ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Narrative Discourse Predictors of Response to Naming Intervention in Aphasia

Dirk B. den Ouden,[a,b] Laura Giglio,[a] Sigfus Kristinsson,[a] Leonardo Bonilha,[a] Deena Schwen Blackett,[c] Brielle C. Stark,[d] Janina Wilmskoetter,[e] and Julius Fridriksson[a]

[a] University of South Carolina, Columbia [b] Chapman University, Irvine, CA [c] University of Central Florida, Orlando [d] Indiana University, Bloomington [e] Medical University of South Carolina, Charleston

ARTICLE INFO

Article History:
Received June 27, 2025
Revision received January 12, 2026
Accepted March 24, 2026

Editor-in-Chief: Julie A. Washington
Editor: William S. Evans

https://doi.org/10.1044/2026_JSLHR-25-00499

ABSTRACT

Purpose: While aphasia treatment studies commonly use picture-naming performance as an outcome measure, narrative discourse better reflects functional language use. Discourse variables may also hold prognostic value for naming treatment response, but their predictive role remains underexplored.
Method: We analyzed baseline and posttreatment narrative discourse samples from 95 chronic stroke survivors with aphasia enrolled in a lexical retrieval intervention study. Participants received 3 weeks each of phonological and semantic naming therapy in a crossover design. The Cinderella story retells were analyzed for a range of discourse features: mean length of utterance, words per minute, verbs per utterance, propositional density, type–token ratio, core lexicon, main concepts analysis, and error ratios. We used univariate (generalized) binomial and linear mixed-effects modeling with multiple predictors to assess whether baseline discourse variables predicted naming gains on the Philadelphia Naming Test and whether discourse variables themselves changed following treatment.
Results: Higher aphasia severity and lower baseline propositional density predicted greater naming gains across time points and treatment types. Gains after phonological therapy were also predicted by higher baseline core lexicon production. Posttreatment discourse showed gains in mean length of utterance, words per minute, and core lexicon, which were maintained at 6 months, while phonological errors declined only at follow-up. Phonological treatment led to increases in words per minute.
Conclusions: Discourse variables reflecting propositional efficiency and lexical appropriateness uniquely predict treatment response beyond general aphasia severity. Lexical retrieval therapy generalizes to improvements in narrative discourse, underscoring the clinical value of incorporating discourse-level measures in aphasia assessment and treatment outcome evaluation.
Supplemental Material: https://doi.org/10.23641/asha.32568822

An important target for aphasia research is to identify prognostic factors for response to treatment and recovery in people with aphasia (PWA; e.g., The REhabilitation and recovery of peopLE with Aphasia after StrokE [RELEASE] Collaborators, 2021). In a relatively large clinical trial, we previously assessed several biographical, neurological, and cognitive–linguistic variables that may predict response to sequential phonological and semantic naming treatment, identifying baseline (BL) aphasia severity and age as primary predictors of response to naming treatment (Kristinsson et al., 2023). In a subsequent analysis of the data from this same cohort, we are now turning our attention to narrative discourse data. To what extent do discourse variables have prognostic value for response to naming treatment, and do phonological or semantic naming interventions affect discourse production at the group level?

Analyzing aphasic discourse offers valuable insight into real-life contextual communication by PWA. Unlike isolated language tasks, discourse analysis examines language in use, capturing the system in action. However, this approach

Correspondence to Dirk B. den Ouden: denouden@chapman.edu.
Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

inherently involves numerous linguistic and extralinguistic variables that are challenging to define and control (Armstrong, 2000). We focus on a subset of all the possible discourse variables that can be easily extracted from transcribed samples and analyzed relatively objectively, that is, without relying on high levels of subjective interpretation.

Discourse-level communication is a central rehabilitation goal for PWA, yet treatment-related improvements in connected speech remain inconsistent and theoretically underspecified. Although impairment-focused therapies, particularly anomia treatments such as semantic feature analysis (SFA; Boyle & Coelho, 1995), reliably improve single-word retrieval, generalization to discourse has historically been modest and highly variable (Cameron et al., 2006; Peach & Reuter, 2010; Silkes et al., 2021). Large-scale modeling confirms that confrontation naming is strongly related to discourse informativeness but that substantial unexplained variance remains, underscoring that discourse reflects more than lexical retrieval alone (Fergadiotis et al., 2019). Systematic reviews further demonstrate that discourse treatment is characterized by heterogeneity in outcome measures, theoretical rationales, and levels of linguistic targeting, with little consensus on how word-level change translates into functional communication. A recent multilevel analysis of SFA outcomes similarly shows minimal and inconsistent effects on standard discourse measures and only weak moderation by nonlinguistic cognitive factors (in particular, cognitive flexibility and problem-solving abilities) in individual speakers, highlighting the potential need for alternative discourse-derived predictors of treatment response (Cavanaugh et al., 2024). In that study, no changes were observed on percentage of semantic errors, lexical diversity, or grammatical complexity (as measured through mean length of utterance [MLU]) in discourse tasks, and only a small change was observed in "informativeness" (number of correct information units divided by the total number of words; Brookshire & Nicholas, 1994).

A variety of methods or prompts can be used to elicit speech in PWA, such as picture description, personal experiences, multiperson conversations, procedural-task descriptions, narrative discourse elicitation, and others. These methods each have their own merits, and the associated discourse samples have different characteristics, reflecting different types of strengths and weaknesses in language production (Armstrong, 2000). Here, we analyzed data from narrative discourse samples elicited by asking participants to retell the Cinderella story (Saffran et al., 1989). In a large global survey of speech pathologists and aphasiology researchers conducted by Stark, Bryant, et al. (2023), 70% of responders rated the Cinderella elicitation method as "valid for collecting and generating spoken discourse samples for analysis." Advantages of the method are that the story is relatively well known, particularly in the North American sampling space of the present study, that it has certain basic story elements that are consistent and thus allow for comparison between participants as well as time points (i.e., the story itself does not change), that the story itself has sufficient content to elicit a representative sample of a speaker's narrative abilities, and that naming/listing of elements is avoided because no visual stimuli can be relied on during the narrative speech production itself. Disadvantages of the Cinderella method are that it taxes memory resources, the story itself has no direct functional relevance to PWA, some speakers may consider it childish, it has a western-hemispheric cultural bias, and the repetitive nature of having to retell the story multiple times is not motivational (in case of repeated-measures experiments or to monitor progression clinically).

More in general, narrative discourse can be problematic as a diagnostic or outcome measure in aphasia because performance varies substantially with elicitation genre and task demands, so differences (or treatment-related change) may partly reflect the task rather than language ability (Bliss & McCabe, 2006). Elicited discourse sample lengths also vary in aphasia, while many discourse indices are sensitive to length or require truncation/standardization, which may reduce comparability and, in some cases, lead to loss of clinically informative data (Fergadiotis & Wright, 2011; Prins & Bastiaanse, 2004). Finally, because narrative outcomes blend multiple linguistic and nonlinguistic influences, they can have limited specificity for isolating mechanisms of impairment or treatment effects, a point emphasized in broader reviews of discourse analysis in aphasia (Bryant et al., 2016).

Despite these challenges, discourse samples do reflect more naturalistic language use than what can be obtained from laboratory language production tasks, and their elicitation is less prone to some inherent downsides, such as the associated cost for certain task batteries, available testing time, and challenges with task comprehension and compliance that may lead to under or overestimation of patient abilities. More or less, unconstrained speech is relatively easy to elicit even in clinical practice, and rapid developments in automated text analysis enable streamlined, consistent, and objective extraction of target variables. Therefore, if the information that can be gathered from discourse analysis is shown to have prognostic value for response to aphasia treatment, this is of practical use to the field. For our understanding of the relationship between language components, it is also of interest to assess which aspects of discourse, if any, predict response to particular treatment approaches.

A systematic review by Dipper et al. (2021) has already shown some evidence for beneficial effects of discourse treatment itself, that is, treatment that is specifically focused on discourse production (see also DeDe & Hoover, 2021). Here, we look at generalization from lexically focused treatment to discourse output, as well as to the prognostic value of discourse variables for response to such treatment. As mentioned above, there is not much evidence, so far, of

generalization from non–discourse-focused treatment to discourse measures (e.g., Cavanaugh et al., 2024; Silkes et al., 2021). Therefore, the retrospective analysis in the current study casts a relatively wide net to include a range of discourse measures that each have their own theoretical motivation for why they might respond to lexically focused intervention. It includes measures such as propositional density (PD), which indexes the density of semantic–syntactic relations rather than lexical accuracy alone, and core lexicon (CL), which captures production of story-relevant content. These offer a means of quantifying how efficiently speakers organize and transmit information in discourse, precisely the level at which functional communication success is realized.

Other biographical as well as test-based prognostic variables have been identified in recent aphasia intervention studies. More advanced patient age is a negative prognostic factor for recovery in general, as well as for response to interventions, possibly accounted for by age-related decreased neuroplasticity and nonmorbid (normal) cognitive decline or its exacerbation after stroke (Kristinsson et al., 2022). For response to naming therapy, pretreatment naming performance and general aphasia severity also predict changes in naming abilities after treatment: Speakers with more severe aphasia tend to respond less positively to intervention (Billot et al., 2022; Braun & Kiran, 2022; Kristinsson et al., 2023; Scimeca et al., 2024). Although Kristinsson et al. (2023) show that this relation holds when linearly modeled, it was noted that the least severely aphasic participants did not show large changes in naming abilities either, likely related to inherently smaller effect sizes in this group but potentially also to a ceiling level of poststroke performance that these patients may have reached during their recovery. Biographical, cognitive–linguistic, and other behavioral variables may also interact with neurological predictors of recovery and treatment response. Therefore, it is expected that combinations of multivariate and multimodal models will ultimately provide the most accurate prognostic tools. However, in the current article, we concentrate on narrative discourse measures as predictors of treatment response, leaving related neurological factors for another occasion.

In addition, even for aphasia intervention approaches that do not specifically target discourse, it remains highly relevant to assess the effects of treatment on relatively unconstrained speech production. Discourse production is argued to reflect "functional" language use to a greater extent than less naturalistic laboratory assessments, such as naming tasks. Thus, for the assessment of generalization of trained abilities to functional language use, it makes sense to turn to discourse analysis. Previous investigations into longitudinal changes to discourse production in aphasia have noted the potential complexity of such patterns, where various discourse variables may interact to create individual patterns of recovery (or decline) that may not be immediately straightforward to interpret (Prins et al., 1978; Ulatowska et al., 1981). Although such individual patterns are important to disentangle, the present study retains its focus on the group level, leveraging a large participant sample and multiple data points, as well as the simultaneous collection of primary outcome data after a language intervention aimed at lexical output. We specifically address the question of the extent to which lexical retrieval training yields generalizing effects on the quality of narrative discourse.

As potential predictors among the many variables that can be distilled from narrative discourse analysis, we selected both microlevel characteristics, reflecting internal linguistic details of the discourse sample, as well as intermediate and macrolevel variables that also reflect the more global structure and content of the sample (cf. Sherratt, 2007). As pointed out by Sherratt (2007) micro- and macrolevels may interact and rely on one another, and their correlation or independence can provide a window on the level of breakdown in PWA, as it is possible to convey globally appropriate content with or without intact microlevel structure, and it is also possible to produce relatively intact grammatical and lexical structure, without conveying the appropriate macrolevel content for the target topic. These variables were therefore selected because each reflects a different aspect of successful discourse production (Stark, Alexander, et al., 2023). Most can be extracted automatically from the discourse transcript, with the only requirement being a human coding of utterance boundaries (MacWhinney et al., 2010).

MLU, intermediate between micro- and macrolevels, reflects both speech fluency and syntactic complexity; PWA who are less fluent will typically not speak in long utterances. Although long utterances may consist of conjunctions (. . . and . . . and . . . etc.) that may not necessarily be syntactically complex, it is more likely for longer utterances to include embedded structures.

The macrolevel number of words per minute (WpM) reflects speech fluency directly. PWA generally speak much slower than typical adults (Nicholas & Brookshire, 1993), and Boyle (2014) estimates the minimal detectable change to be nine WpM, as a measure of improvement beyond noise.

The intermediate-level number of verbs per utterance (VpU) was included as a proxy for syntactic complexity, more directly reflecting sentence structure than MLU. A greater number of verbs within each utterance implies a generally richer structure that includes embedded clauses (cf. Thorne & Faroqi-Shah, 2016).

The macrolevel PD reflects the number of propositions made relative to the total number of words in the sample (Bryant et al., 2013). Although it has been used as a proxy for lexical-semantic informativeness (e.g., Bryant et al., 2013), Webster and Morris (2019) report that PD in

PWA was only weakly related to listener ratings of informativeness; the strongest associations they found with perceived informativeness were for the number of correct information units and the number of propositions, not for PD per se. The measure used here is based on the application by Brown et al. (2008), going back to the work of Kintsch (1974). Kintsch's notion of PD is approximated by counting the number of verbs, adjectives, adverbs, prepositions, and conjunctions (but not nouns), divided by the total number of words (MacWhinney, 2000; Snowdon et al., 1996). As such, PD primarily indexes semantic–syntactic complexity or the density of relational/functional material.

Type–token ratio (TTR) is a macrolevel measure of lexical diversity and straightforward to calculate (the number of different lexical items divided by the total number of lexical items), with the caveat that the measure has been noted to be susceptible to variations in language-sample size (e.g., Fergadiotis & Wright, 2011). Various sophisticated computational approaches have been suggested to address this shortcoming (Covington & McFall, 2010; Fergadiotis et al., 2013), each with their own advantages and disadvantages but always inherently at the cost of the straightforward interpretability offered by TTR. Most closely resembling TTR, the moving-average TTR reduces sample-length effects in neurologically intact speakers (Covington & McFall, 2010). However, we opted to retain the traditional TTR here, because its sensitivity to discourse length may not necessarily be a limitation in aphasic speech. Lexical diversity in aphasia is commonly conceptualized as reflecting global efficiency of lexical access and retrieval across an entire discourse sample, rather than a locally stationary stylistic property (Fergadiotis & Wright, 2011). In aphasia, failures of lexical retrieval are often clustered, discourse production is frequently curtailed, and speakers may show progressive lexical exhaustion or increased reliance on high-frequency forms as discourse unfolds. Measures such as MTTR, which average lexical diversity across local sliding windows, may therefore attenuate construct-relevant variance by normalizing away global scarcity effects. In addition, MTTR has been shown to be sensitive to the selected window size (Covington & McFall, 2010; Fergadiotis et al., 2013), adding a level of uncertainty to the measure and thereby reducing its direct interpretability. Still, it must be noted that, as a ratio, TTR can be influenced both by an increase in the total number of words produced (which would lead to a lower TTR), as well as by an increase in the variety of words produced (which would lead to a higher TTR). Therefore, the qualitative interpretation of any changes to TTR is less straightforward than for some of the other discourse variables, and this measure is best considered in the context of other variables.

The macrolevel CL score is a measure of *lexical appropriateness*, in that it reflects the use of lexical items that are appropriate to the specific context of the discourse elicitation, in this case, the Cinderella story. Therefore, this measure is context dependent and can only be applied to discourse types for which the CL has been established based on normative data from unimpaired speakers (Kim et al., 2019).

The other variables required human coding. Phonological, semantic, and unrelated error types are microlevel variables that reflect lexical accuracy at different functional levels of production. The macrolevel main concept analysis (MCA) score reflects the accuracy of the concepts that are brought up in the discourse sample (Kong et al., 2016; Richardson & Dalton, 2016). As such, like the CL score, it is specific to the particular context of the elicitation method. The MCA score serves as a proxy for the macrolevel informativeness or conceptual appropriateness of the discourse sample, a crucial component of functional communication. Arguably, the effects of treatment generalizing to any of these variables may reflect improvements that are functionally relevant to PWA and their daily language use.

Although we did not formulate specific a priori hypotheses about the relations between (semantic vs. phonological) lexical intervention and generalization to particular discourse variables, we did choose to focus on variables that could reasonably be expected to be affected by facilitation of lexical retrieval. This expectation is less strong for microlevel variables reflecting fine-grained morphosyntactic markers (such as verb inflections for tense and number/person agreement), while these also require a separate and more complex level of analysis. For that reason, such microlevel variables were not included in the analysis presented here.

The present study had three aims: (a) to assess the relation between narrative discourse variables and naming abilities at BL, (b) to identify narrative discourse variables that have predictive value for response to naming treatment, and (c) to assess the effects of phonological versus semantic naming treatment on narrative discourse production. We have focused on variables that require minimal manual coding, with the exception of manually marking utterance endings, error types, and main concepts.

Method

Participants

A total of 107 stroke survivors with aphasia in the chronic stage (more than 6 months poststroke) participated in a study on the identification of predictors of response to treatment focused on lexical retrieval (Predicting Outcomes of Language Rehabilitation [POLAR]). The study was conducted at the University of South Carolina and the Medical

University of South Carolina. Institutional review boards at both universities approved all study procedures (University of South Carolina Protocol No. Pro00105675, Medical University of South Carolina Protocol No. Pro00058579), and informed consent was obtained from participants before study enrollment. We refer to Kristinsson et al. (2023) for full details on the study and its participants. For 12 participants, we were missing either BL naming data or narrative discourse samples at BL, due to technical issues with the recording equipment or human error in applying the correct settings. The current report is based on the 95 participants for whom we were able to collect both naming data and narrative discourse samples at BL. See Table 1 for demographic details of the study sample.

Treatment Study (POLAR)

Participants received a total of 6 weeks of therapy aimed at lexical retrieval abilities. Through randomized assignment, approximately half of the group (n = 52) received 3 weeks of phonologically focused treatment, while the other half (n = 43) received 3 weeks of semantically focused treatment, followed by a 2-week break before switching to the other type of treatment. Each 3-week treatment phase consisted of a total of 15 hr of treatment, spread over 15 sessions on different days (Kristinsson et al., 2023). Different training stimuli were used in each phase, consisting of 50 nouns and 10 verbs (for a total of 120 training stimuli). "Phonological" treatment targeted the retrieval of word forms and consisted of a combination of phonological components analysis (Leonard et al., 2008), a phonological production training task, and a phonological judgment training task. "Semantic" treatment targeted lexical content and consisted of a combination of SFA (Boyle, 2017; Boyle & Coelho, 1995), a semantic barrier task, and Verb Network Strengthening Treatment (Edmonds, 2014; Edmonds et al., 2009).

The primary outcome measure for the POLAR treatment study was change in picture-naming performance relative to BL, quantified as raw change on the Philadelphia Naming Test (PNT; Roach et al., 1996) and assessed after the first treatment phase (Tx1), immediately before the start of the second phase (ITBL), immediately after the second

Table 1. Baseline demographics and clinical characteristics of the study sample (N = 95).

Variable	Range	*M*/count	*SD*
Age (years)	29–80	60.6	11.1
Female		39 (41%)	
Male		55 (59%)	
Race			
Black		20 (21%)	
White		74 (79%)	
Handedness			
Right		82 (87.2%)	
Left		11 (11.7%)	
Ambidextrous		1 (1.1%)	
Education (years)	12–20	15.5	2.3
Time since stroke onset (months)	10–241	49.8	52.4
NIH Stroke Scale score	0–18	6.39	3.98
WAB-R Aphasia Quotient	14.5–93.1	60.9	22.4
Baseline PNT correct	0–99%	46.3%	34.2%
Apraxia of speech (binary)[a]		52 (55.3%)	
ASRS apraxia of speech severity	0–4	1.55	1.54
Expressive agrammatism (binary)[b]		15 (16%)	
Aphasia type by WAB-R			
Anomia		28 (29.7%)	
Broca's		39 (41.5%)	
Conduction		15 (16.0%)	
Global		4 (4.3%)	
Transcortical motor		1 (1.1%)	
Wernicke's		7 (7.4%)	

Note. NIH = National Institutes of Health; WAB-R = Western Aphasia Battery–Revised; PNT = Philadelphia Naming Test; ASRS = Apraxia of Speech Rating Scale.

[a]Based on clinical judgment and ASRS score. [b]Based on clinical judgment and morphosyntactic quality of discourse.

treatment phase (Tx2), 1 month after treatment completion (1mo), and 6 months after treatment completion (6mo). None of the training stimuli were part of the PNT. The first 45 participants completed two PNTs within the same week at each time point, for a total of 12 PNTs. For these participants, we used the first PNT at each time point in the present analysis, unless the first administration was incomplete, in which case we used the second administered test (eight cases). In one participant, half of the PNT was administered in the first session at each time point; and the other half, in the second session at each time point, so we consolidated these data into a single PNT item set for each time point. Likewise, for two other participants, the BL PNT assessment was spread out over two separate days within the same week, so we consolidated these data into single BL PNT sessions for the analyses.

All treatment was delivered by certified and experienced speech-language pathologists who also administered all behavioral outcome measures. Transcription and scoring of all behavioral outcome measures were performed in randomized order by trained graduate student assistants who were blinded to the participants' treatment phases and time points.

Narrative Discourse Procedure and Variables

At the same six time points as the PNT collection, participants were asked to (re)tell the Cinderella story as a sample of their narrative discourse abilities. Following the standard procedure for this test (see MacWhinney et al., 2011), participants were shown a picture book of the Cinderella story, with the words covered up. They were asked to use the pictures to remind themselves of the story, which was confirmed to be familiar to all participants. After that, the picture book was removed, and participants were asked to recount the story, in as much detail as possible. All samples were video-recorded for subsequent transcription.

Transcription followed the Codes for the Human Analysis of Transcripts (CHAT) coding system, after which samples were analyzed using Computerized Language ANalysis (updated versions between 2016 and 2021) for automated extraction of discourse variables (MacWhinney, 2000). Consistent with the CHAT coding system, utterances were defined as "a main clause together with its related dependent clauses, adjuncts, and adverbial phrases." Conjoined independent clauses were treated as separate utterances. Dependent clauses (e.g., clauses that cannot stand alone due to the absence or elision of a subject) were not considered complete utterances. Thus, in the CHAT coding system, "She put on the dress and went to the ball" is coded as a single utterance, whereas "She put on the dress / and she went to the ball" is coded as two separate utterances. Trained graduate assistants manually coded word production errors as "phonological" (which included all nonword errors, as well as real-word errors phonologically related to the target), "semantic" (all real-word errors semantically related to the target), or "unrelated" (all real-word errors without either a phonological or semantic relation to the target). The calculation of CL scores was automated through a dedicated script (Cavanaugh et al., 2021a), based on previous work on the Cinderella story (Dalton et al., 2020). Likewise, to extract the MCA score, we used an automated script (Cavanaugh et al., 2021b; Richardson & Dalton, 2016) through which trained undergraduate assistants indicated which Cinderella story concepts were included in the samples and judged them for accuracy and completeness according to the MCA protocol.

Table 2 lists the BL group-level descriptive data for the discourse variables analyzed in the current article. Variables such as the noun/verb ratio or the open-class/closed-class ratio also reflect lexico-syntactic aspects of discourse and have been used to characterize speakers qualitatively. However, we chose not to include such ratios, in order to maximize the leverage of the large sample size offered by the POLAR study, as the use of ratios between independent variables can potentially lead to noncomputable values (e.g., in cases where a speaker does not produce any verbs or closed-class words). The presence of "NA" values in the data set limits the ability to use multiple-regression models to assess the effects of multiple variables simultaneously.

Statistical Analyses

Data Preprocessing

With 95 participants and six potential time points for each participant, we had a total of 570 potential observations,

Table 2. Baseline descriptive statistics for the discourse variables of interest.

Variable	Min	Max	*M*	*SD*
Duration (s)	13	1,142	228	190
Total no. of utterances	2	133	28.9	27.6
Total no. of words	3	817	187	175
MLU in words	1	13.6	5.97	2.86
Words per minute	2.63	166	50.1	34.6
Verbs per utterance	0	2.5	1.03	0.598
Propositional density	0	0.63	0.414	0.119
Type/token ratio	0.21	1	0.463	0.18
Ratio phonological errors	0	0.732	0.117	0.16
Ratio semantic errors	0	0.2	0.0135	0.0262
Ratio unrelated errors	0	0.2	0.0151	0.0295
Core lexicon score	0	77	29.1	21.2
MCA composite score	0	61	19.9	18.4

Note. MLU = mean length of utterance; MCA = main concepts analysis.

but since not every participant had both PNT and discourse data at each time point, the actual number of observations was 539 (time points missing: 5.4%: six at Tx1, two at ITBL, three at Tx2, six at 1mo, and 14 at 6mo). In addition to recording errors, missing data at time points after the BL session were sometimes due to scheduling conflicts or errors. All discourse variables were centered and scaled to *z* scores prior to statistical analysis.

Naming Accuracy Changes After Treatment

Because our sample of 95 PWA comprised a subsample of the participants included in Kristinsson et al. (2023), we also tested changes to naming accuracy relative to BL, to confirm the general treatment response reported in our previous work and to ensure that any discourse BL values or changes to discourse variables would be accurately associated with significant changes to naming accuracy, where relevant. For this, we fitted a binomial regression model with glmer, from the lme4 package (Version 1.1-31; Bates et al., 2015) in R (Version 4.2.2; R Core Team, 2024). This maximizes the power afforded by the available item-by-item naming responses. Binomial item-level naming accuracy was the dependent variable, as a function of a four-level factor time point, including the BL (base), posttreatment (Tx2), 1-month (1mo), and 6-month (6mo) time points. All level-specific effects of this factor were assessed relative to BL. Random intercepts were added for subjects and items. As a follow-up, we rotated all levels of the time-point variable to assess pairwise differences between the three posttreatment time points.

For this model, we also calculated percent accuracy at all time points, with the intercept functioning as the BL. Because the generalized linear mixed model was estimated using a logit link, model coefficients represent log odds. To obtain interpretable accuracy values, log odds were converted to predicted probabilities using the inverse-logit transformation, and probabilities were multiplied by 100 to yield percent accuracy. Predicted accuracy for each time point was calculated as the inverse-logit of the model intercept plus the corresponding time-point coefficient.

Predictive Value of BL Discourse Variables on BL Naming Accuracy

For the analysis that focused on PNT naming accuracy at BL as predicted by discourse variables, we again used a binomial regression with glmer, from the lme4 package (Version 1.1-31; Bates et al., 2015) in R (Version 4.2.2; R Core Team, 2024), with random intercepts for subjects and items. Western Aphasia Battery–Revised Aphasia Quotient (WAB-R AQ; also centered and scaled) was included in all these models, as our previous work had shown it to be a strong predictor of treatment response, and we wanted to assess the predictive value of discourse variables beyond "general" aphasia severity (Kristinsson et al., 2023). We assessed collinearity between independent variables by calculating variance inflation factors (VIFs) and identifying variables with VIF > 5. Variables with the highest value above this threshold were sequentially removed until no variables met the criterion. Probability values associated with the resulting model's effect sizes were calculated with the R package lmerTest (Version 1.1-3; Kuznetsova et al., 2017).

Predictive Value of BL Discourse Variables on Posttreatment Changes to Naming Accuracy

To assess the extent to which BL discourse variables predicted treatment response in terms of changes to naming accuracy relative to BL, we used separate binomial regressions following the same methodology as above, with item-level naming accuracy as the binomial dependent variable. For ease of interpretation, separate models were created for the three different outcome time points: immediately, 1 month, and 6 months posttreatment. Each model included a two-level time-point variable, allowing for comparison of the BL with the outcome time points (Tx2, 1mo, or 6mo). In these models, the effect of a specific discourse variable on changes to naming accuracy was reflected in its interaction with the two-level time-point variable. The random-effects structure for these models included crossed random intercepts for subjects and items. In addition, participants were assigned random slopes for time point, allowing the effect of time to vary across individuals. This structure accounts for both participant-level differences in BL accuracy and participant-level variability in sensitivity to the intervention while controlling for word-level variation in item difficulty.

To assess discourse predictors of treatment-specific response (phonological vs. semantic treatment), we used the participant-specific time points for PNT performance before and after the two respective treatment phases. As outlined above, for half of the participants, phonological treatment was administered in the first phase, followed by semantic treatment in the second phase, while for the other half, this order was reversed. For the Tx1, the pre–post time points were the BL and the time point immediately following the Tx1. For the second treatment phase, the pre–post time points were the time point immediately preceding (ITBL) and immediately following the Tx2. The statistical models for these outcome measures were the same as those outlined above. Earlier analyses of the POLAR data showed no signs of an effect of treatment order on naming outcomes (Kristinsson et al., 2021), so to maximize model simplicity, we did not add a treatment order factor to these analyses.

Effects of Treatment on Narrative Discourse

To assess the effect of treatment on our selected individual discourse variables, we conducted separate linear regressions for each discourse variable, using lme from the

lme4 package (Version 1.1-31; Bates et al., 2015) in R (Version 4.2.2; R Core Team, 2024), with random intercepts for subjects. As for the analysis of naming changes by time point, the independent variable in these models was always a four-level factor time, including the BL (base), posttreatment (Tx2), 1-month (1mo), and 6-month (6mo) time points. All level-specific effects of this factor were assessed relative to BL. We performed similar analyses for the immediate effects of phonological and semantic therapy. For these analyses, we used the participant-specific scores at the BL and Tx1 time points (Phase 1) and at the ITBL and Tx2 time points (Phase 2). We then modeled separate linear regressions for each discourse variable, with the two-level factor time (pre, post) and random intercepts for subjects, separately for the effects of phonological and semantic treatment. Because this entailed 10 separate analyses for each time point or treatment type, we applied Bonferroni corrections for multiple comparisons, yielding an alpha level of .005 for each of these analyses.

In a supplementary analysis, we assessed more directly whether effects on the discourse measures differed immediately after phonological versus semantic treatment. For this analysis, we again used the participant-specific scores between the BL and Tx1 time points (Phase 1) and between the ITBL and Tx2 time points (Phase 2), as above. We then modeled separate linear regressions for each discourse variable, with the factors treatment type (phonological, semantic) and time (pre, post), and random intercepts for subjects. This allowed us to assess the interaction between treatment type and change scores on individual discourse variables. As before, because this entailed 10 separate analyses (one for each discourse variable), we applied a Bonferroni correction for multiple comparisons, yielding an alpha level of .005 for these analyses.

Results

Note that in describing results from our regression models, we use the terms "predict" and "predictor" to refer to independent variables that are statistically associated with the dependent variable. This terminology does not necessarily imply a direct or unidirectional causal relationship. In our reporting of the results from the mixed-effects models, we have mostly followed recommendations from Meteyard and Davies (2020). Tables for the primary analyses include beta estimates, z or t values, confidence intervals, and p values for fixed effects, as well as variance and standard deviations for modeled random effects.

Naming Accuracy Changes After Treatment

The binomial regression model used to predict the likelihood of a correct response as a function of time point showed significantly higher rates of correct responses at all time points post-BL: Tx2 ($\beta = .297$, $SE = 0.061$, $z = 4.90$, $p < .001$), 1mo ($\beta = .325$, $SE = 0.056$, $z = 5.84$, $p < .001$), and 6mo ($\beta = .339$, $SE = 0.064$, $z = 5.27$, $p < .001$). Table 3 presents the full model outcomes. Releveling of the time-point factor revealed no differences among the three post-BL accuracy rates. Accuracy went from 33.6% at BL to 40.5% posttreatment, remaining at 41.6% and 41.2% at 1 and 6 months posttreatment, respectively.

Predictive Value of BL Discourse Variables on Naming Accuracy

For clarity and interpretation, Figure 1 first presents the multiple-correlation matrix for the BL discourse variables, WAB-R AQ, and naming performance on the PNT. Discourse variables largely correlate with one another, as well as with aphasia severity and naming scores, with the notable exception of the ratio of semantic errors and, to a lesser extent, unrelated errors. The latter is inversely correlated with WAB-R AQ and PNT and positively with TTR. Strong correlations are noted between WAB-R AQ and PNT scores ($r^2 = .81$) and between MLU and VpU ($r^2 = .79$).

In the full binomial regression model that tested discourse variables predicting naming accuracy at BL, there was initially high covariance between variables, with VIF scores above 5 for both MLU (VIF = 6.5) and VpU (VIF = 7.3). Removal of VpU, which had the highest VIF value, reduced the covariance, ensuring that no variables had VIF values above 3, indicating acceptable (low) collinearity. The initial collinearity proved to be the same for the other models predicting naming improvement, shown below, with removal of VpU solving all collinearity issues. Thus, all results reported below, on discourse predictors of naming improvement, are based on models from which VpU has been removed to resolve collinearity.

Based on the reduced model, significant predictors of BL naming accuracy were high WAB-R AQ ($\beta = 2.384$, $SE = 0.162$, $z = 14.73$, $p < .001$), low WpM ($\beta = -.323$, $SE = 0.149$, $z = -2.18$, $p < .05$), a low ratio of phonological errors ($\beta = -.413$, $SE = 0.192$, $z = -2.15$, $p < .05$), and a low ratio of unrelated errors ($\beta = -.301$, $SE = 0.153$, $z = -1.96$, $p < .05$). Table 4 presents the full model outcomes.

Predictive Value of BL Discourse Variables on Posttreatment Changes to Naming Accuracy

Immediately Posttreatment

We again observed the main effects on overall naming accuracy immediately posttreatment for WAB-R AQ, WpM, and the phonological error ratio, but of specific interest here were the interactions between BL discourse variables and the time-point variable. These interactions revealed that higher

Table 3. Model output for naming accuracy changes after treatment, relative to baseline.

Δ PNT accuracy							
Observations: 61,570; groups: subjects, 95; items, 176; marginal R^2/conditional R^2 .002/.743							
Fixed effects	**β estimate**	***SE***	***z***	***p***	**CI low**	**CI high**	**% Acc**
(Intercept)	–.678	0.300	–2.26	**.024**	–1.267	–0.089	33.6
Time: postTx2	.297	0.061	4.9	**< .001**	0.178	0.415	40.5
Time: 1mo	.325	0.056	5.84	**< .001**	0.216	0.435	41.2
Time: 6mo	.339	0.064	5.27	**< .001**	0.213	0.466	41.6
Random effects	**Effect**	**Variance**		***SD***	**corr**	**corr**	**corr**
Items	(Intercept)	0.822		0.907			
Subject	(Intercept)	8.040		2.836			
	Time: postTx2	0.159		0.399	0.20		
	Time: 1mo	0.120		0.347	0.41	0.93	
	Time: 6mo	0.178		0.422	0.35	0.60	0.70

Note. Significant *p* values are indicated in bold. PNT = Philadelphia Naming Test; CI = confidence interval; Acc = accuracy; PostTx2 = time point immediately following the second (final) treatment phase; 1mo = time point 1 month after the end of treatment; 6mo = time point 6months after the end of treatment; corr = correlation coefficient.

Figure 1. Correlation matrix for discourse variables, aphasia severity, and naming performance at baseline. Nonsignificant correlations ($p > .05$) are marked with crossed-out correlation coefficients. CL = core lexicon; unrel. errors = number of unrelated errors per word; sem. errors = number of semantic errors per word; phon. errors = number of phonological errors per word; PD = propositional density; VpU = verbs per utterance; TTR = type–token ratio; WpM = words per minute; MLU = mean length of utterance; WAB-R AQ = Western Aphasia Battery–Revised Aphasia Quotient; PNT = Philadelphia Naming Test; MCA = main concepts analysis.

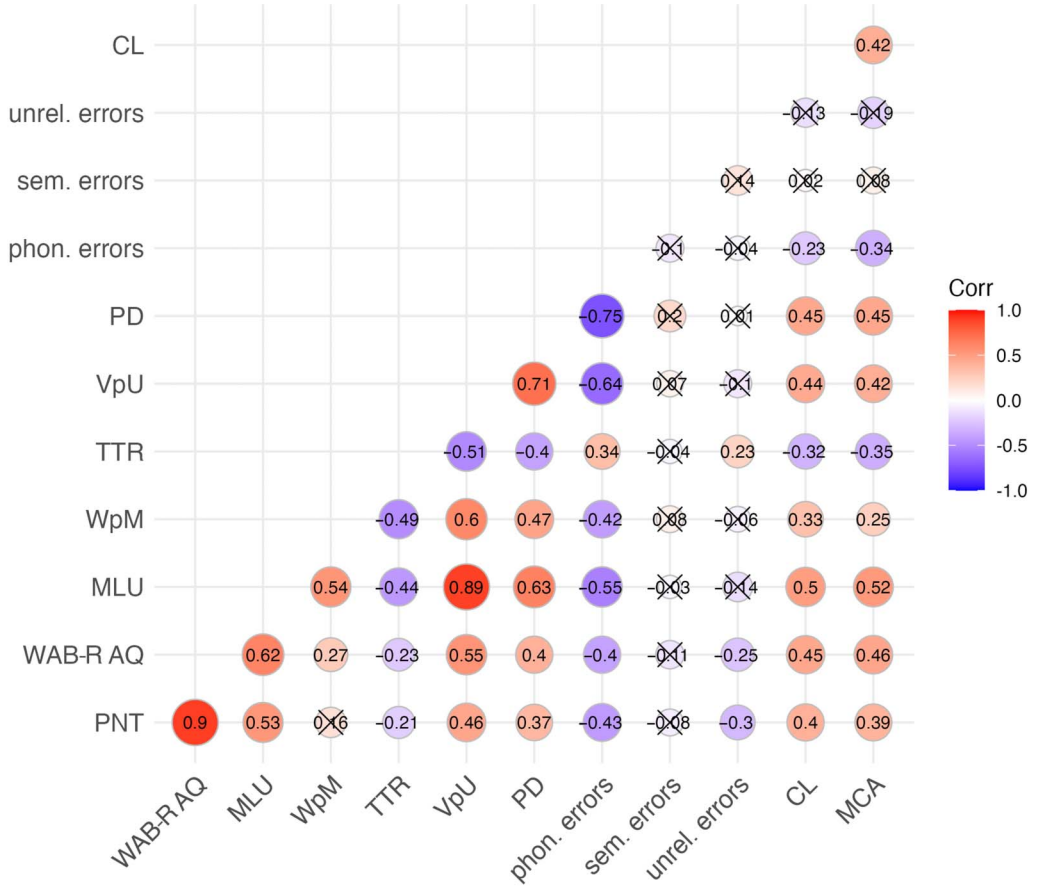


Table 4. Model output for predictors of Philadelphia Naming Test (PNT) naming accuracy at baseline.

PNT accuracy at baseline						
Obs.: 16,258; groups: subjects, 95; items, 176; marginal R^2/conditional R^2 .539/.708						
Fixed effects	**Estimate**	***SE***	***z***	***p***	**CI low**	**CI high**
(Intercept)	–0.640	0.136	–4.70	**< .001**	–0.907	–0.373
WAB-R AQ	2.384	0.162	14.73	**< .001**	2.066	2.701
MLU in words	–0.115	0.202	–0.57	.571	–0.511	0.282
Words per minute	–0.323	0.149	–2.18	**.030**	–0.614	–0.032
Propositional density	0.085	0.200	0.42	.671	–0.308	0.477
Type–token ratio	0.007	0.146	0.05	.962	–0.278	0.292
Ratio phonological errors	–0.413	0.192	–2.15	**.032**	–0.790	–0.037
Ratio semantic errors	–0.010	0.106	–0.10	.924	–0.218	0.198
Ratio unrelated errors	–0.301	0.153	–1.96	**.050**	–0.601	–0.001
Core lexicon	0.057	0.149	0.39	.699	–0.234	0.349
MCA composite score	–0.212	0.153	–1.39	.166	–0.511	0.088
Random effects	Effect	Variance	*SD*			
Subject	(Intercept)	1.065	1.032			
Items	(Intercept)	0.837	0.915			

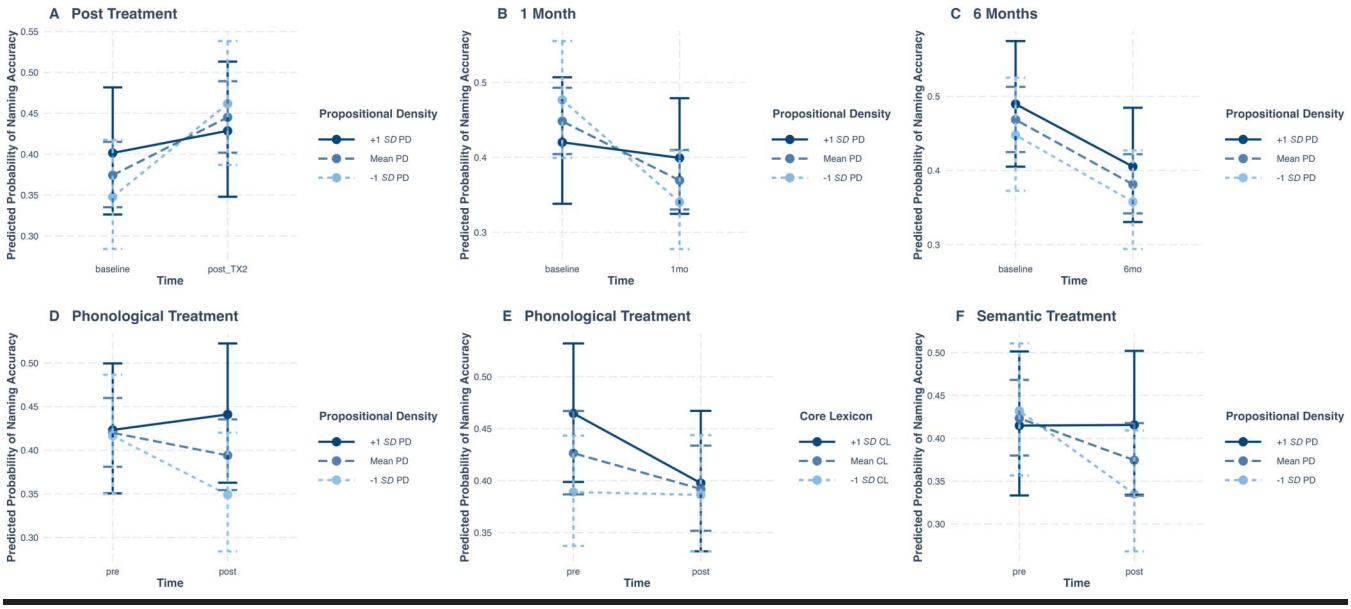
Note. Significant *p* values are indicated in bold. Obs. = observations; CI = confidence interval; WAB-R AQ = Western Aphasia Battery–Revised Aphasia Quotient; MLU = mean length of utterance; MCA = main concepts analysis.

BL WAB-R AQ (β = .211, *SE* = 0.087, *z* = 2.43, *p* < .05) and lower BL PD (β = −.182, *SE* = 0.088, *z* = −2.08, *p* < .05) were significant predictors of improved naming accuracy immediately post treatment. Figure 2A visualizes the predictive effect of BL PD on posttreatment improvement to naming accuracy. Supplemental Material S1 provides the full model outcomes.

1 Month Posttreatment

At 1 month posttreatment, we again observed the main effects on overall naming accuracy for WAB-R AQ, WpM, and phonological error ratio, as well as for the unrelated error ratio. Interactions between BL discourse variables and the time-point variable revealed that higher BL WAB-R AQ

Figure 2. Predictive effects of discourse variables on change in picture-naming accuracy from baseline, with baseline performance of the predictors plotted at –1 *SD*, 0, and +1 *SD*. (A) Baseline propositional density (PD) prediction of naming change immediately after treatment; (B) baseline PD prediction of naming change 1 month after treatment; (C) (absence of) baseline PD prediction of naming change 6 months after treatment; (D) baseline PD prediction of naming change after phonological treatment; (E) baseline core lexicon (CL) prediction of naming change after phonological treatment; (F) baseline PD prediction of naming change after semantic treatment.



(β = .254, *SE* = 0.078, *z* = 3.25, *p* < .05) and low BL PD (β = −.241, *SE* = 0.091, *z* = −2.65, *p* < .05) were significant predictors of improved naming accuracy immediately posttreatment (see Figure 2B). Supplemental Material S2 provides the full model outcomes.

6 Months Posttreatment

We again observed the main effects on overall naming accuracy at 6 months posttreatment for WAB-R AQ, WpM, phonological error ratio, and, to a lesser extent, for the unrelated error ratio. The only significant BL predictor of improved naming accuracy at 6 months posttreatment was higher BL WAB-R AQ (β = .218, *SE* = 0.094, *z* = 2.31, *p* < .05), without any additional predictive value offered by the discourse variables (for PD and the comparison to the interactions observed at the two previous posttreatment time points, see Figure 2C). Supplemental Material S3 provides the full model outcomes.

Phonologically Focused Treatment

Based on the reduced model, we again observed the main effects on overall naming accuracy for WAB-R AQ, WpM, and the unrelated error ratio. Interactions between BL discourse variables and the time-point variable revealed that lower BL PD (β = −.180, *SE* = 0.074, *z* = −2.43, *p* < .05) and higher BL CL values (β = .131, *SE* = 0.059, *z* = 2.23, *p* < .05) were significant predictors of improved naming accuracy after phonological treatment (see Figures 2D and 2E). Supplemental Material S4 provides the full model outcomes.

Semantically Focused Treatment

We again observed main effects on overall naming accuracy for WAB-R AQ and unrelated error ratio, as well as for MCA scores. The interactions between BL discourse variables and the time-point variable revealed that the only significant predictor of improved naming accuracy after semantic treatment was lower BL PD (β = −.208, *SE* = 0.081, *z* = −2.55, *p* < .05), as shown in Figure 2F. Supplemental Material S5 provides the full model outcomes.

In summary, posttreatment naming responses are most strongly and consistently predicted by lower BL PD, an effect that remains present 1 month after treatment but disappears after 6 months, at which point improved naming relative to BL is only predicted by milder aphasia at BL. Specific response to phonologically and semantically focused treatments is also predicted by lower BL PD. As is visible from Figure 2, part of the predictive effect of PD on naming improvement is driven by an association of low BL PD with relatively low BL naming, which then improves after treatment. Additionally, higher BL CL scores contributed to the prediction of a positive response to phonological treatment. Higher BL WAB-R AQ scores predicted treatment response at all time points, as well as after phonological therapy, but not after semantic therapy.

Effects of Treatment on Narrative Discourse

Table 5 shows the results of the linear regression models testing changes to discourse variables immediately posttreatment, at 1 month posttreatment, and at 6 months posttreatment, as well as separately in response to phonological and semantic treatment. Relative to BL, MLU increased immediately after the two treatment cycles (posttreatment; β = .247, *t* = 3.289, *p* < .005), and this effect was maintained at 1 month (β = .233, *t* = 3.103, *p* < .005) and 6 months (β = .225, *t* = 2.989, *p* < .005) posttreatment. Likewise, WpM increased after treatment (β = .195, *t* = 4.306, *p* < .005), and this increase was maintained at 1 month (β = .176, *t* = 3.889, *p* < .005) and 6 months (β = .181, *t* = 3.998, *p* < .005) posttreatment. TTR decreased posttreatment (β = −.226, *t* = −3.076, *p* < .005), and this effect was still significant at 1 month postonset (β = −.232, *t* = −3.115, *p* < .005) but no longer met the corrected statistical threshold at 6 months (β = −.175, *t* = −2.383, *p* = .018). By contrast, a reduction in phonological errors was only observed at 6 months posttreatment (β = −.13, *t* = −2.914, *p* < .005). Like MLU and WpU, CL increased immediately posttreatment (β = .28, *t* = 4.598, *p* < .005), and this increase was maintained at 1 month (β = .214, *t* = 3.947, *p* < .005) and at 6 months (β = .262, *t* = 4.305, *p* < .005) posttreatment. Weaker effects that did not meet the corrected statistical threshold were an increase in VpU immediately posttreatment (β = .166, *t* = 2.322, *p* = .021) and an increase in PD immediately posttreatment (β = .172, *t* = 2.404, *p* = .017), which was still visible at 1 month posttreatment (β = .187, *t* = 2.631, *p* = .01). Figure 3 visualizes the changes in the discourse variables, across time points.

In response to 3 weeks of phonological therapy, WpM increased significantly (β = .143, *t* = 3.508, *p* < .005), and we also observed a numerical increase in VpU that did not meet the corrected statistical threshold (β = .143, *t* = 2.25, *p* = .027). Likewise, CL marginally increased after both phonological therapy (β = .123, *t* = 2.259, *p* = .026) and semantic therapy (β = .079, *t* = 2.164, *p* = .033), but these effects did not meet the corrected threshold.

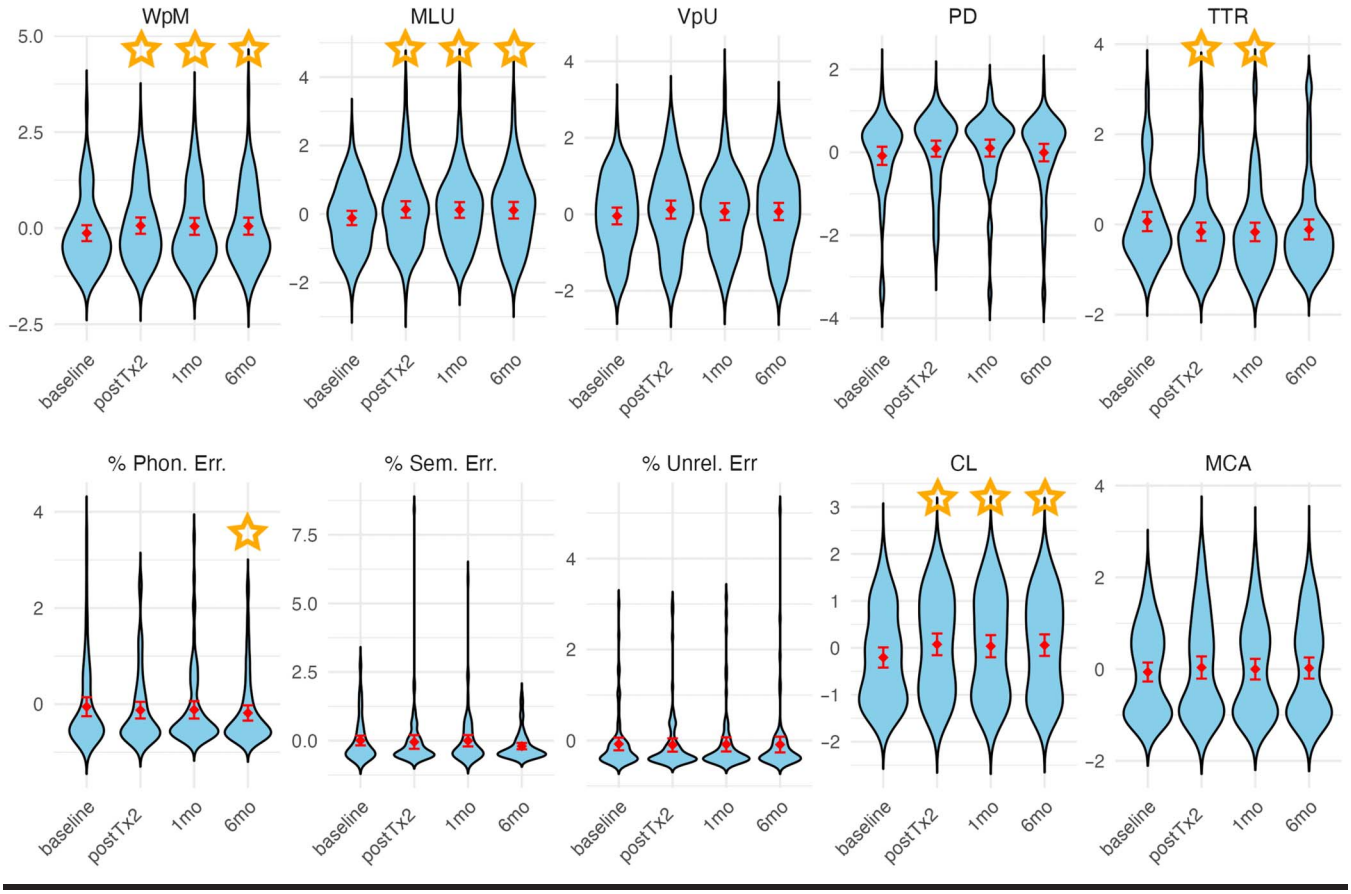
The supplementary analysis directly comparing response to phonological versus semantic therapy did not reveal interaction effects between time and treatment that survived Bonferroni correction. We did observe weak above-threshold interactions for WpM (β = −.140, *t* = −2.285, *p* = .023) and the ratio of phonological errors (β = −.158, *t* = −2.0739, *p* = .039). The nonsignificant interaction between time and treatment for WpM was driven by a numerically greater increase in WpM after phonological treatment than after semantic treatment, but for phonological errors, the nonsignificant interaction was driven by a numerical increase after phonological treatment, versus a numerical decrease after semantic treatment (see Supplemental Material S7).

Table 5. Output for linear regression models of generalization of treatment effects to narrative discourse variables, measured immediately posttreatment, at 1 month posttreatment, at 6 months posttreatment, and separately in response to phonological and semantic treatments.

	Post-TX			1 month			6 months			Phonological TX			Semantic TX		
Variable	**β**	***t***	***p***	**β**	***t***	***p***	**β**	***t***	***p***	**β**	***t***	***p***	**β**	***t***	***p***
MLU	.247	3.289	**.001**	.233	3.103	**.002**	.225	2.989	**.003**	.105	1.755	.083	.024	0.239	.812
VpU	*.166*	*2.322*	*.021*	.114	1.595	.112	.116	1.624	.106	*.143*	*2.25*	*.027*	–.02	–0.312	.756
WpM	.195	4.306	**< .001**	.176	3.889	**< .001**	.181	3.998	**< .001**	.143	3.508	**< .001**	.003	0.071	.943
PD	*.172*	*2.404*	*.017*	*.187*	*2.631*	*.010*	.078	1.094	.275	.14	1,863	.066	–.025	–0.322	.748
TTR	–.226	–3.076	**.002**	–.232	–3.155	**.002**	*–.175*	*–2.383*	*.018*	–.029	–0.403	.787	–.13	–1.974	.056
Sem. err.	–.052	–0.421	.674	–.008	–0.069	.945	–.204	–1.662	.098	–.06	–0.686	.494	.164	1.279	.204
Phon. err.	–.071	–1.602	.111	–.065	–1.448	.149	–.13	–2.914	**.004**	.074	1.225	.224	–.084	–1.608	.111
Unrel. err	–.022	–0.313	.754	–.008	–0.117	.907	–.012	–0.176	.860	–.078	–0.597	.552	.052	0.656	.514
CL	.28	4.598	**< .001**	.241	3.947	**< .001**	.262	4.305	**< .001**	*.123*	*2.259*	*.026*	*.079*	*2.164*	*.033*
MCA	.098	1.246	.214	.062	0.788	.413	.087	1.099	.273	–.03	–0.658	.512	.084	1.557	.123

Note. Effects that survive Bonferroni correction ($p < .005$) are displayed in bold font. Effects that pass only an uncorrected threshold ($p < .05$) are displayed in italics. TX = treatment; MLU = mean length of utterance; VpU = verbs per utterance; WpM = words per minute; PD = propositional density; TTR = type–token ratio; Sem. err. = number of semantic errors per word; Phon. err. = number of phonological errors per word; Unrel. err. = number of unrelated errors per word; CL = core lexicon; MCA = main concepts analysis.

Figure 3. Violin plots showing scaled values for discourse variables at different time points. Means and 95% confidence intervals are shown in red. Significant differences from baseline, after Bonferroni correction, are marked with a star. WpM = words per minute; MLU = mean length of utterance; VpU = verbs per utterance; PD = propositional density; TTR = type–token ratio; postTx2 = time point immediately following the second (final) treatment phase; 1mo = time point 1 month after the end of treatment; 6mo = time point 6 months after the end of treatment; % Phon. Err. = percentage of phonological errors per word; % Sem. Err. = percentage of semantic errors per word; % Unrel. Err. = percentage of unrelated errors per word; CL = core lexicon; MCA = main concepts analysis.



The full results of this analysis are shown in Supplemental Material S6.

Discussion

Posttreatment Naming Changes

Confirming the results reported for the full data set in Kristinsson et al. (2023), participants showed significant improvement in naming accuracy following intervention, with accuracy rates significantly higher at all post-BL time points (immediately posttreatment, 1 month, and 6 months). No differences were observed between these posttreatment time points, suggesting the maintenance of an immediate posttreatment improvement up to 6 months. An improvement over the previously reported results is that the present study applied an item-level logistic regression, whereas the previously published results were based on aggregated scores by participant. Naming accuracy, as measured with the PNT, was the primary outcome measure for the POLAR study from which these data were derived, and the PWA in the study were in the chronic phase poststroke, where task performance is typically relatively stable in the absence of specific interventions or practice. However, we reiterate that no control groups or multiple baselines were included, since the primary objective of the study was not to assess treatment efficacy but rather to identify individual predictors of response to treatment.

Relation Between Discourse Measures and Picture-Naming Performance

BL predictors of higher naming accuracy included less severe aphasia (WAB-R AQ), lower speech fluency (WpM), and higher ratios of phonological and unrelated errors in descriptive discourse. Given that narrative discourse production is relatively more naturalistic and functional than

object-picture naming, it is important to find evidence of the correlation between these two tasks. At the very least, this implies a level of convergent validity for using picture-naming accuracy as a measure of treatment outcome in aphasia. The present results suggest that when aphasia severity is kept constant, speakers with greater word-finding difficulty on a naming task may actually produce fewer lexical-level errors in their narrative discourse production. This may well be accounted for by these speakers' avoidance of challenging lexical targets in spontaneous speech production.

Interestingly, less fluent speech (i.e., lower WpM) was associated with better naming performance. It must be noted that fluency is also an important substrate of the WAB-R AQ, our measure of overall aphasia severity. It directly affects scores for spontaneous speech, which comprise 40% of the AQ score, so that PWA who produce less speech will have a lower WAB-R AQ, all else being equal (see also Kristinsson et al., 2025). By itself, WpM was not correlated with PNT scores and only weakly with WAB-R AQ ($r^2 = .07$, $p < .05$); however, in the full model that takes other variables into account, the independent association between nonfluent speech and better naming performance comes to light. This effect may reflect the notion that, given similar levels of aphasia severity, speakers who produce more words more rapidly may actually also be producing less lexically accurate speech. Similarly, if WAB-R AQ is kept constant, nonfluent PWA may be less likely than fluent PWA to produce errors on the PNT.

Somewhat surprising was the absence of a relation between PNT scores and the ratio of semantic errors produced in narrative discourse. This warrants a closer examination that takes PNT error types into account, which is outside the scope of the present article. We note that the ratio of semantic errors also did not correlate with other discourse variables, nor with overall aphasia severity. It is possible that the apparently exceptional role of semantic errors here is influenced by the relatively low numbers of this error type in our data. This may be related to our study sample, which was low on PWA who might be expected to produce larger numbers of semantic paraphasias, such as those with Wernicke's aphasia. However, only the difference between the ratios of phonological errors and the other error types is significant (semantic, $t(94) = 6.1063$, $p < .001$; unrelated, $t(94) = 6.0501$, $p < .001$), with no significant difference between semantic and unrelated error types, $t(94) = 0.4212$, $p = .675$.

BL Discourse Variables Predictive of Treatment Response

Importantly, posttreatment improvement in picture naming accuracy was consistently predicted by higher BL WAB-R AQ scores and, at earlier time points, by lower BL PD in discourse. Furthermore, response to phonologically focused treatment was predicted by both low PD and high CL scores, while semantic treatment effects were driven solely by low BL PD. These findings confirm the role of aphasia severity and highlight the additional role of macrolevel discourse variables as predictors of treatment response, with BL PD emerging as a robust and temporally sensitive predictor.

PD correlates weakly with PNT performance at BL ($r^2 = .14$, $p < .05$), but this relation is positive, as are its correlations with aphasia severity and all other discourse variables except TTR (negative correlation). However, high PD does not independently predict PNT when severity and other variables are taken into account in the full model (see above) and a lower BL PD turns out to be the strongest predictor of response to treatment as reflected in naming performance.

Across studies, PD is consistently lower in PWA than in controls and is related to severity in complex ways. Bryant et al. (2013) reported a positive correlation between WAB-R AQ and PD, such that more severe aphasia was associated with lower PD. Fromm et al. (2016) showed that PD is particularly low in Broca's aphasia and that PD best distinguishes Broca's (and, to some extent, transcortical motor) aphasia from other subtypes. Importantly, they found no linear association between PD and severity once aphasia type was taken into account. In the aging/dementia literature, low PD in early-life writing has repeatedly been interpreted as an index of reduced cognitive reserve and shown to predict later-life cognitive decline and Alzheimer pathology (cf. Snowdon et al., 1996).

Speculating further on the combination of mild aphasia and low PD, it is possible that the latter may in fact be a reflection of an adaptive mechanism. Although at first sight, higher PD in discourse may be considered a positive reflection of cognitive and linguistic abilities in the speaker, it must be noted that the PD measure by itself does not take into account the appropriateness of the propositions. Speakers with aphasia may also adapt their communicative strategies, for example, by using simpler structures or circumlocutions, to maintain intelligibility and convey intent (Armstrong, 2000). Where such adaptation leads to lower PD counts, this may actually positively reflect preserved pragmatic or executive abilities, rather than linguistic deterioration. The relative preservation of these abilities, then, particularly coupled with mild aphasia and the production of lexical targets that are appropriate to the topic, may be predictive of higher response to treatment. We note that this purely speculative account is no more than that, at this point, and would be subject to further hypothesis testing.

In all, these findings indicate that low PD may reflect nonfluent grammatical output (as observed in many speakers

with Broca's aphasia), reduced semantic/syntactic complexity, and, in other populations, lower cognitive reserve. As in Kristinsson et al. (2023), we deliberately did not include "aphasia type" as a predictor in the present study, since we aimed to focus on measurable individual behavioral variables rather than clusters of symptoms. However, we cannot rule out the possibility that the prognostic value of PD in our study is partly driven by good responders who are individuals with relatively mild nonfluent/Broca-like profiles and thus tend to have lower PD.

Response to phonological therapy is additionally strongest in PWA who produce high levels of lexical targets appropriate to the context. In our study, the production of story-specific key lexical items is captured separately by the CL measure, and our main interaction of interest for phonological treatment was precisely low PD plus high CL. Our data therefore suggest that participants with relatively sparse propositional structure but robust production of appropriate lexical targets benefit most from naming treatment.

Generalized Changes to Discourse Variables After Treatment

Although the treatment in the POLAR study focused on lexical retrieval and production, it was associated with broad improvements in discourse production. Immediately posttreatment, chronic PWA showed significant gains in MLU, WpM, and CL production, with all effects maintained through the 6-month follow-up. TTR decreased significantly through 1 month posttreatment, suggesting numerically increased lexical output (a greater number of words, relative to the number of different words), although this effect had attenuated by 6 months. Phonological error rates, by contrast, only decreased significantly at the 6-month mark. Weak numerical increases in verb production (VpU) and PD were observed posttreatment, though these did not meet corrected significance thresholds and should therefore be considered with extra caution. As before, these observations all suggest a generalizing effect of language intervention at the lexical level on the fluency (WpM, TTR), syntactic complexity (MLU and perhaps VpU), lexical accuracy (CL), and perhaps the informativeness (PD) of narrative discourse in aphasia.

Phonological treatment, which was focused on word forms over word meaning, significantly increased WpM, while semantic treatment did not yield significant generalizations to changes in discourse variables. A direct comparison between the effects of phonological versus semantic treatment did not reveal significant differences, so we emphasize that there was no evidence of a greater impact of phonological versus semantic treatment. Numerically, phonological treatment was associated with greater improvements in WpM than semantic treatment. Therefore, improved lexical access at the form level may have had a more substantive direct impact on the fluency of narrative speech production than improved access to semantic representations, but this is a speculation subject to more targeted testing.

Above the statistical threshold, there were weak numerical increases in VpU after phonological treatment, while both phonological and semantic therapies were nonsignificantly associated with CL gains. Phonological error ratios showed divergent effects, declining after semantic therapy but actually rising after phonological therapy. Especially given the overall positive changes to both naming itself as well as to other discourse variables, we are inclined to speculate that the increased phonological error rates may reflect a positive change as well, in particular reduced self-consciousness or error avoidance. PWA who are more self-confident and less anxious about their functional communication abilities are less prone to suppress or avoid making errors. This may lead to an increased number of word form errors, at least relative to other error types, but should still be considered a positive outcome that is relevant to functional communication and living with aphasia. Some support for this speculative account is found in the observation of a (nonsignificant) rise in the ratio of semantic errors after semantic treatment ($\beta = .082$, $z = 1.283$, $p = .199$), where the same mechanism may have been at play.

Limitations

The present study only considered story retelling as an elicitation method for narrative discourse, but previous studies have shown that discourse characteristics differ between genres (cf. Stark, 2019). Although we do maintain that narrative discourse likely forms a more appropriate reflection of functional communication abilities than picture description and it is more likely to elicit larger and therefore more representative speech samples than procedural discourse, all of these methods naturally reflect different language abilities and are therefore relevant to examine. In particular, we have focused solely on monodirectional language output, without considering conversational interactive speech, which places a greater emphasis on the pragmatic and improvisational aspects of discourse production and comprehension (Damico et al., 1999; Doedens & Meteyard, 2020).

We have necessarily focused on a limited set of discourse variables but have not taken into account others that may be of interest, such as microlevel variables reflecting morphosyntactic accuracy or more macrolevel variables such as cohesion (Ellis et al., 2005; Zhang et al., 2021). Such measures are of interest as well but typically do require more extensive human (and therefore subjective) coding than the variables we have investigated here. Our larger

research program is aimed not only at improving our understanding of the nature of aphasic language deficits but also at developing efficient automated measures that may have more direct clinical applications.

With respect to any changes to discourse measures after treatment, the same caveats apply as for the overall POLAR study from which these data are derived: No control conditions or multiple baselines were applied, so the possibility that changes are due to repetition effects cannot be ruled out. Again, however, discourse production was not trained, and all participants were in the chronic phase poststroke, where consistent treatment gains are typically not achieved through mere task repetition, especially if the task is applied weeks and months apart.

Conclusions

In a large sample of stroke survivors with different types and severity levels of aphasia, we have shown that discourse variables, particularly the macrolevel PD and, to a lesser extent, CL variables, offer modest, incremental prognostic value over severity and BL naming performance for the prediction of response to naming treatment. While aphasia severity remains the most consistent predictor across time, the added predictive power of discourse-level features underscores their relevance for treatment planning and outcome evaluation. Moreover, lexical retrieval therapy was found to yield generalizable improvements in spontaneous discourse, affecting fluency, syntactic complexity, and lexical accuracy. Phonological therapy aimed at the word form level was associated with generalized improvement in speech fluency (WpM). These findings support the integration of narrative discourse analysis into aphasia rehabilitation research and clinical practice, for both its diagnostic insight and its utility in tracking functional gains beyond isolated language tasks.

Author Contributions

Dirk B. den Ouden: Conceptualization, Supervision, Investigation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Laura Giglio:** Conceptualization, Writing – review & editing. **Sigfus Kristinsson:** Conceptualization, Investigation, Writing – review & editing. **Leonardo Bonilha:** Conceptualization, Writing – review & editing. **Deena Schwen Blackett:** Conceptualization, Writing – review & editing. **Brielle C. Stark:** Conceptualization, Data curation, Investigation, Writing – review & editing. **Janina Wilmskoetter:** Conceptualization, Data curation, Methodology, Project administration, Supervision, Writing – review & editing. **Julius Fridriksson:** Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Data Availability Statement

The data that support the findings of this study are available upon reasonable request from the corresponding author.

Acknowledgments

This study was supported by a research grant from the National Institute on Deafness and Other Communication Disorders, DC014664 (awarded to Julius Fridriksson, principal investigator).

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Billot, A., Lai, S., Varkanitsa, M., Braun, E. J., Rapp, B., Parrish, T. B., Higgins, J., Kurani, A. S., Caplan, D., Thompson, C. K., Ishwar, P., Betke, M., & Kiran, S. (2022). Multimodal neural and behavioral data predict response to rehabilitation in chronic poststroke aphasia. *Stroke, 53*(5), 1606–1614. https://doi.org/10.1161/STROKEAHA.121.036749

Bliss, L. S., & McCabe, A. (2006). Comparison of discourse genres: Clinical implications. *Contemporary Issues in Communication Science and Disorders, 33*(Fall), 126–167. https://doi.org/10.1044/cicsd_33_F_126

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research, 57*(3), 966–978. https://doi.org/10.1044/2014_jslhr-l-13-0171

Boyle, M. (2017). Semantic treatments for word and sentence production deficits in aphasia. *Seminars in Speech and Language, 38*(1), 52–61. https://doi.org/10.1055/s-0036-1597256

Boyle, M., & Coelho, C. (1995). Application of semantic feature analysis as a treatment for aphasic dysnomia. *American Journal of Speech-Language Pathology, 4*(4), 94–98. https://doi.org/10.1044/1058-0360.0404.94

Braun, E. J., & Kiran, S. (2022). Stimulus- and person-level variables influence word production and response to anomia treatment for individuals with chronic poststroke aphasia. *Journal of Speech, Language, and Hearing Research, 65*(10), 3854–3872. https://doi.org/10.1044/2022_jslhr-21-00527

Brookshire, R. H., & Nicholas, L. E. (1994). Test-retest stability of measures of connected speech in aphasia. *Clinical Aphasiology, 22,* 119–133. https://aphasiology.pitt.edu/163/

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40*(2), 540–545. https://doi.org/10.3758/brm.40.2.540

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

Bryant, L., Spencer, E., Ferguson, A., Craig, H., Colyvas, K., & Worrall, L. (2013). Propositional idea density in aphasic discourse. *Aphasiology, 27*(8), 992–1009. https://doi.org/10.1080/02687038.2013.803514

Cameron, R., Wambaugh, J., Wright, S., & Nessler, C. (2006). Effects of a combined semantic/phonologic cueing treatment on word retrieval in discourse. *Aphasiology, 20,* 269–285. https://doi.org/10.1080/02687030500473387

Cavanaugh, R., Dalton, S. G., & Richardson, J. (2021a). *coreLexicon: An open-source web-app for scoring core lexicon analysis* (R Package Version 0.0.1.0000). https://github.com/aphasia-apps/coreLexicon

Cavanaugh, R., Dalton, S. G., & Richardson, J. (2021b). *mainConcept: An open-source web-app for scoring main concept analysis* (R package version 0.0.1.0000). https://github.com/aphasia-apps/mainConcept

Cavanaugh, R., Dickey, M. W., Hula, W. D., Fromm, D., Golovin, J., Wambaugh, J., Fergadiotis, G., & Evans, W. S. (2024). Determinants of multilevel discourse outcomes in anomia treatment for aphasia. *Journal of Speech, Language, and Hearing Research, 67*(9), 3094–3112. https://doi.org/10.1044/2024_jslhr-24-00030

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Dalton, S. G. H., Kim, H., Richardson, J. D., & Wright, H. H. (2020). A compendium of core lexicon checklists. *Seminars in Speech and Language, 41*(1), 45–60. https://doi.org/10.1055/s-0039-3400972

Damico, J. S., Oelschlaeger, M., & Simmons-Mackie, N. (1999). Qualitative methods in aphasia research: Conversation analysis. *Aphasiology, 13*(9–11), 667–679. https://doi.org/10.1080/026870399401777

DeDe, G., & Hoover, E. (2021). Measuring change at the discourse-level following conversation treatment: Examples from mild and severe aphasia. *Topics in Language Disorders, 41*(1), 5–26. https://doi.org/10.1097/tld.0000000000000243

Dipper, L., Marshall, J., Boyle, M., Botting, N., Hersh, D., Pritchard, M., & Cruice, M. (2021). Treatment for improving discourse in aphasia: A systematic review and synthesis of the evidence base. *Aphasiology, 35*(9), 1125–1167. https://doi.org/10.1080/02687038.2020.1765305

Doedens, W. J., & Meteyard, L. (2020). Measures of functional, real-world communication for aphasia: A critical review. *Aphasiology, 34*(4), 492–514. https://doi.org/10.1080/02687038.2019.1702848

Edmonds, L. A. (2014). Tutorial for Verb Network Strengthening Treatment (VNeST): Detailed description of the treatment protocol with corresponding theoretical rationale. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders, 24*(3), 78–88. https://doi.org/10.1044/nnsld24.3.78

Edmonds, L. A., Nadeau, S. E., & Kiran, S. (2009). Effect of Verb Network Strengthening Treatment (VNeST) on lexical retrieval of content words in sentences in persons with aphasia. *Aphasiology, 23*(3), 402–424. https://doi.org/10.1080/02687030802291339

Ellis, C., Rosenbek, J. C., Rittman, M. R., & Boylstein, C. A. (2005). Recovery of cohesion in narrative discourse after left-hemisphere stroke. *Journal of Rehabilitation Research and Development, 42*(6), 737–746. https://doi.org/10.1682/jrrd.2005.02.0026

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modeling confrontation naming and discourse informativeness using structural equation modeling. *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

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

Fromm, D., Greenhouse, J., Hou, K., Russell, G. A., Cai, X., Forbes, M., Holland, A., & MacWhinney, B. (2016). Automated proposition density analysis for discourse in aphasia. *Journal of Speech, Language, and Hearing Research, 59*(5), 1123–1132. https://doi.org/10.1044/2016_jslhr-l-15-0401

Kim, H., Kintz, S., Zelnosky, K., & Wright, H. H. (2019). Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders, 54*(1), 62–78. https://doi.org/10.1111/1460-6984.12432

Kintsch, W. (1974). *The representation of meaning in memory* (PLE: Memory) (1st ed.). Psychology Press. https://doi.org/10.4324/9781315794563

Kong, A. P.-H., Whiteside, J., & Bargmann, P. (2016). The main concept analysis: Validation and sensitivity in differentiating discourse produced by unimpaired English speakers from individuals with aphasia and dementia of Alzheimer type. *Logopedics Phoniatrics Vocology, 41*(3), 129–141. https://doi.org/10.3109/14015439.2015.1041551

Kristinsson, S., Basilakos, A., den Ouden, D. B., Cassarly, C., Spell, L. A., Bonilha, L., Rorden, C., Hillis, A. E., Hickok, G., Johnson, L., Busby, N., Walker, G. M., McLain, A., & Fridriksson, J. (2023). Predicting outcomes of language rehabilitation: Prognostic factors for immediate and long-term outcomes after aphasia therapy. *Journal of Speech, Language, and Hearing Research, 66*(3), 1068–1084. https://doi.org/10.1044/2022_jslhr-22-00347

Kristinsson, S., Basilakos, A., Elm, J., Spell, L. A., Bonilha, L., Rorden, C., den Ouden, D. B., Cassarly, C., Sen, S., Hillis, A., Hickok, G., & Fridriksson, J. (2021). Individualized response to semantic versus phonological aphasia therapies in stroke. *Brain Communications, 3*(3), Article fcab174. https://doi.org/10.1093/braincomms/fcab174

Kristinsson, S., den Ouden, D. B., Rorden, C., Newman-Norlund, R., Johnson, L., Wilmskoetter, J., Gleichgerrcht, E., Hillis, A. E., Hickok, G., Fridriksson, J., & Bonilha, L. (2025). Partial least squares multimodal analysis of brain network correlates of language deficits in aphasia. *Brain Communications, 7*(3), Article fcaf246. https://doi.org/10.1093/braincomms/fcaf246

Kristinsson, S., den Ouden, D. B., Rorden, C., Newman-Norlund, R., Neils-Strunjas, J., & Fridriksson, J. (2022). Predictors of therapy response in chronic aphasia: Building a foundation for personalized aphasia therapy. *Journal of Stroke, 24*(2), 189–206. https://doi.org/10.5853/jos.2022.01102

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Leonard, C., Rochon, E., & Laird, L. (2008). Treating naming impairments in aphasia: Findings from a phonological components analysis treatment. *Aphasiology, 22,* 923–947. https://doi.org/10.1080/02687030701831474

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

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

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

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112,* Article 104092. https://doi.org/10.1016/j.jml.2020.104092

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, Language, and Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Peach, R., & Reuter, K. (2010). A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. *Aphasiology, 24,* 971–990. https://doi.org/10.1080/02687030903058629

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

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

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richardson, J. D., & Dalton, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology, 30*(1), 45–73. https://doi.org/10.1080/02687038.2015.1057891

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rational. *Clinical Aphasiology, 24,* 121–133.

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

Scimeca, M., Peñaloza, C., & Kiran, S. (2024). Multilevel factors predict treatment response following semantic feature-based intervention in bilingual aphasia. *Bilingualism: Language and Cognition, 27*(2), 246–262. https://doi.org/10.1017/s1366728923000391

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

Silkes, J. P., Fergadiotis, G., Graue, K., & Kendall, D. L. (2021). Effects of phonomotor therapy and semantic feature analysis on discourse production. *American Journal of Speech-Language Pathology, 30*(1S), 441–454. https://doi.org/10.1044/2020_ajslp-19-00111

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA, 275*(7), 528–532. https://doi.org/10.1001/jama.1996.03530310034029

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

Stark, B. C., Alexander, J. M., Hittson, A., Doub, A., Igleheart, M., Streander, T., & Jewell, E. (2023). Test–retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research, 66*(7), 2316–2345. https://doi.org/10.1044/2023_jslhr-22-00266

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2023). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology, 37*(5), 761–784. https://doi.org/10.1080/02687038.2022.2039372

The REhabilitation and recovery of peopLE with Aphasia after StrokE (RELEASE) Collaborators. (2021). Predictors of post-stroke aphasia recovery: A systematic review-informed individual participant data meta-analysis. *Stroke, 52*(5), 1778–1787. https://doi.org/10.1161/strokeaha.120.031162

Thorne, J., & Faroqi-Shah, Y. (2016). Verb production in aphasia: Testing the division of labor between syntax and semantics. *Seminars in Speech and Language, 37*(1), 23–33. https://doi.org/10.1055/s-0036-1571356

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language, 13*(2), 345–371. https://doi.org/10.1016/0093-934X(81)90100-0

Webster, J., & Morris, J. (2019). Communicative informativeness in aphasia: Investigating the relationship between linguistic and perceptual measures. *American Journal of Speech-Language Pathology, 28*(3), 1115–1126. https://doi.org/10.1044/2019_ajslp-18-0256

Zhang, M. Y., Geng, L. Y., Yang, Y. N., & Ding, H. W. (2021). Cohesion in the discourse of people with post-stroke aphasia. *Clinical Linguistics & Phonetics, 35*(1), 2–18. https://doi.org/10.1080/02699206.2020.1734864