of the American Medical Association, 275*(7), 528–532. https://doi.org/10.1001/jama.275.7.528

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

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

Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Allyn & Bacon.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29. https://doi.org/10.1037/h0071663

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. https://doi.org/10.1016/j.bandl.2016.08.004

Trupe, E. H. (1984). *Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification*. Paper presented at the Clinical Aphasiology Conference, Baltimore, MD, United States.

Vermeulen, J., Bastiaanse, R., & Van Wageningen, B. (1989). Spontaneous speech in aphasia: A correlational study. *Brain and Language, 36*(2), 252–274. https://doi.org/10.1016/0093-934X(89)90064-3

Wagenaar, E., Snow, C., & Prins, R. (1975). Spontaneous speech of aphasic patient: A psycholinguistic analysis. *Brain and Language, 2,* 281–303. https://doi.org/10.1016/S0093-934X(75)80071-X

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

Wright, H. H., & Capilouto, G. J. (2012). Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology, 26*(5), 656–672. https://doi.org/10.1080/02687038.2012.676855

Wright, H. H., Koutsoftas, A., Fergadiotis, G., & Capilouto, G. (2010). Coherence in stories told by adults with aphasia. *Procedia—Social and Behavioral Sciences, 6,* 111–112. https://doi.org/10.1016/j.sbspro.2010.08.056