Effects of Semantic Feature Type, Diversity, and Quantity on Semantic Feature Analysis Treatment Outcomes in Aphasia

William S. Evans[1, 2*], Robert Cavanaugh[1, 2], Michelle L. Gravier[1, 2], Alyssa M. Autenreith[1], Patrick J. Doyle[1, 2], William D. Hula[1, 2], & Michael Walsh Dickey[1, 2]

[1] VA Healthcare System, Pittsburgh, PA, USA
[2] University of Pittsburgh, Pittsburgh, PA, USA
[3] California State University at East Bay, Hayward, California, USA

Purpose: Semantic Feature Analysis (SFA) is a naming treatment found to improve naming performance for both treated and semantically-related untreated words in aphasia. A crucial treatment component is the requirement that patients generate semantic features of treated items. This paper examined the role feature generation plays in treatment response to SFA in several ways: it attempted to replicate preliminary findings from Gravier et al. (2018), which found features generation predicted treatment-related gains for both trained and untrained words. It examined whether feature diversity or the number of features generated in specific categories differentially affected SFA treatment outcomes.
Method: SFA was administered to 44 participants with chronic aphasia daily for 4 weeks. Treatment was administered to multiple lists sequentially in a multiple-baseline design. Participant-generated features were captured during treatment and coded in terms of feature category, total average number of features generated per trial, and total number of unique features generated per item. Item-level naming accuracy was analyzed using logistic mixed-effect regression models.
Results: Producing more participant-generated features was found to improve treatment response for trained but not untrained items in SFA, in contrast to Gravier et al. (2018). There was no effect of participant-generated feature diversity or any differential effect of feature category on SFA treatment outcomes.
Conclusions: Patient-generated features remains a key predictor of direct training effects and overall treatment response in SFA. Aphasia severity was also a significant predictor of treatment outcomes. Future work should focus on identifying potential non-responders to therapy, and explore treatment modifications to improve treatment outcomes for these individuals.

Introduction

Naming impairments are pervasive and salient in people with aphasia (PWA; Goodglass & Wingfield, 1997), and their remediation has been the focus of a great deal of theoretical and clinical research (Howard, Patterson, Franklin, Orchard-lisle, & Morton, 1985; Nickels, 2002). Impairment-focused interventions that are designed to improve word-retrieval abilities among PWA typically focus on one of two stages of lexical retrieval: phonological encoding or lexical-semantic selection (Foygel & Dell, 2000; Schwartz, Dell, Martin, Gahl, & Sobel, 2006). Interventions at both levels involve repeated structured practice, with activities that are intended to facilitate access to representations that are critical to that stage of lexical retrieval. For example, Kendall and colleagues' phono-motor treatment for anomia (Kendall et al., 2008) engages PWA in a wide variety of activities that provide multimodal reinforcement and practice with retrieving and producing speech sounds (e.g., repetition and production of minimal pairs, observing clinician production of speech sounds, having PWA watch their own speech-sound production in a mirror). With sufficient practice, these activities should promote success in phonological encoding of words (both treated and untreated) that contain those sounds.

Semantically-oriented treatments use a similar logic but focus on the lexical-semantic stage of word retrieval. Semantic Feature Analysis treatment (SFA; Boyle & Coelho, 1995) and closely-related semantic feature verification treatments (Kiran, Sandberg, & Abbott, 2009; Kiran & Thompson, 2003) have patients generate or answer questions about

*Corresponding Author: William S. Evans
will.evans@pitt.edu

Evans William S., Cavanaugh Robert, Gravier Michelle L., Autenreith Alyssa M., Doyle Patrick J., Hula William D., & Dickey Michael Walsh. (2020). Effects of Semantic Feature Type, Diversity, and Quantity on Semantic Feature Analysis Treatment Outcomes in Aphasia. American Journal of Speech-Language Pathology. https://doi.org/10.1044/2020_AJSLP-19-00112

semantic features associated with treated items. For example, in SFA, PWA are presented with pictures of objects that they have difficulty retrieving, and are guided to produce associated semantic features in a number of different categories (Figure 1), on the assumption that generation of these features will increase activation of the hard-to-retrieve lexical-semantic representations (Collins & Loftus, 1975), and thereby facilitate successful word retrieval. There is significant evidence that semantic-feature-based treatments do promote improved retrieval of treated words (Boyle, 2010; e.g., Kiran & Thompson, 2003; Quique, Evans, & Dickey, 2019). In addition, repeatedly accessing these features should both: (a) strengthen the conceptual-semantic network and (b) increase activation of other words that share accessed features. Importantly, these properties of semantically-oriented naming treatments should promote generalization. Increased activation of semantically related but untreated words should engender *response generalization* (improved naming of related but untreated stimuli), whereas general strengthening of the semantic network – particularly the portion of the network associated with treated items – should promote *stimulus generalization* (improved naming of treated stimuli in other contexts). There is evidence that semantically-oriented treatment often results in response generalization (Boyle, 2010; Kiran & Thompson, 2003; Lowell, Beeson, & Holland, 1995). The evidence for stimulus generalization following such treatment, or for more general improvements in word retrieval in connected discourse, is less clear (Antonucci, 2009; Kristensson, Behrns, & Saldert, 2014; Rider, Wright, Marshall, & Page, 2008).

Further supportive evidence for the positive effect of repeated practice accessing semantic features during SFA comes from the preliminary findings of a VA RR&D-sponsored clinical trial (NCT02005016) reported by Gravier and colleagues (2018). This work examined the effects of practice-related predictors on response to intensive SFA in participants with chronic aphasia. Participants received 4 weeks of intensive SFA treatment during which they generated features in five semantic categories: superordinate category (group), physical properties (description), use (function), typical location, and personal association (see Figure 1). An examination of naming-probe performance in a preliminary sample of participants revealed that the average *number* of participant-generated semantic features per trial was predictive of naming improvements in response to SFA. Critically, this relationship held not only for treated items but for untreated but related items, whereas other practice-related predictors (such as the total number of treatment trials or total hours of treatment) did not. These findings suggested that *structured practice promoting high-dosage, repeated access to semantic features is especially important for promoting response generalization during semantically-oriented naming treatment*. They are also consistent with the

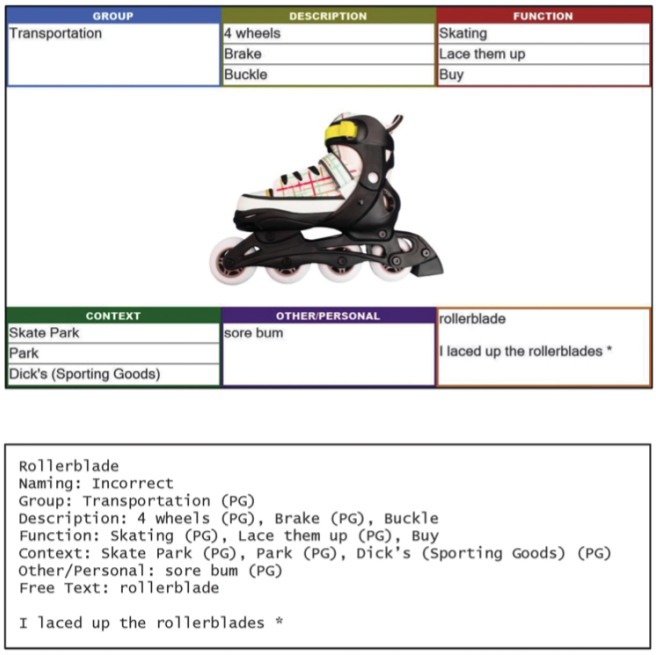


Figure 1. (Reprinted from Graver et al., 2018). Sample semantic feature analysis (SFA) trial presented via a computer with associated program output. The SFA chart (top) is visible for the duration of the trial, whereas the trial information summary (bottom) is generated as a text document and output after each session. (PG) indicates that a feature was "participant generated." The star next to the sentence indicates that the participant was able to generate a correct sentence (minimally containing the subject and verb and two semantically related elements) without clinician assistance. Note that the item name (in the "free text" box) is available during the sentence generation task.

hypothesized connection between semantic feature generation and improved lexical retrieval described above. However, this novel finding was the result of exploratory analyses on a small sample (N=17). As such, the first aim of the current study was to replicate this specific finding in the full sample of forty-four participants from this VA RR&D-sponsored clinical trial, which has now been completed.

In addition to updating the results of Gravier et al. (2018), this paper also aims to extend the analyses to address some additional remaining questions. First, while many participants were relatively consistent in the specific features they generated across trials, others were noted to generate a wider variety of features throughout the course of treatment. That is, some participants who generated more features per trial may have done so by repeatedly accessing the same features, whereas others may have generated a wider/more

diverse array of target-related features. This observation led us to question whether the positive effect of feature generation reported in Gravier et al. (2018) is a consequence of repeated access to the *same* features, or a consequence of greater feature diversity – that is, activating a broader array of features and engendering greater spread of activation within the semantic network (Collins & Loftus, 1975). If the number of unique features generated by individuals during treatment is associated with greater treatment response, this may indicate that activating features that cover a broader area within a semantic network is important for SFA treatment response. This finding would be consistent with claims that accessing atypical features may be especially beneficial for semantic feature-based naming treatments, precisely because they spread activation across a wider area within a semantic network (Kiran, 2008; Kiran & Thompson, 2003). Therefore, a second aim of this study was to examine whether diversity in feature generation predicted treatment response.

Finally, another question that arose from Gravier et al. (2018) stems from the observation that the effect of feature generation was collapsed across the five feature categories employed in SFA (see Figure 1). However, both cognitive models and clinical impression suggest that perhaps not all semantic features are equal in their facilitative effects. For example, personal associations are also strong retrieval cues. Rogers, Kuiper and Kirker (1977) found that participants showed better recall for words when they were asked to make judgments regarding their personal experience or connection to those words, than when they were asked to judge them based on phonological or semantic features. This *self-reference effect* is a robust finding in the memory and learning literature (Symons & Johnson, 1997), and may therefore help facilitate successful lexical retrieval in aphasia. Separate findings (Bower and Winzenz, 1970) found that college-aged adults learning word pairs benefitted most in recognition and recall tasks from imagery cues (visualizing an image of a tree and a boat and imagining them interacting when learning the pair *tree-boat*), more than repetition or creating a sentence containing the two words. Both these findings align with the clinical impression of our study team that physical-property and personal-association features often appear to act as especially strong retrieval cues for PWA receiving SFA treatment. Thus, the third goal of this study was to explore associations between participant-generated features within specific feature categories and SFA treatment outcomes.

Study Aims

In sum, the current study has three aims:

(1) To replicate the specific finding from Gravier et al. (2018) on the effects of participant-generated features on SFA treatment outcomes using our full study sample.

(2) To examine whether feature diversity, or the number of unique features generated, predicts changes in SFA treatment outcomes.

(3) To examine how the number of participant-generated features in specific feature categories (i.e., superordinate category/group, physical properties/description, use/function, typical location, and personal association) affect Treatment outcomes in SFA.

Method

Results from this clinical trial (NCT02005016) have been previously reported in this journal. For detailed descriptions of stimuli and treatment procedures, please refer to Gravier et al. (2018).

Participants

Forty-four adults with chronic aphasia due to unilateral left-hemisphere stroke greater than 6 months post-onset completed the study. Participants were recruited from the Western Pennsylvania Research Registry, the Audiology and Speech Pathology Research Registry maintained by the VA Pittsburgh Healthcare System (VAPHS), clinician referral, and the VA Pittsburgh's Program for Intensive Residential Aphasia Treatment and Education (PIRATE). No participants enrolled in this study received any concurrent speech-language treatment outside of the study-related sessions for the duration of the study.

To be included in the study, participants were required to score below the modality mean T-score of 70 on the Comprehensive Aphasia Test, which measures overall aphasia severity (CAT; Swinburn, Porter, & Howard, 2004). The final range of CAT mean modality T-score for participants the full dataset was 44.33 to 64.17 (mean = 52.33, standard deviation = 4.72). Participants who had a history of progressive neurological disease, nervous system injury or disorder prior to the stroke, or the presence of a severe motor speech disorder were excluded from the study. Three participants voluntarily withdrew from the study during treatment, and data from two of these participants were excluded from analysis. The third of these participants (S2) withdrew after 23 sessions due to reduced stamina resulting in an inability to tolerate the intensive treatment schedule. His data were initially excluded from Gravier et al. (2018), but have been included here in the full sample here because the total number of treatment sessions he received was similar to that of participants who completed the trial (session number average: 28.7, range: 20-37; see Supplementary Materials Table S3). Summary demographic information and language and cognitive test scores for all participants who completed the study is provided

in Table 1. Individual demographics and test scores are provided in Supplemental Material Tables S1 and S2.

Stimuli

For participants S1–S5, four treatment lists were generated, each with 10 items from three semantic categories. A list of 10 semantically related items and 10 semantically unrelated items were generated for each treatment list to assess generalization, which were balanced for number of syllables. However, for these first five participants, the time burden of administering daily naming probes prohibitively limited the amount of time available for treatment. Therefore, for the remaining participants, five-item treatment and semantically related generalization lists were generated, and the semantically unrelated generalization list was eliminated.

Items on the treatment lists were determined by performance on a picture naming task, consisting of 194 full-color photographs across eight semantic categories. The naming task was administered three times, with the third administration consisting only of items participants had named incorrectly on one of the first two administrations. Initially, an item had to be named incorrectly twice in order to be selected as a treatment or generalization item. However, this criterion was modified half-way through the trial, allowing items that had been named incorrectly only once to be selected if necessary to generate a full treatment list. This modification allowed us enroll more participants with mild aphasia to increase the diversity of our sample.

Treatment Description

Participants received individual SFA treatment (Boyle & Coelho, 1995; Coelho, McHugh, & Boyle, 2000) 4–5 days per week for 4 weeks, in two daily sessions of approximately 120 min each. Treatment was administered using a computer program, the Interactive Multimodality Assessment and Treatment Toolbox (IMATT; Winans-Mitrik et al., 2013), and only one list was targeted during each session. Stimuli were presented randomly within each list and were only repeated after the entire list had been presented.

For each treatment trial, participants were first presented with a picture of the target and asked to name it aloud within 20 seconds. Participants were then asked to generate semantic features for the target in five categories: superordinate category (group), physical properties (description), use/action (function), location (context), and personal association; see Figure 1). A three-level cueing hierarchy was used to elicit features, consisting of general prompt (e.g., "How would you describe this?"), followed by a relevant directed question (e.g., "What does this feel like?"), followed by a binary forced-choice question (e.g., "Is this item smooth or rough?"). A feature was provided by the clinician if the participant failed to respond correctly within the cueing hierarchy. Participants were encouraged to generate three features in the "description," "function," and "context" categories, and one feature in the "group" and "personal association" categories. Features were scored as *participant-generated* if they were provided verbally, either independently or after one the first two cueing levels.

At the end of each trial, participants were asked to name the target item again with clinician feedback and/or modeling for incorrect or absent responses, and the clinician read the most salient feature from each of the five feature categories aloud to them. The participant was asked to name the item one last time and if the participant was still unable to provide an accurate response, the clinician provided the correct response. At the end of each trial, participants were asked to generate a sentence using the target word with up to two cues to assist with generation or modification of the response. If the participant was unable to verbally generate a complete, correct sentence, they were asked to repeat one provided by the clinician.

All assessment and treatment procedures were provided by licensed and certified speech-language pathologists. Five treating clinicians administered the protocol over the course of the 4-year study. Each participant received the majority of treatment from a single clinician, with rare exceptions of a second clinician providing treatment coverage due to clinician illness. Treatment fidelity was monitored by a member of the study staff not providing treatment by reviewing short video-recorded segments of treatment for adherence to the SFA protocol using a treatment fidelity checklist. No deviations from the protocol were noted over the course of treatment.

Naming Probes

To assess naming maintenance and generalization, treated lists were probed daily, and all lists were probed every fourth day. Probes were administered prior to the initiation of treatment for the day. Naming probes were recorded and scored both online by the clinician and off-line by a blinded second rater. Interrater reliability was calculated as percent agreement in naming accuracy scores. If interrater reliability $< 90\%$, a third rater would also score the probe, and final accuracy would be determined via consensus. Otherwise, if interrater reliability $> 90\%$, the clinician scoring was used. In all instances for naming probes, interrater reliability was $> 90\%$.

Experimental Design

Treatment was administered targeting each list sequentially in a multiple-baseline design across behaviors and participants. Participants progressed to the next treated list when they named 90% (S1–S5) or 80% (S6–S44) of treated items accurately on three of four consecutive probes, or if the list was trained for a maximum of 8 days. Each list was also

	Mean	*SD* [1]
Age (years)	62	12
Race	5 AA[2], 2 AA/NA, 37 C, 1 H	
Gender	5 Female, 39 Male	
Education (years)	14.86	3.18
Months Post-Onset	62.5	57.2
CAT[3] Mean Modality T	52.33	4.72
PNT total	111.95	39.92
PNT S-weight	0.025	0.0097
PNT P-weight	0.022	0.0059
Pyramids and Palm Trees: 3 Picture	0.92	0.05
PALPA 02: same	0.92	0.14
PALPA 02: different	0.89	0.15
PALPA 15: written overall	0.69	0.15
PALPA 15: Auditory Overall	0.89	0.09
PALPA 49: Overall	0.8	0.1
PALPA 50: Overall	0.8	0.13

Table 1. Summary demographics and test scores for all participants (n = 44)
Note: SD = Standard Deviation; 2 AA = African-American, NA = Native American, C=Caucasian, H= Hispanic; 3 CAT = Comprehensive Aphasia Test , 4 Philadelphia Naming Test (Roach et al., 1996), 5 Pyramids and Palm Trees (Howard & Patterson, 1992), 6 Psycholinguistic Assessments of Language Processing in Aphasia (Kay et al., 1996): 02 – Minimal Pair Same/Different Judgement, 15: Word Rhyme Judgement, 49: Auditory Synonym Judgement, 50: Written Synonym Judgement

treated for a minimum of 4 days regardless of treatment probe accuracy. A total of 19/44 participants (S2–S5, S10–12, S14, S20, S21, S26, S27, S30, S31, S34, S36, S40, S42, and S43) failed to advance to list 3 during treatment and matching generalization lists for these participants were excluded from analysis.

Analysis

Entry and exit naming probe response accuracy data were used for this analysis. Entry probes were conducted immediately prior to the first treatment session and exit probes were conducted the morning after the final treatment session. Trial-level data were extracted from the IMATT computer program (Winans-Mitrik et al., 2013) and used to calculate the following variables of interest.

For Aim 1, we replicated the dependent variable from Gravier et al., (2018) by calculating the average number of patient-generated features per trial for each treated item. For Aim 2, we hand-coded all participant-generated features for each participant in order to identify the total number of unique features produced across treatment sessions. Participant-generated features were considered the *same* if they differed only in their grammatical morphology (e.g. "wash" vs. "washing"), differed in the inclusion of a pronoun or article (e.g. "deflate" vs. "deflate it"), if they were full or reduced versions of the same proper noun (e.g. "Washington D.C." vs. "D.C."), or if they were direct synonyms (e.g. "fuel tank" vs. "gas tank"). Participant-generated features were considered *different* if descriptors or adjectives were added (e.g. "play" vs. "play music") due to the differences in semantic information. After coding, we calculated the number of unique participant-generated features per trial for each item in each feature category. For Aim 3, we calculated the average number of patient-generated features per trial for each treated item, grouped separately by the feature category (e.g. "description") that elicited it.

Item-level naming probe data were analyzed using multilevel generalized linear regression with a logistic link function in R 3.6.1 (R Core Team, 2019) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014). The data structure was composed of naming accuracy for each trial (probe word) at entry and exit for each subject, for both treated and untreated words. These observations served as the predicted (outcome) variable in all models. Each of these observations was accompanied by the subject's aphasia severity (CAT Modality Mean T-Score) at entry and the primary fixed effect of interest for the model (e.g., the average number of patient generated features). These served as the predictor variables in the models.

Unlike more traditional repeated-measures techniques such as analysis of variance, this "mixed effect" modeling approach permits appropriate handling of categorical response data (Jaeger, 2008) and unbalanced designs which contain varying

Table 2. Fixed effects structure of evaluated models

Aim	Practice-related fixed effects	x Probe time	x Item Condition +	Covariate
1	Average no. of patient-generated features / trial			
2	Unique no. of patient generated features / trial			
3	Average no. of context features / trial Average no. of description features / trial Average no. of function features / trial Average no. of group features / trial Average number of personal association features / trial	Entry/Exit	Treated/ Untreated	Comprehensive Aphasia Test modality mean T-score

Note All models additional included the crossed random effects intercepts for subjects and items.

Table 2. Fixed effects structure of evaluated models
Note: All models additional included the crossed random effects intercepts for subjects and items

numbers of observations per participant or condition (Baayen, Davidson, & Bates, 2008). Furthermore, multilevel models improve the ability to make accurate inferences about the populations and effects of interest by accounting for variation in both participants and items simultaneously via crossed random effects (Baayen et al., 2008).

Using this approach, separate multilevel models were run for each practice-related fixed effect of interest. For Aim 1, the models included the average number of participant-generated features per trial for each item. For Aim 2, the models included the number of *unique* participant-generated features per trial for each item, grouped by feature category. For Aim 3, five individual models included the average number of participant-generated features per trial for each item, analyzed separately by feature category. In each model, this practice-related effect was crossed with the fixed effects of time point (entry/ exit) and item condition (treated/ untreated), in a three-way interaction. All models also included aphasia severity (CAT modality mean T-score) as a main effect covariate, as aphasia severity has been shown to affect treatment response in general (Conroy, Sage, & Lambon Ralph, 2009; Robey, 1998) and SFA response in particular (Quique et al., 2019). Therefore, each of these models tested how a given practice-related factor moderated the effects of treatment, controlling for aphasia severity. In these models, the use of the logistic link function means that each model predicted total number of correct responses on naming probes in a given session, based on the random effects and fixed effects of interest.

In terms of random effects structures, all models included crossed random effects intercepts for participants and items[^1]. Probe time was reference-coded for "exit" and item type reference-coded for "treated" in all models.[^2] Model fixed effects were plotted and interpreted using sjPlot (Lüdecke, 2018b), with 95% confidence intervals estimated via the ggeffects package (Lüdecke, 2018a). Full fixed and random effect model specifications for each model are listed in Table 2.

Results

Results for Aim 1:

The goal of Aim 1 was to replicate preliminary findings of participant-generated features on the naming outcomes in SFA found in Gravier et al. (2018) using the full study sample. Results for the average number of participant-generated features per trial are reported in Table 3. The main effects of time point ($beta$ = 2.14, p <0.001), item type ($beta$ = 1.54, p <0.001) and their interaction ($beta$ = 1.75, p <0.001) were significant, indicating that treated items improved significantly from entry to exit and that treated items improved more than untreated items. The main effect of aphasia severity was significant ($beta$ = 0.57, p <0.001), suggesting that less severe aphasia is associated with better treatment response. The main effect of the average number of features generated per trial ($beta$ = 0.45, p <0.001) and the interaction between features generated and time point ($beta$ = 0.37, p <0.001) were significant, indicating that generating more features for an item improved naming accuracy for treated items from entry to exit. Both the two-way interaction between the average number of features generated per trial and item type ($beta$ = 0.22, p = 0.005), and the three-way interaction between features generated, item type, and time point ($beta$ = 0.29, p = 0.006) were significant, suggesting that the effect of generating more features was stronger for treated than untreated items.

These results are not consistent with those reported in Gravier et al. (2018), which found that the number of participant-generated features per trial positively predicted naming accuracy from entry to exit for treated and untreated

items, and specifically, that the effect was not different between treated and untreated items. When comparing current results to previous findings (i.e., Figure 2 here compared to Gravier et al.'s Figure 3, Panel D), differences appear to have been mostly driven by a stronger relationship in the current analyses between the number of participant-generated features per trial and naming probe performance at *entry*, with more features generated during treatment associated with better pre-treatment baseline performance. This effect was significant for untreated items ($beta = 0.15$, $p = 0.008$) and at the level of a non-significant tend for treated items ($beta = 0.09$, $p = 0.13$). In Gravier et al. (2018), there was no significant relationship between participant-generated features per trial and naming probe performance at entry. Therefore, this increased association at entry in the current analysis appears to have attenuated differences between entry and exit, especially for untreated items. To further examine this effect in the current dataset, we also looked at the nested two-way interaction models between time point and features generated, separately for treated and for untreated items. These models show that the interaction effect between time point and features generated was significant for treated items ($beta = 0.36$, $p < 0.001$) but not significant for untreated items ($beta = 0.08$, $p = 0.282$; see Table 3 for full models). Therefore, for the final data set reported here, there is no evidence that generating more features is related to treatment gains on untreated items.

Results for Aim 2

The goal of Aim 2 was to examine whether feature diversity, or the number of unique features generated, predicts changes in SFA treatment outcomes. Results for the unique number of participant-generated features per trial are reported in Table 4. As in the model for Aim 1, significant main effects were found for time point ($beta = 1.11$, $p < 0.001$) and item type ($beta = 0.62$, $p < 0.001$), and a significant two-way interaction was found between time point and item type ($beta = 0.90$, $p < 0.001$). The main effect of the unique number of features generated was also significant ($beta = 0.40$, $p < 0.001$), indicating that generating more unique features during treatment was associated with greater naming accuracy of treated items at exit. However, Figure 3 shows that the same basic positive relationship between the number of unique features generated and naming accuracy was present both at entry and exit and for treated and untreated items. In addition, the lack of significant two-way interactions between time point and the number of unique features generated ($beta = -0.05$, $p = 0.49$), between item type and the number of unique features generated ($beta = 0.18$, $p = 0.074$), and the lack of significant three-way interaction between all three fixed effects ($beta = -0.05$, $p = 0.63$) further supports the conclusion that there was no relationship between the number of unique semantic features produced and any *practice-related* changes in naming performance, since effects are observed at entry, before treatment began.

Results for Aim 3

The goal of Aim 3 was to examine how the number of participant-generated features in specific feature categories (i.e., superordinate category/group, physical properties/description, use/function, typical location, and personal association) affects treatment outcomes in SFA. Results for the average number of participant-generated features per trial by feature category are reported in Table 5. The crucial test for these models was the three-way interaction between time point, treatment condition, and features generated, separately by feature category. After Bonferroni correction for multiple comparisons ($alpha = .01$), this three-way interaction was significant for the description and function categories but not for the context, group, and personal association categories. Despite these differences in statistical significance, visualization of description, function, context, and group feature categories revealed the same basic relationships between time point, treatment condition, and features generated (Figure 4, panels A through D).

In contrast, the personal-association feature category exhibited a different pattern, and the three-way interaction was not significant. Referring to (Figure 4, Panel E), this null three-way interaction appears to be driven by differences in slope for trained items at exit in the personal-association model, compared to the effects depicted for other feature categories. Increasing the number of personally-relevant features appears to be associated with a shallower slope for treated items at exit, possibly driven by the fact that individuals who generated very few personal-association features still showed *some* direct training gains for treated items. In other words, the reduction in treatment gains associated with generating fewer personal-association features does not appear to be as great as the penalty associated with producing fewer of the other feature types. The shallower effect of feature generation on treated items means that treated and untreated items matched in slope at exit, and this seems to account for most of what is driving this null 3-way interaction. It does not suggest effects of generalization or direct training for this feature type.

	Overall			Treated			Untreated		
Predictors	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	0.93	0.65 – 1.22	**<0.001**	0.97	0.67 – 1.26	**<0.001**	-0.64	-0.93 – -0.34	**<0.001**
Time Point	2.14	1.84 – 2.43	**<0.001**	2.17	1.86 – 2.49	**<0.001**	0.4	0.12 – 0.68	**0.005**
Item Condition	1.54	1.24 – 1.84	**<0.001**						
Features Generated	0.45	0.32 – 0.58	**<0.001**	0.44	0.31 – 0.58	**<0.001**	0.27	0.13 – 0.40	**<0.001**
CAT Mean Modality T	0.57	0.37 – 0.76	**<0.001**	0.64	0.41 – 0.87	**<0.001**	0.46	0.23 – 0.70	**<0.001**
Time Point * Item Condition	1.75	1.35 – 2.15	**<0.001**						
Time Point * Features Generated	0.37	0.22 – 0.52	**<0.001**	0.36	0.21 – 0.51	**<0.001**	0.08	-0.07 – 0.23	0.282
Item Condition * Features Generated	0.22	0.07 – 0.37	**0.005**						
Time Point * Item Condition *	0.29	0.08 – 0.50	**0.006**						
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	0.61_{Target}			0.68_{Target}			0.80_{Target}		
	$0.24_{\text{Participant}}$			$0.21_{\text{Participant}}$			$0.26_{\text{Participant}}$		

Table 3. Generalized Linear Models for the Average Number of Features Generated per Trial and Nested Models for Treated and Untreated Items.
CI = confidence interval; CAT = Comprehensive Aphasia Test. Bolded text indicates significance at $alpha < .05$.

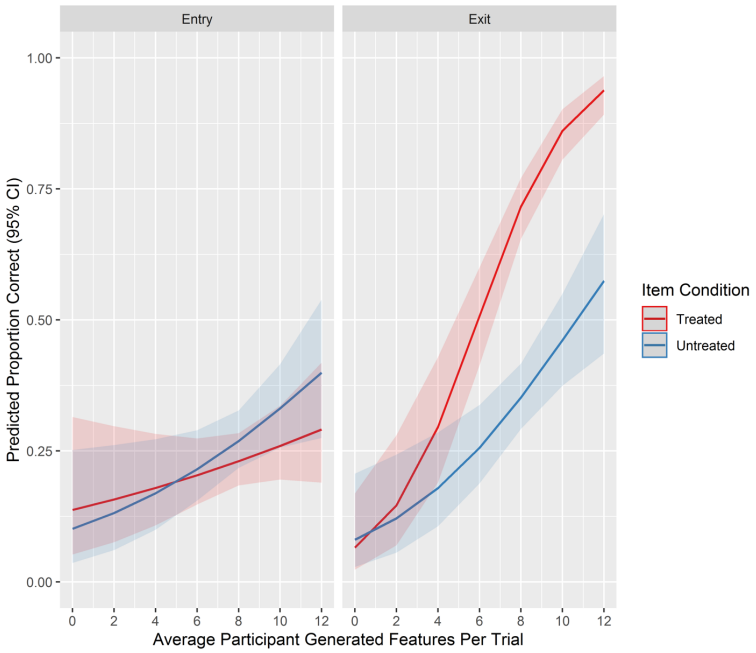


Figure 2. Predicted proportion of correct naming responses by average number of participant-generated features per trial at entry and exit probes.
Note: The y-axis reflects the estimated proportion of naming probes correct for the average participant in this sample. Therefore, Figure 2 depicts how naming probe performance at entry and exit would predicted to change for a participant with average (moderate) aphasia severity, based on producing more or less average features per trial.

Predictors	*Log-Odds*	*CI*	*p*
(Intercept)	-0.2	-0.50 – 0.09	0.181
Time Point	1.11	0.96 – 1.26	**<0.001**
Item Condition	0.62	0.40 – 0.85	**<0.001**
Features Generated	0.4	0.18 – 0.61	**<0.001**
CAT Mean Modality T	0.73	0.47 – 0.99	**<0.001**
Time Point * Item Condition	0.9	0.70 – 1.09	**<0.001**
Time Point * Features Generated	-0.05	-0.18 – 0.09	0.491
Item Condition * Features Generated	0.18	-0.02 – 0.38	0.074
Time Point * Item Condition * Features Generated	-0.05	-0.23 – 0.14	0.632
Random Effects			
σ^2			3.29
$\tau_{00\ Target}$			0.52
$\tau_{00\ Participant}$			0.6

Table 4. Generalized Linear Model for the Unique Number of Features Generated per Trial
CI = confidence interval; CAT = Comprehensive Aphasia Test. Bolded text indicates significance at $alpha < .05$.

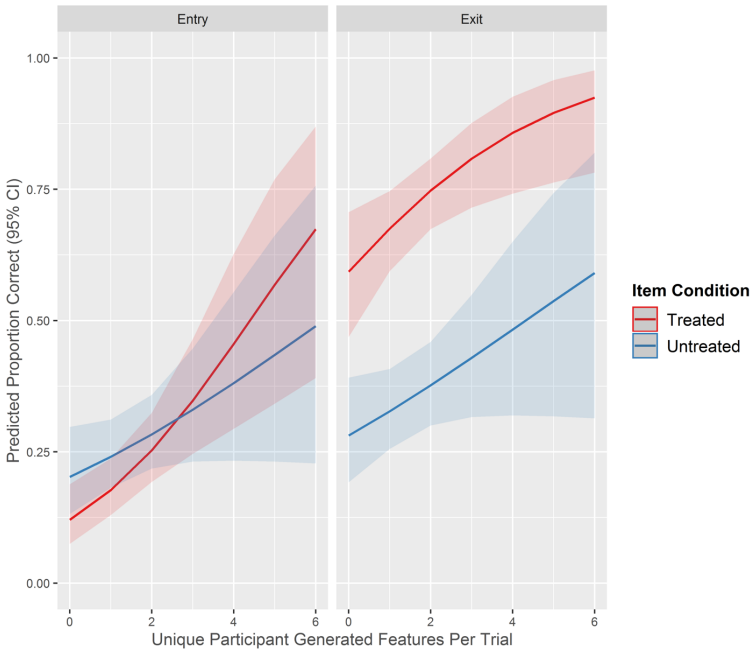
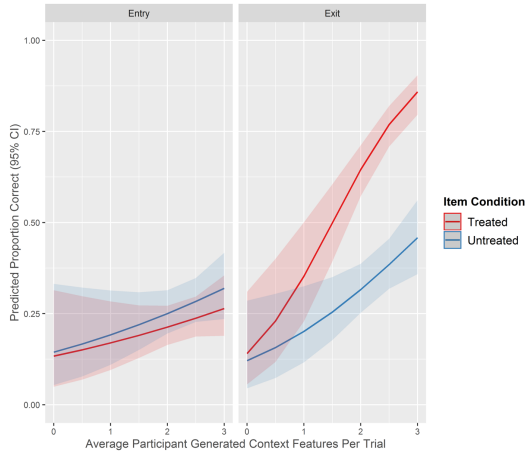
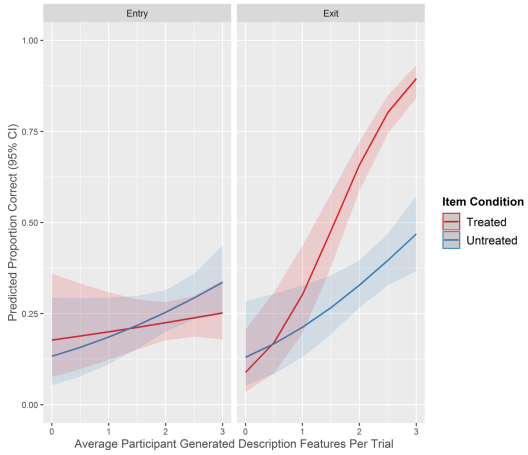


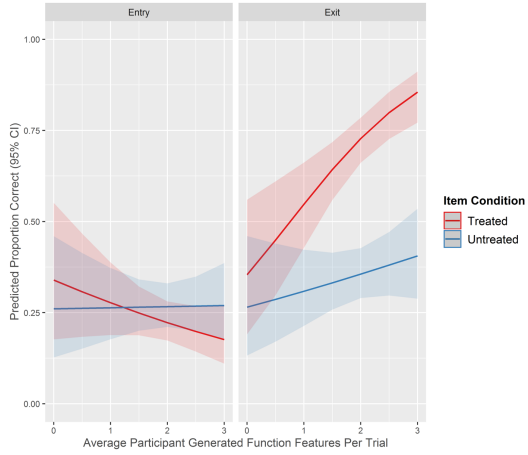
Figure 3. Predicted proportion of correct naming responses by unique number of participant-generated features per trial at entry and exit probes.



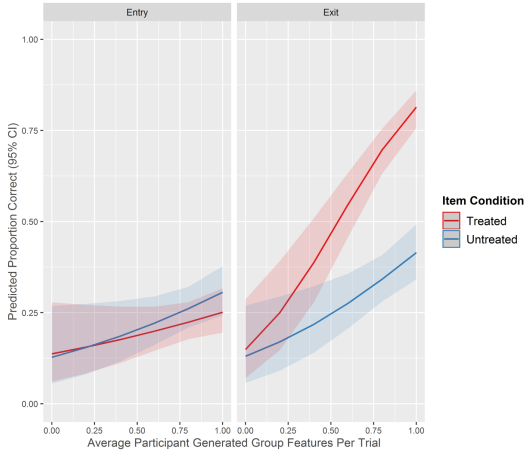
(a)



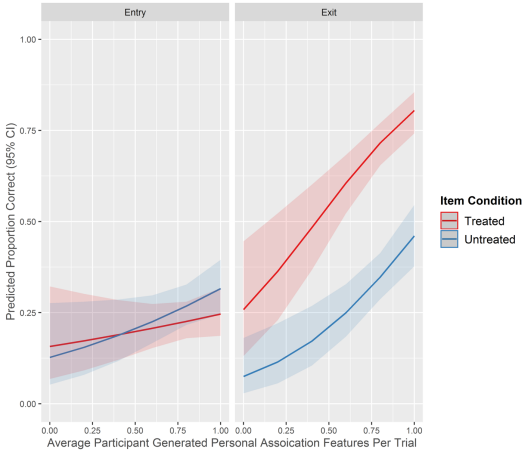
(b)



(c)



(d)



(e)

Predicted proportion correct naming responses for each feature category by the average number of participant-generated features per trial at entry and exit probes. CI = confidence interval.

	Group			Personal Association			Context			Description			Function		
Predictors	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	0.91	0.62 – 1.20	**<0.001**	0.91	0.62 – 1.19	**<0.001**	0.92	0.63 – 1.21	**<0.001**	0.93	0.64 – 1.23	**<0.001**	0.94	0.63 – 1.25	**<0.001**
Time Point	2.13	1.84 – 2.43	**<0.001**	2.14	1.85 – 2.43	**<0.001**	2.15	1.86 – 2.44	**<0.001**	2.14	1.85 – 2.44	**<0.001**	2.17	1.88 – 2.46	**<0.001**
Item Condition	1.53	1.22 – 1.83	**<0.001**	1.55	1.25 – 1.85	**<0.001**	1.53	1.23 – 1.83	**<0.001**	1.54	1.24 – 1.84	**<0.001**	1.54	1.24 – 1.84	**<0.001**
Features Generated	3.22	2.27 – 4.17	**<0.001**	2.47	1.46 – 3.48	**<0.001**	1.21	0.77 – 1.64	**<0.001**	1.49	1.06 – 1.92	**<0.001**	0.79	0.38 – 1.21	**<0.001**
CAT Mean Modality T	0.56	0.35 – 0.77	**<0.001**	0.67	0.47 – 0.86	**<0.001**	0.63	0.42 – 0.83	**<0.001**	0.59	0.38 – 0.79	**<0.001**	0.72	0.48 – 0.96	**<0.001**
Time Point * Item Condition	1.74	1.34 – 2.14	**<0.001**	1.78	1.38 – 2.17	**<0.001**	1.76	1.36 – 2.15	**<0.001**	1.74	1.35 – 2.14	**<0.001**	1.76	1.37 – 2.16	**<0.001**
Time Point * Features Generated	2.47	1.34 – 3.60	**<0.001**	1.91	0.65 – 3.17	**0.003**	0.92	0.42 – 1.43	**<0.001**	1.34	0.85 – 1.83	**<0.001**	1.08	0.62 – 1.55	**<0.001**
Item Condition * Features Generated	1.66	0.50 – 2.83	**0.005**	0.12	-1.23–1.47	0.865	0.6	0.08 – 1.12	0.024	0.9	0.39 – 1.41	**0.001**	0.58	0.11 – 1.05	0.016
Time Point * Item Condition * Features Generated	2.02	0.39 – 3.65	0.015	0.71	-1.16–2.58	0.456	0.66	-0.06–1.38	0.072	1.15	0.47 – 1.83	**0.001**	0.89	0.24 – 1.53	**0.007**
Random Effects															
σ^2	3.29			3.29			3.29			3.29			3.29		
τ_{00}	0.57 $_{\text{Target}}$			0.57 $_{\text{Target}}$			0.59 $_{\text{Target}}$			0.60 $_{\text{Target}}$			0.57 $_{\text{Target}}$		
	0.28 $_{\text{Participant}}$			0.28 $_{\text{Participant}}$			0.29 $_{\text{Participant}}$			0.27 $_{\text{Participant}}$			0.43 $_{\text{Participant}}$		

Table 5. Generalized Liner Model for the Average Number of Features Generated per Trial within each Feature Category
CI = confidence interval; CAT = Comprehensive Aphasia Test. Bolded text indicates significance at *alpha* < .01.

Discussion

The rationale for this paper was to replicate the preliminary findings reported in Gravier et al. (2018), and to further examine how specific aspects of participant-generated features affect treatment outcomes in semantic feature analysis (SFA). Specifically, exploratory analyses on our preliminary dataset in Gravier et al. (2018) found that participants who generate more semantic features during SFA treatment show greater improvements in naming performance. Furthermore, these improvements were roughly equal for both treated and untreated items, which suggested that focusing on generating features during SFA might facilitate treatment generalization. Thus, in Aim 1, we updated this preliminary analysis with the full dataset from our VAPHS clinical trial. In Aim 2, we examined how feature diversity, measured in terms of the number of unique features generated, predicted treatment-related gains in naming performance. Finally, in Aim 3, we examined whether participant-generated features within different feature categories differentially predicted treatment-related gains in naming performance.

Although Gravier et al. (2018) found that the average number of participant-generated features predicted treatment gains both for treated and untreated related words, the current replication only found evidence for a direct training effect: generating more features during SFA appears to predict treatment gains for treated items, but not for untreated items. This may be an instance of "regression to the mean." Specifically, if an initial effect size (in this case, the effect of participant-generated features on treatment outcomes for untreated items) is over-estimated due to random variation, additional testing tends to attenuate findings by producing values closer to the true mean (Barnett, van der Pols, & Dobson, 2005). This change is not likely due to any changes in aphasia severity of enrolled PWA over the course of the study due to shifts in item selection criteria, since analyses specifically controlled for this by including aphasia severity as a covariate.

Since baseline aphasia severity has been shown to predict treatment response in SFA (Quique et al., 2019), one potential interpretation suggested by a helpful reviewer is that the features generated variable may represent a more fine-grained measure of baseline naming ability, which showed a predictive effect even after controlling for aphasia severity. However, the average number of patient-generated features per trial showed only moderate correlations with aphasia severity as measured by CAT mean modality T-score (r =.45; r^2 = .20) and naming ability as assessed by the Philadelphia Naming Test (r =.56; r^2 = .31). Given that both the CAT and the PNT are reliable tests and the large amount of trial-level data that went into the patient-generated features measure, it appears to reflect something distinct about individual performance above-and-beyond overall aphasia severity and naming ability. As mentioned in the results section, one change between the current findings and the preliminary analyses in Gravier et al. (2018) was an increase in the strength of the relationship between participant-generated features and naming probe performance at baseline. If the ability to generate features during treatment is associated with the ability to name probes at entry, before treatment occurs, then this relationship is likely at least partially driven by individual differences that exist prior to intervention, and is not purely a practice-related factor.

This highlights a confound in the current study design, in that the number of participant-generated features was not specifically controlled. This does not allow us to determine whether feature generation serves as a direct mechanism of action, or is instead a proxy for underlying individual differences in language ability which are themselves responsible for differences in SFA outcomes. For example, if PWA with relatively spared semantic systems only generated a few features, would they still show the same level of treatment gains? Conversely, if modifications to the SFA protocol allowed PWA with more severe semantic deficits generate more features, would they show improved treatment gains? Therefore, a critical next step in this line of research is to specifically manipulate the number of features generated to control for individual differences in PWA.

Regardless, the absence of a clear effect of feature generation on improvement for untreated items is somewhat surprising, given both the theoretical motivation of SFA and the body of evidence demonstrating that word retrieval is facilitated by presentation or repetition of semantically-related material (see e.g., Nelly, 2012, for review). This finding may suggest that SFA operates at least in part via mechanisms other than automatic priming of semantic representations, such as SFA-prompted self-cueing strategies (as originally suggested by Boyle & Coelho, 1995). Further work is needed to disentangle the contributions of these different mechanisms.

Aim 2 examined how feature diversity, or the number of unique features generated, predicted changes in naming performance. This hypothesis was motivated by the notion that greater feature diversity may engender greater spreading activation across a network (Boyle & Coelho, 1995; Collins & Loftus, 1975), thereby producing larger treatment gains. Furthermore, positive effects of greater feature diversity (i.e., sampling features more broadly within the semantic network), would also be consistent with claims that accessing atypical features is especially beneficial for semantic feature-based naming treatments (Kiran, 2008; Kiran & Thompson, 2003).

However, the current findings did not provide any evidence that generating more unique features per trial has any specific effect on treatment gains for either treated or untreated items. This is in contrast to the Aim 1 findings, where the number

of features generated (not accounting for diversity), was predictive of treatment gains for treated items. Together, these findings draw attention to an inherent tension between feature diversity and repeated practice in treatments like SFA. Specifically, for a given dosage, there is a direct tradeoff between the number of different feature-target pairs that can be practiced overall and the amount that specific feature-target pairs can be strengthened via repetition. While failing to reject the null hypothesis here does not allow us "accept the null" and conclude that there is no true underlying relationship between participant-generated feature diversity and treatment outcomes, the small effect sizes (see log-odds estimates and 95% confidence intervals for the crucial interaction terms in Table 4) do suggest even if such a relationship exists, it is unlikely to large enough to be of much clinical significance.

As in the Aim 1 analyses, it should again be noted that we did not experimentally manipulate feature diversity, and that these conclusions are only supported within the natural range of variation in feature generation that was engendered by our clinical design. Modifying SFA to intentionally target an extremely limited or extremely broad set of features might produce values outside our sample range, and different outcomes as a result; this is a question for future study.

Aim 3 examined how generation of different feature categories predicted changes in naming performance for treated and untreated items in response to SFA. We hypothesized that the personal association and description (physical-property) feature categories would demonstrate particularly robust effects on treatment gains, based on previous findings from the memory and learning literatures (Bower & Winzenz, 1970; Rogers et al., 1977; Symons & Johnson, 1997).

In evaluating the crucial three-way interaction between time point (entry vs. exit), item type (treated vs. untreated), and number of participant-generated features within each feature category, models testing the description and function categories showed significant interactions after Bonferroni correction, while the models examining group, personal association, and context did not. With the exception of personal association, visualization of these models revealed the same basic pattern as the omnibus model described under Aim 1: specifically, that generating more features in each of these four feature categories appears to predict treatment gains, and does so primarily for treated items. While the null 3-way interaction found in the personal association feature category model appears to be an exception to this overall trend, visualization of the personal-category model findings (see Figure 4, Panel E) were not suggestive of greater generalization effects for individuals who generated more personal-association features. Instead, the absence of a three-way interaction appears to reflect a smaller penalty for generating fewer personal-association features, particularly for treated items. Overall, the current findings provide little evidence that individual feature categories appear to have a distinctive (stronger *or* weaker) effect on treatment gains in SFA. While our null findings for Aims 2 or 3 did not provide support for our initial hypotheses, we would like to explicitly note the importance of reporting them here in order to minimize effect estimate biases in the published literature (Ioannidis, 2006).

Clinical implications

Based on our findings, we offer the following clinically-relevant conclusions and recommendations.

1. **SFA improves naming ability for both trained words and semantically-related untrained words.** In this clinical trial of 44 PWA, SFA treatment improved confrontation picture naming for both trained and semantically-related untrained words, with larger effects for trained words. This is consistent with previous meta-analysis findings (Quique et al., 2019) and with another recent clinical trial reporting treatment effects of SFA on 30 PWA (Kendall et al., 2019), both of which also reported outcomes at the group level. In a recent systematic review, Efstratiadou et al. (2018) found that only 40% of previous single-subject case series studies of SFA reported generalization to semantically-related untrained stimuli, suggesting that individual variability may account for why generalization effects are modest at the group level. The source of these individual differences (i.e., person-level predictors of treatment response) should be investigated further.

2. **Patient-generated features is a key predictor of treatment response in SFA.** In the current study, generating more features was associated larger direct training effects, controlling for aphasia severity. In addition, review of individual participant performance in terms of responders and non-responders revealed a helpful performance-based clinical benchmark: *no PWA who averaged less than 5 features per trial (out of 11 opportunities) over the course of treatment demonstrated any notable treatment gains in treated or untreated items*. This suggests that a patient who cannot generate many features on average during standard SFA treatment is unlikely to show gains. While our specific benchmark is currently based on the average number of participant-generated features across the entire study intervention, the general principle (very poor feature generation = poor treatment prognosis) should still apply to clinicians trialing SFA during diagnostic treatment.

3. **Patient-generated semantic feature diversity did not predict treatment outcomes in this SFA clinical trial.**

There were no significant effects of feature diversity on treatment outcomes, and estimated effect sizes based on model fixed effects were quite small. Therefore, varying feature diversity within the range seen in our sample is unlikely to produce any clinically-meaningful differences in improving treatment outcomes for PWA. Considering this second point in conjunction with the first leads to the following clinical recommendation: when using SFA with a patient, if attempting to elicit novel semantic features starts to become an overly time-consuming process, it may be better to focus on repeated practice of fewer features to maximize feature quantity over feature diversity.

4. **Aphasia severity continues to be a significant predictor of treatment outcomes in SFA.** Aphasia severity was consistently found as a significant predictor of naming probe performance in the models presented above (as well as in the previous findings by Gravier et al., 2018). For example, in the Aim 1 model evaluating the impact of the average number of participant-generated features per trial, the odds of a correct response for treated items increased 1.77 times for each 1 standard-deviation increase in CAT modality-mean T-score (aphasia severity). In other words, more severe PWA were less likely to improve in response to our SFA intervention. This finding contributes to the existing evidence that aphasia severity plays a prognostic role in response to intervention (e.g., Lambon Ralph, Snell, Fillingham, Conroy, & Sage, 2010; Lee, Kaye, & Cherney, 2009), including response to SFA (Quique et al., 2019). Therefore, a patient's overall level of aphasia severity should be considered when selecting SFA as a potential intervention, as patients with more severe aphasia are less likely to demonstrate treatment-related gains.

Limitations and future directions

As noted above, a central limitation to this paper is that none of the key variables of interest associated with feature generation were experimentally manipulated; these analyses reflect natural variation in the data, thus limiting our ability to draw strong, causal conclusions about potential mechanisms of action in SFA. In addition, confounds in the amount of dosage provided between treatment lists required us to control for the total number of trials in our patient-generated features variables instead of further investigating this factor directly. Therefore, future work should experimentally-manipulate key practice-related variables of interest to better dissociate individual differences in underlying cognitive-linguistic ability from differences in behavioral during therapy and better control treatment list exposure.

While the current results suggest that PWA who cannot generate at least an average of 5/10 semantic features per trial are not good candidates for SFA, it is unclear how these same individuals would respond to related semantically-based naming treatments such as feature verification, where features do not need to be verbally produced (Kiran et al., 2009; Kiran & Thompson, 2003). Direct comparisons could be a focus of future work.

In regards to the Aim 3 analyses examining differences between participant-generated feature categories, three limitations should be noted: first, the number of features participants were prompted to generate differed between categories. Participants were encouraged to generate three features for the context, function, and description categories, but only one feature for the personal-association and group categories. Between-category differences of this type should be controlled in future work. Second, while category *prompts* were clearly distinct, participant-generated features themselves did not always neatly correspond to a single category, and participants occasionally used the same feature for multiple categories. For example, a participant generated the same feature, "in Bob's backyard," in response to both the personal association and location category prompts. Since our data were coded in direct relation to the category prompt, our findings are more directly relevant to how SFA fractions semantic conceptual space, and do less to address detailed semantic relationships between individual feature-target pairs themselves. Future work could look at issues of semantic overlap and semantic relationships more specifically by investigating the effects of feature-target semantic similarity on treatment response (Mandera, Keuleers, & Brysbaert, 2017). Third, Aim 3 made use of 3-way interaction models analyzed in parallel for each feature category. Strong conclusions about differences between these models would require directly comparing differences statistically. However, such comparisons would have necessitated the use of 4-way interaction models, introducing both interpretive issues and power considerations, which negatively affects the strength of our conclusions for Aim 3.

Conclusions

Overall, results from our full study sample show that increasing the number of participant-generated features improves treatment response for trained but not untrained items in SFA. However, there were large individual differences in the number of features generated by participants in our sample, and it is not clear if the current findings are attributable to individual differences in cognitive-linguistic ability or differences in practice-related factors. Therefore, future work investigating these effects should experimentally control the numbers of features generated by participants to determine whether feature generation *causally* contributes to treatment gains in SFA (Boyle, 2010).

While we find no evidence that participant-generated feature diversity or feature category differentially affect treatment outcomes in SFA, several clinically relevant conclusions may still be drawn from the current study. Patient-generated features remains a key predictor of treatment response in SFA, especially for direct training, and PWA who cannot generate a sufficient number of features on average (5 out of 11 opportunities in the current clinical trial), do not appear to respond well to SFA. Aphasia severity continues to be a significant predictor of treatment outcomes in SFA as well. Future work should focus on determining whether person-related factors that can help us identify, pre-intervention, potential non-responders to therapy, and explore treatment modifications to improve treatment outcomes for these individuals.

Acknowledgments

We are grateful to the participants with aphasia who took part in the clinical trial (NCT02005016) whose results are partially reported here. We would also like to acknowledge the contribution of Elisabeth Ashmore in help with data processing and semantic feature coding, our study research participants, and present and past members of the Program for Intensive Residential Aphasia Treatment and Education clinical and research staff, including Rebecca Ruffing, Ronda L. Winans-Mitrik, Shannon Austerman Hula, Emily Boss, Angela Grzybowski, Mary Sullivan, Brooke Swoyer, and Beth Friedman.

Funding statement: This research was supported by VA Rehabilitation Research and Development Award I01RX000832 to Michael Walsh Dickey and Patrick J. Doyle, and a VA Pittsburgh Healthcare System Geriatric Research Education and Clinical Center Pilot Project Awarded to William S. Evans. The contents of this article do not represent the views of the Department of Veterans Affairs of the U.S. Government.

Conflict of interest statement: The authors receive salary from their supporting institutions. They have declared that no competing interests existed at the time of publication.

References

Antonucci, S. M. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology*, *23*(7–8), 854–866. https://doi.org/10.1080/02687030802634405

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59* (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, *20*(2), 119–120. https://doi.org/10.3758/BF03335632

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: what's in a name? *Topics in Stroke Rehabilitation*, *17*(6), 411–422. https://doi.org/10.1310/tsr1706-411

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

Coelho, C. A., McHugh, R. E., & Boyle, M. (2000). Semantic feature analysis as a treatment for aphasic dysnomia: A replication. *Aphasiology*, *14*(2), 133–142. https://doi.org/10.1080/026870300401513

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Conroy, P., Sage, K., & Lambon Ralph, M. A. (2009). Errorless and errorful therapy for verb and noun naming in aphasia. *Aphasiology*, *23*(11), 1311–1337. https://doi.org/10.1080/02687030902756439

Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A Systematic Review of Semantic Feature Analysis Therapy Studies for Aphasia. *Journal of Speech, Language, and Hearing Research*, *61*(May 2018), 1261–1278.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182–216. https://doi.org/10.1006/jmla.2000.2716

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

Gravier, M. L., Dickey, M. W., Hula, W. D., Evans, W. S., Owens, R. L., Winans-Mitrik, R. L., & Doyle, P. J. (2018). What Matters in Semantic Feature Analysis:

Practice-Related Predictors of Treatment Response in Aphasia. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *27*(1S), 438–453. https://doi.org/10.1044/2017_AJSLP-16-0196

Howard, D., Patterson, K., Franklin, S., Orchard-lisle, V., & Morton, J. (1985). The facilitation of picture naming in aphasia. *Cognitive Neuropsychology*, *2*(1), 49–80. https://doi.org/10.1080/02643298508252861

Howard, D., & Patterson, K. (1992). The Pyramids and Palm Trees Test: A test of semantic access from words and pictures. Pearson Assessment.

Ioannidis, J. P. A. (2006). Journals should publish all "null" results and should sparingly publish "positive" results. *Cancer Epidemiology Biomarkers & Prevention*, *15*(1), 186–186. https://doi.org/10.1158/1055-9965.EPI-05-0921

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kay, J., Lesser, R., & Coltheart, M. (1996). Psycholinguistic assessments of language processing in aphasia (PALPA): An introduction. *Aphasiology*, *10*(2), 159–180. https://doi.org/10.1080/02687039608248403

Kendall, D. L., Moldestad, M. O., Allen, W., Torrence, J., & Nadeaue, S. E. (2019). Phonomotor Versus Semantic Feature Analysis Treatment for Anomia in 58 Persons With Aphasia: A Randomized Controlled Trial. *Journal of Speech Language and Hearing Research*, 1–19.

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Gonzalez Rothi, L. J., & Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language*, *105*(1), 1–17. https://doi.org/10.1016/j.bandl.2007.11.007

Kiran, S. (2008). Typicality treatment for naming deficits in aphasia: why does it work? *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, *18*(1), 6. https://doi.org/10.1044/nnsld18.1.6

Kiran, S., Sandberg, C., & Abbott, K. (2009). Treatment for lexical retrieval using abstract and concrete words in persons with aphasia: Effect of complexity. *Aphasiology*, *23*(7), 835–853. https://doi.org/10.1080/02687030802588866

Kiran, S., & Thompson, C. K. (2003). The role of semantic complexity in treatment of naming deficits: training semantic categories in fluent aphasia by controlling exemplar typicality. *Journal of Speech, Language, and Hearing Research*, *46*(4), 773–787.

Kristensson, J., Behrns, I., & Saldert, C. (2014). Effects on communication from intensive treatment with semantic feature analysis in aphasia. *Aphasiology*, 1–22. https://doi.org/10.1080/02687038.2014.973359

Lambon Ralph, M. A., Snell, C., Fillingham, J. K., Conroy, P., & Sage, K. (2010). Predicting the outcome of anomia therapy for people with aphasia post CVA: both language and cognitive status are key predictors. *Neuropsychological Rehabilitation*, *20*(2), 289–305. https://doi.org/10.1080/09602010903237875

Lee, J. B., Kaye, R. C., & Cherney, L. R. (2009). Conversational script performance in adults with non-fluent aphasia: Treatment intensity and aphasia severity. *Aphasiology*, *23*(7–8), 885–897. https://doi.org/10.1080/02687030802669534

Lowell, S., Beeson, P. M., & Holland, A. L. (1995). The efficacy of a semantic cueing procedure on naming performance of adults with aphasia. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *4*(4), 109. https://doi.org/10.1044/1058-0360.0404.109

Lüdecke, D. (2018a). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *The Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

Lüdecke, D. (2018b). sjPlot: Data Visualization for Statistics in Social Science (R package version 2.4.1.9000) \[Computer software\].

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In *Basic processes in reading* (pp. 272-344). Routledge.

Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, *16*(10–11), 935–979. https://doi.org/10.1080/02687030244000563

Quique, Y. M., Evans, W. S., & Dickey, M. W. (2019). Acquisition and Generalization Responses in Aphasia Naming Treatment: A Meta-Analysis of Semantic Feature Analysis Outcomes. *American Journal of Speech-Language Pathology / American Speech-Language-*

Hearing Association, *28*(1S), 230–246. https://doi.org/10.1044/2018_AJSLP-17-0155

R Core Team. (2019). R: A language and environment for statistical computing (3.6.1) \[Computer software\]. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research*, *41*(1), 172–187. https://doi.org/10.1044/jslhr.4101.172

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*(9), 677–688. https://doi.org/10.1037//0022-3514.35.9.677

Schwartz, M., Dell, G., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*, *54*(2), 228–264. https://doi.org/10.1016/j.jml.2005.10.001

Swinburn, K., Porter, G., & Howard, D. (2004). Comprehensive aphasia test. *Comprehensive Aphasia Test*.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: a meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Winans-Mitrik, R., Chen, S., Owens, R., Hula, W. D., Eichhorn, K., Ebrahimi, M., & Doyle, P. J. (2013). Tele-rehabilitation solutions for patients with aphasia. *Poster Presented at the Annual Meeting of the Association of Veterans Affairs Speech-Language Pathologists, San Francisco, CA*.

Supplementary Materials

1 Table S1: Participant Demographics

Participant	Age (years)	Race[2]	Gender[3]	Education (years)	Etiology	MPO[5]	CAT Mean Modality T[6]
S1	42	AA, NA	M	15	ischemic LH CVA w/ HC[4]	55	53.67
S2	77	C	M	12	ischemic LH CVA	33	46.83
S3	62	C	M	23	ischemic LH CVA	89	50
S4	51	C	M	13	ischemic LH CVA w/ HC	75	46.33
S5	68	C	F	12	ischemic LH CVA	199	47.5
S6	66	C	M	18	ischemic LH CVA	86	64.17
S7	52	C	M	12	ischemic LH CVA	10	56
S8	24	C	M	16	LH cerebral aneurysm	10	58
S9	78	H	M	25	ischemic LH CVA	16	52.5
S10	64	C	M	11	ischemic LH CVA w/ HC	84	47.5
S11	45	AA	M	12	ischemic LH CVA	120	49.83
S12	72	C	M	14	ischemic LH CVA	93	50.83
S13	51	AA	M	16	ischemic LH CVA	39	51.33
S14	75	C	M	13	ischemic LH CVA	14	47.83
S15	70	C	M	14	ischemic LH CVA	172	50.16
S16	48	C	M	14	ischemic LH CVA	8	50.16
S17	74	C	M	20	ischemic LH CVA	15	52.83
S18	67	C	F	12	ischemic LH CVA w/ HC	8	59
S19	71	C	M	16	ischemic LH CVA	35	54.83

S20	63	C	M	14	ischemic LH CVA	77	46.5
S21	62	AA	M	14	hemorrhagic LH CVA	151	46.5
S22	63	C	M	18	ischemic LH CVA	74	48.67
S23	65	C	M	20	ischemic LH CVA	64	58.5
S24	57	C	F	17	ischemic LH CVA	102	56.33
S25	49	C	M	12	ischemic LH CVA	94	52.67
S26	68	C	M	12	ischemic LH CVA	161	49.16
S27	50	C	M	12	ischemic LH CVA	111	48.33
S28	71	C	F	15	ischemic LH CVA	114	51.5
S29	44	AA	F	15	ischemic LH CVA	33	62
S30	59	C	M	20	ischemic LH CVA	25	52.83
S31	68	C	M	18	ischemic LH CVA	41	56.5
S32	73	C	M	14	ischemic LH CVA w/ HC	18	53.5
S33	53	C	M	14	ischemic LH CVA	7	55.33
S34	61	AA	M	13	ischemic LH CVA	245	47.83
S35	72	C	M	16	ischemic LH CVA	7	46.83
S36	68	C	M	16	ischemic LH CVA	59	52
S37	67	C	M	14	ischemic LH CVA	54	52
S38	70	C	M	10	ischemic LH CVA	18	58.83
S39	31	C	M	14	ischemic LH CVA	29	59.17
S40	68	C	M	12	ischemic LH CVA	14	52.83

S41	69	C	M	12	ischemic LH CVA	7	60.33
S42	76	C	M	16	ischemic LH CVA	15	44.33
S43	71	C	M	12	hemorrhagic LH CVA	64	49.5
S44	54	AA, NA	M	16	ischemic LH CVA	6	50
AVG (SD[1])	62 (12)		5F, 39M	14.86 (3.18)		62.5 (57.2)	52.33 (4.72)

[1] SD = Standard Deviation; [2] AA = African-American, NA = Native American, C=Caucasian, H= Hispanic; [3] M = Male, F= Female; [4] LH = left hemisphere, HC = hemorrhagic conversion; [5] MPO = months post onset; [6] CAT = Comprehensive Aphasia Test

Table S2: Language Testing Results

Participant	Philadelphia Naming Test*			Pyramids and Palm Trees**	PALPA 02***[a]		PALPA 15[b]		PALPA 49[c]	PALPA 50[d]
	Total Correct	s-weight	p-weight	3-Picture	Same	Different	Written Overall	Auditory Overall	Overall	Overall
S1	122	0.0232	0.0257	0.942	0.917	0.556	0.533	0.85	0.767	0.733
S2	3	0.0001	0.0113	0.885	0.972	0.694	0.483	0.733	0.7	0.833
S3	142	0.0400	0.0163	0.96	1	0.9444	0.62	0.98	0.82	0.95
S4	72	0.0219	0.0194	0.7885	0.9444	0.9722	0.5	0.9167	0.6667	0.5333
S5	58	0.0151	0.0151	0.94	0.97	1	0.47	0.85	0.72	0.82
S6	156	0.0331	0.0282	0.9615	1	1	0.9833	1	0.9333	0.8667
S7	144	0.0275	0.0238	0.9615	0.9167	0.9722	0.75	0.8833	0.8167	0.8167
S8	141	0.0275	0.0375	0.923	0.972	0.972	0.833	0.983	0.883	0.9
S9	110	0.0213	0.0263	0.8846	0.1389	0.8056	0.5833	0.7	0.7833	0.7833
S10	45	0.0151	0.0225	0.9038	0.9722	0.9444	0.5667	0.9833	0.7833	0.8167
S11	137	0.0331	0.0182	0.7692	0.8611	1	0.6833	1	0.8667	0.9167
S12	86	0.0182	0.0219	0.9615	0.9167	1	0.6667	0.9333	0.7833	0.75
S13	133	0.0275	0.0213	0.9038	1	1	0.55	0.7833	0.7333	0.5333
S14	74	0.0213	0.0126	0.8846	1	0.9722	0.6	0.9833	0.8166	0.6666
S15	124	0.0257	0.0207	0.923	0.8611	0.5555	0.8333	0.8333	0.6	0.8166
S16	85	0.0163	0.0288	0.9615	1	0.9722	0.8666	0.9833	0.9166	0.9833
S17	141	0.0356	0.0194	0.9038	1	1	0.8833	1	0.9833	0.9667
S18	141	0.0263	0.0275	0.9615	1	1	0.95	0.9833	0.9333	0.95
S19	159	0.03	0.035	0.9038	0.9444	1	0.8833	0.9167	0.9167	0.95
S20	74	0.02	0.0151	0.8269	1	0.83	0.4	0.9	0.5667	0.62
S21	88	0.0219	0.0157	0.9039	1	0.67	0.5667	0.85	0.8667	0.85
S22	41	0.0369	0.0244	0.9423	0.8611	0.8333	0.65	0.8167	0.7	0.7
S23	157	0.04	0.0255	1	0.9167	1	0.8833	0.8333	0.8833	0.9167
S24	135	0.0388	0.0151	0.9615	1	1	0.7833	1	0.85	0.85
S25	130	0.0263	0.03	0.8461	1	0.9722	0.6167	0.9667	0.7	0.6833
S26	119	0.025	0.025	0.9423	0.8333	0.6944	0.4833	0.9167	0.8	0.7333

S27	92	0.0157	0.0257	0.9423	0.9444	0.8333	0.6167	0.85	0.7333	0.75
S28	132	0.0288	0.0213	0.9615	1	0.8889	0.55	0.9167	0.8833	0.8667
S29	161	0.04	0.0263	0.9423	1	1	0.78	0.9167	0.9	0.8833
S30	98	0.0213	0.02	0.9615	1	1	0.65	0.1	0.8667	0.7667
S31	121	0.0238	0.0263	0.9038	0.9167	1	0.90	0.9667	0.9333	0.9333
S32	106	0.0238	0.0176	0.9615	0.9167	0.4722	0.6167	0.65	0.9167	0.9667
S33	119	0.0207	0.0282	0.9615	1	1	0.7667	0.95	0.65	0.75
S34	69	0.0182	0.0157	0.9615	0.8889	0.8611	0.5	0.9	0.7	0.6833
S35	103	0.0188	0.02	0.8846	0.8055	0.6389	0.8	0.8	0.85	0.8833
S36	156	0.0388	0.0213	1	0.75	0.8889	0.8333	0.9667	0.8167	0.7833
S37	148	0.0394	0.0219	0.9423	0.9444	0.9722	0.8	0.9833	0.7333	0.8333
S38	147	0.0306	0.0244	0.9038	1	0.6389	0.6667	0.9833	0.7667	0.65
S39	145	0.0388	0.0213	0.9230	0.9722	1	0.95	0.9833	0.9	0.7833
S40	136	0.0275	0.0263	0.9615	0.9444	0.9444	0.5333	0.7333	0.7667	0.6167
S41	154	0.04	0.0188	0.9038	0.8611	0.9444	0.8	0.95	0.8833	0.9
S42	8	0.0007	0.0157	0.8846	0.7778	0.9167	0.5333	0.8667	0.5667	0.63
S43	125	0.0219	0.0306	0.7692	1	0.9722	0.5833	0.6167	0.6667	0.5
S44	89	0.0188	0.0213	0.9423	0.8055	0.9444	0.6333	0.9333	0.7667	0.7
AVG (SD)	111.95 (39.92)	0.02526 (0.00966)	0.02255 (0.00593)	0.92 (0.05)	0.92 (0.14)	0.89 (0.15)	0.69 (0.15)	0.89 (0.09)	0.8 (0.1)	0.8 (0.13)

*Philadelphia Naming Test (PNT; Roach et al., 1996), ** Pyramids and Palm Trees (PPT, Howard & Patterson, 1992), ***PALPA (Psycholinguistic Assessments of Language Processing in Aphasia, Kay et al., 1996): a - Minimal Pair Same/Different Judgement, b – Word Rhyme Judgement, c – Auditory Synonym Judgement, d – Written Synonym Judgement

Table S3: Performance Variables (summarized across all treated items)

Participant	Sessions	Trials	Minutes	Trials/ Hour	Features*	Features* /Trial
S1	36	694	3884	10.72	4987	7.19
S2	23	246	2122	8.63	438	1.78
S3	37	411	2933	8.4	3388	8.24
S4	30	369	2943	7.52	1625	4.4
S5	28	274	2299	7.15	1656	6.04
S6	30	344	3330	6.2	2274	6.61
S7	32	347	4011	5.19	1985	5.72
S8	30	541	3698	8.78	4608	8.52
S9	32	522	3767	8.31	3361	6.44
S10	28	258	2927	5.29	834	3.23
S11	28	443	2990	8.9	4365	9.85
S12	29	320	2915	6.59	2346	7.33
S13	29	333	2837	7.04	2376	7.14
S14	28	511	3524	8.7	2851	5.06
S15	28	703	3045	13.85	6190	8.81
S16	26	670	3045	13.2	6609	9.86
S17	31	854	3585	14.29	8531	9.98
S18	27	446	3014	8.88	3104	6.96
S19	30	1044	2770	22.61	9492	9.09
S20	32	420	3188	7.9	2354	5.60
S21	26	437	2795	9.38	4008	9.17
S22	26	457	2835	9.67	4305	9.42
S23	28	714	3015	14.21	7472	10.46
S24	30	665	2970	13.43	6791	10.21
S25	29	770	2840	16.27	7205	9.36
S26	29	546	2910	11.26	4728	8.66
S27	25	650	2970	13.13	5399	8.31

S28	27	1029	3305	18.68	10467	10.17
S29	30	393	2681	8.79	3513	8.94
S30	30	603	3145	11.5	4347	7.21
S31	30	594	3215	11.08	4852	8.17
S32	27	729	2767	15.81	7198	9.87
S33	29	797	2947	16.22	6652	8.35
S34	26	540	3042	10.65	3833	7.1
S35	31	847	3486	14.58	7322	8.64
S36	28	1302	3188	24.5	12949	9.95
S37	26	324	3130	6.21	2983	9.21
S38	28	1366	2987	27.44	13676	10.01
S39	26	741	2906	15.29	7683	10.37
S40	27	629	2592	4.12	5600	8.99
S41	31	1293	3424	2.64	13764	10.65
S42	20	269	2786	10.35	280	1.04
S43	32	666	3440	5.16	4618	6.93
S44	29	761	3003	3.94	6832	8.98

*Features = participant-generated features